Kollokationen

Lexikalische Akquisition und lexikografische Verwertung

Am Beispiel der Substantiv-Verb Kollokationen im Portugiesischen und dem Programmpaket PECCI

Heike Stadler

2006

Diplomarbeit
Nr. 45

Institut für Maschinelle Sprachverarbeitung (IMS)
Azenbergstraße 12
D-70174 Stuttgart

Erstgutachter: PD Dr. Ulrich Heid
Zweitgutachter: Prof. Dr. Hinrich Schütze

Bearbeitungszeitraum: 18.7.2005-10.1.2006

Vielen Dank an meine Betreuer Ulrich Heid und Hinrich Schütze

Inhaltsverzeichnis

0. Einleitung 1

1. Kollokationen und Kookkurrenzen 3

2. Computerlinguistische Methoden der corpusbasierten Kollokationsakquisition 11
   2.1. Statistische Assoziationsmaße für Kookkurrenzen 11
   2.2. Linguistische Corpusaufbereitung 17
   2.3. Akquisitionsmethoden 19
      2.3.1. Smadjas Xtract 21
      2.3.2. Seretan/Nerima/Wehrli 23
      2.3.3. Heid et al. 24
   2.4. Automatische semantische Klassifikation von Kollokationen 27
      2.4.1. Lexikalische Funktionen und Kollokationen 28
      2.4.2. Das Klassifikationsverfahren von Wanner 34

3. Kollokationen in lexikografischer Theorie und Praxis 38
   3.1. Kollokationen und verwandte Wortkombinationen 38
      3.1.1. Oxford Dictionary of Current Idiomatic English (Cowie) 38
      3.1.2. BBI Combinatory Dictionary of English (Benson) 39
      3.1.3. Hausmann 41
      3.1.4. Grossmann/Tutin 45
      3.1.5. Konzeptuelle Kollokationen (Heid) 46
      3.1.6. Kollokationssystematik 49
   3.2. Präsentation von Kollokationen in Wörterbüchern 50
      3.2.1. Der Ort der Kollokation im Wörterbuch 50
      3.2.2. Kollokationen in allgemeinsprachlichen Wörterbüchern 55
      3.2.3. Informationstypen in (Kollokations)wörterbüchern 56
      3.2.4. Aufnahmekriterien und Usus in Kollokationswörterbüchern 64
   3.3. Divergenzen bei der Übersetzung von Kollokationen 69
   3.4. Äquivalenzbeziehungen zwischen deutschen und portugiesischen Substantiven 72

4. Lusitanistik und Kollokationen 77
   4.1. Forschungsüberblick 77
   4.2. Portugiesische linguistische Ressourcen im Internet 82

5. Extraktion der Substantiv-Verb Kollokationen - das Programmpaket PECCI 85
   5.1. Corpusbeschreibung und Corpusaufbereitung 85
   5.2. PECCIs Programmarchitektur 87
   5.3. Evaluierung der Extraktionsergebnisse 92
      5.3.1. Anfragen und Ergebnisse bei PECCI und der Linguateca 92
      5.3.2. Verbindividuelle Suchraumeinstellung 96
      5.3.3. Syntaktische Relationen der Kollokationen 100
      5.3.4. Weitere Evaluierungsmöglichkeiten 102

6. PECCIs Anwendung in der Lexikografie am Beispiel der Substantive der Gefühle 103
6.1. Aspekte der Kollokationsbeziehungen und computationelle lexikografische Darstellung von Kollokationen 104
6.1.1. Interne Kollokationsrelationen 104
6.1.2. Kollokationen in Online-Wörterbüchern 108
6.1.3. Darstellung der verbalen Kollokationen der Gefühlssubstantive 113
6.1.3.1. Form der Wörterbuchartikel 113
6.1.3.2. Vom PECCI-Output zum Wörterbucheintrag 115
6.1.3.3. Semantische Beschreibung der Kollokationen 119
6.2. Wörterbucheinträge portugiesischer Gefühlssubstantive 125
alegria, ciúme(s), esperança (em), inveja, susto, medo, ódio
6.3. Differenzierung polysemer und synonymer Substantive anhand der Kookkurrenzdaten 148
pena, inclinação, raiva, fúria
6.4. Varietätenspezifische Kollokationen 152

7. Clustering der portugiesischen Gefühlssubstantive 153
7.1. Das Clusterverfahren K-Means 153
7.2. Exemplarische Anwendungen von Clusterverfahren in der Computerlinguistik 155
7.3. Die Ergebnisse des Clustering und die Klassifikation der Gefühlssubstantive anhand semantischer Eigenschaften bei Mel'čuk und Wanner (1994) 157

Literaturverzeichnis 168

Anhang

A Quellcode
A1 Quellcode: Cetemp 178
A2 Quellcode: Cetenf 180

B PECCI
B1 PECCI: Programmdokumentation 183
B2 PECCI: Installationshinweise 190

C Ausgabedateien
C1 SentimentoCetemp/AusgabeNomina - t-score, MI (partiell) 191
C2 SentimentoCetemp/AusgabeNomina - log-like, t-score (partiell) 200
C3 SentimentoCetemp/AusgabeNomina4Scores (partiell) 204
C4 SentimentoCetemp/AusgabeWortfeld (partiell) 205
C5 SentimentoCetemp/AusgabeVerben (partiell) 209
C6 SentimentoCetemp/Samplerelevanz (partiell) 210
C7 SentimentoCetemp/ClusterNomen/Ergebniskmeans100 211

Abbildungsverzeichnis

Abb. 1 Precision-Kurven für PNV-Daten (Evert/Krenn 2001: 191) 17
Abb. 2 Recall-Kurven für PNV-Daten (Evert/Krenn 2001: 191) 17
Abb. 3 Tiefensyntaktische Beziehungen der Funktionsverben zu ihren Argumenten (Mel'čuk 1998: 39) 30
Abb. 4 Typologie der Wortkombinationen (Hausmann 1984: 399) 41
Abb. 5 Kombinatorik linguistischer Objekte (Hausmann 2005: 8) 43
Abb. 6 Kollokationssystematik und Definitionskriterien 49
Abb. 7 Informationstypen für Substantiv-Verb Kollokationen (Heid 2004:731) 57
Abb. 8 Kollokationale Divergenzen 69
Abb. 9 Übersetzungsäquivalenz der 40 untersuchten Gefühlssubstantive im portugiesischen Wortfeld und bei Mel'čuk/Wanner (1994) 75
Abb.10 PECCIs Programmarchitektur 90

Notationen

N	Anzahl der Tokens in einem Text
$\bar{x}$	Mittelwert
μ	Erwartungswert
s^2	Varianz
$P(A \mid B)$	Wahrscheinlichkeit von A unter der Bedingung B
b$(k; n,x)$	Binomialverteilung
$\binom{n}{k}$	Binomialkoeffizient
$\log a$	Logarithmus von a
$I(x\,;y)$	Mutual Information
$\vec{x}$	Reellwertiger Vektor: $\vec{x} \in \mathbb{R}^n$
$\mathsf{P}(A)$	Potenzmenge von A
$\lvert\vec{x}\rvert$	Euklidische Länge von $\vec{x}$

0. Einleitung

Eine Kollokation im allgemeinsprachlichen Gebrauch ist eine "Ordnung nach der Reihenfolge" (*Duden* 2005). Zurück geht der Begriff auf das lateinische *collocatio* 'Stellung, Anordnung'. In der Linguistik wurde der Begriff erst nach der Mitte des 20. Jahrhunderts prominent. Trotzdem (oder gerade deswegen?) erscheint es unmöglich eine Kollokationsdefinition zu geben, die die verschiedenen Bereiche der Kollokationsforschung vereint. Ihre Wurzeln hat diese in mehreren Richtungen der Linguistik: "in der statistisch-syntagmatisch vorgehenden Schule des britischen Kontextualismus, in den eher philologisch orientierten, aus der kontrastiven Linguistik und Lexikografie erwachsenden Kollokationsmodellen und nicht zuletzt in den Forschungen zu semantischen Vereinbarkeitsbeziehungen und begrifflichen Relationen" (Steyer 2000: 103).
Ein Kollokationsbegriff, der zunächst verdeutlichen soll, um welche Art von Wortverbindungen es sich hier handeln kann, stammt aus der Computerlinguistik:

> - a collocation is a binary combination of lexical items,
> - a collocation possesses a coherent syntactic structure, i.e. the base and the collocate always possess the same grammatical function with respect to each other,
> - a collocation is a lexically restricted word combination, i.e. it cannot be constructed using universal (semantic) selectional restriction rules; rather, the base predetermines the set of lexical items (collocates) it may appear with in a combination based on idiosyncratic, collocation type-specific grounds.
>
> (Wanner 2004: 98)

Kollokationen umfassen Wortkombinationen wie: *Fahrrad fahren, Spaziergang machen, Zähne putzen, Hoffnung hegen, verfaulte Zähne, gelbe Zähne, starker Raucher, schwer verletzt, Wutanfall* und *Knoblauchzehe.* Für den Muttersprachler erscheinen die Ausdrücke trivial, darüber, dass man kein **schwerer Raucher* ist und man keine **beigen Zähne* hat, macht er sich keine Gedanken. Erst im interlingualen Vergleich fällt der sprachimmanente Unterschied eklatant ins Auge: im Englischen **reitet* man sein *Fahrrad* (*ride a bike)* im Portugiesischen **geht* man damit (*ir de bicicleta),* der *Spaziergang* wird **genommen (*engl. *take a walk)* oder **gegeben* (port. *dar um passeio*), die Zähne werden **gebürstet* (engl. *brush the teeth)* oder **gewaschen* (port. *lavar os dentes*). Im Portugiesischen **wärmt* man die *Hoffnung* und *hegt* sie nicht *(acalentar esperança)*, die *verfaulten Zähne* sind **verdorben* (*dentes estragados*) und eine *Knoblauchzehe* entspricht einem **Knoblauchzahn* (*dente de alho).* Problematisch wird es im zwischensprachlichen Vergleich auch mit den Wortarten und der Wortanzahl. Im Deutschen *ist* man *eifersüchtig,* im Portugiesischen **hat* man *Eifersüchte* (*ter ciúmes*). Das Englische *a gust of anger* hat im Französischen mit *une bouffée de colère* ein Äquivalent, der **Wutstoß* wird aber im Deutschen zum *Wutanfall*, die Präposition entfällt.
Gegenstand dieser Arbeit sind im Besonderen die Substantiv-Verb Kollokationen im Portugiesischen. Sie werden mit Hilfe des Perl-Programms PECCI (Program for the Extraction of Collocations and Cluster Information) aus den mit SGML-Tags aufbereiteten Corpora *Cetempúblico* und *Cetenfolha* extrahiert. Weiter restringiert wird der Untersuchungsbereich auf die paradigmatische Behandlung von Substantiven der Gefühle. Hier sei jedoch angemerkt, dass eine Erweiterung um andere Wortfelder in PECCI generisch angelegt ist.

Die Auswahl der Emotionsnomina als untersuchtem Feld orientiert sich an einem 1994 erschienenen Artikel von Igor Mel'čuk und Leo Wanner: "Lexical Co-occurrence and Lexical Inheritance. Emotion Lexemes in German: A Lexicographic Case Study". Darin wird die Kombinatorik von 40 Gefühlssubstantiven des Deutschen mit bestimmten Verben mittels lexikalischer Funktionen dargestellt. Lexikalische Funktionen bieten ein Modell, mit dem Kollokationen in einer Metasprache zu systematisieren sind. Die Daten für das Deutsche dienen als Ausgangspunkt für die Lexikoneinträge der portugiesischen Gefühlssubstantive.

Kapitel 1 gibt zunächst einen Überblick über die Verwendung der Termini 'Kollokation' und 'Kookkurrenz' in den letzten 50 Jahren bis heute. Computerlinguistische Methoden der corpusbasierten lexikalischen Akquisition von Kollokationen werden in Kapitel 2 vorgestellt, ihre Resultate bilden die Basis für die weitere Nutzung von Kollokationsdaten. Statistische Assoziationsmaße, linguistische Corpusaufbereitung, exemplarische Akquisitionsverfahren und die automatische semantische Klassifikation von Kollokationen sind hier Themen. Kapitel 3 zeigt Kollokationen in lexikografischer Theorie und Praxis: die Abgrenzung zu verwandten Wortkombinationen und die Darstellung in Kollokationswörterbüchern und allgemeinsprachlichen Wörterbüchern werden diskutiert, mögliche Divergenzen bei der Übersetzung von Kollokationen illustriert, und die Äquivalenzbeziehungen zwischen den deutschen und portugiesischen Gefühlssubstantiven charakterisiert.

Kapitel 4 stellt Forschungsansätze zu Kollokationen aus der Lusitanistik vor, sowie portugiesische linguistische Ressourcen im Internet, die auch die Corpora zur Verfügung stellen, aus denen die Kollokationen mit PECCI extrahiert werden. Kapitel 5 beschreibt die untersuchten Corpora, deren Aufbereitung zur Weiterverarbeitung, das Programmpaket PECCI zur Akquisition der Substantiv-Verb Kollokationen und die Evaluierung der Extraktionsergebnisse.

Kapitel 6 beleuchtet verschiedene Aspekte der Kollokationsbeziehungen, wie die internen Kollokationsrelationen, die auf semantischen Kriterien beruhen, weitere Darstellungsmöglichkeiten von Kollokationen in elektronischen Wörterbüchern, und motiviert außerdem die Darstellungsform der verbalen Kollokationen portugiesischer Gefühlssubstantive. Die von PECCI extrahierten Substantiv-Verb Kollokationen werden für einige der untersuchten portugiesischen Gefühlssubstantive exhaustiv in dem gewählten Lexikonformat verzeichnet, das Konzept der Wörterbucheinträge ist an die Theorie der lexikalischen Funktionen angelehnt. Beispielhaft werden Einträge von Gefühlssubstantiven des Portugiesischen gegeben, in deren Darstellung als Vergleichsdaten auch die Einträge der entsprechenden deutschen Emotionsnomina bei Mel'čuk und Wanner (1994) und, falls vorhanden, die Übersetzungen der Kollokationen aus verschiedenen (deutsch-)portugiesischen Wörterbüchern integriert sind. Desweiteren wird die Beziehung zwischen Wortbedeutungsunterscheidung polysemer und synonymer Substantive und den Kookkurrenzdaten aufgezeigt, sowie Beispiele für die varietätenspezifische Verwendung von Kollokationen in Portugal und Brasilien gegeben. In Kapitel 7 werden Clusterverfahren auf die Frequenzen der Kookkurrenzen angewandt und die Ergebnisse analysiert. Der Sinn ist auch bei diesen Verfahren über die automatisierte Auswertung von Sprachdaten, die (maschinellen) lexikografischen Möglichkeiten zu erweitern.

1. Kollokationen und Kookkurrenzen

> Zu seinem Erstaunen stellt Momo fest, dass derzeit eine Art Krieg stattfindet, ein Terminologiekrieg, der Krieg um die Besetzung des linguistischen Terminus *Kollokation*. Es stehen sich gegenüber der basisbezogene Kollokationsbegriff, wie er für das Fremdsprachenlernen und die darauf ausgerichtete Lexikografie unverzichtbar ist - und auf der anderen Seite der computerlinguistische Kollokationsbegriff, der damit jede Art von Clusterbildung meint. Der computerlinguistische Kollokations-begriff ist nicht auf das Fremdsprachenlernen ausgerichtet. Er ist dafür verwertbar, weil er den basisbezogenen Kollokationsbegriff als Teilmenge in sich birgt, aber dazu bedarf es vieler weiterer Operationen, die vorerst niemand durchführt.
> (Hausmann 2004: 320-321)

Den Grund für die Auseinandersetzung Hausmanns mit dem Terminus 'Kollokation' liefert die Existenz zweier Wörterbücher des Englischen, die dem Titel nach ähnliches versprechen und doch ganz verschiedenes bieten. *A Dictionary of English Collocations: Based on the Brown Corpus* von Göran Kjellmer erschien 1994 und zeigt eine Ansammlung von einigen Wörtern, die vor oder nach dem gesuchten Ausdruck stehen:

	EF	IF	RF	TC	DI	
HOPE CTy45; CF150; CTe 84						
HOPE FOR b	6	8		4	2	
HOPE FOR d	2	2		L	0	
HOPE OF b	10	22		6	2	...
ALL *HOPE* a	2	2		2	1	
ANY *HOPE* a	2	2		2	1	
BEST *HOPE* FOR ab	2	2	*	2	4	...
IN THE *HOPE* OF ab	5	5	*	5	5	
IS THE *HOPE* OF gab	2	2	*	2	4	
THE ONLY *HOPE* a	3	3	*	B	3	
A NOTE OF *HOPE* aba	2	2	*	2	4	

(*A Dictionary of English Collocations* 1994)

Begleitet werden die Wörter von Minuskeln, die ein für Kollokationen entworfenes klassifikatorisches System widerspiegeln, und verschiedenen Frequenzangaben.[1] Das *Oxford Collocations Dictionary for Students of English* (2002) hingegen ist eher auf die Bedürfnisse eines menschlichen Wörterbuchbenutzers zugeschnitten. Den interessiert in der Regel nicht eine Liste mit beliebigen frequenten Kontexten eines Wortes, er möchte vor allem im Bereich des Fremdspracherwerbs präzisere und prägnantere Angaben. Diese findet er im *Oxford Collocations Dictionary for Students of English:*

> **hope** *noun*
>
> 1 belief that sth you want will happen
> · ADJ. **considerable, fervent, great, high** *a feeling of considerable hope* ◆*It is my ...*
> · QUANT. **flicker, glimmer, ray, spark** *I looked at her an felt a glimmer of hope.*
> · VERB + HOPE **be full of, cherish, entertain, have, see** *Lord Mountbatten secretly cherished hopes that Charles would mary his granddaughter.* | **express, voice** *The Mexican president expressed hope for cooperation on trade.* | **share** | **pin** *He ...*
> · HOPE + VERB **lie, rest** *Her only hope lay in escape.* | **grow, rise** *Hopes of a peaceful end to the strike are now growing.* | **flare (up), spring (up), surge** *Hope ...*
> · PREP. **beyond ~** *damaged beyond hope of repair.* **in ~ of, in the ~ that** *I am ...*
> · PHRASES **every/little/no/some hope of sth** *We have every hope of completing ...*

1 Das recht ausgiebige Klassifikationsschema und die Aufschlüsselung der Frequenzangaben werden in der Einführung des Wörterbuchs beschrieben.

hope *verb*
· ADV. **desperately, fervently, really, sincerely, very much** *hoping desperately that their missing son would come home.* ◆ *I sincerely hope that you will be successful.*
· VERB + HOPE **(not) dare (to)** *I scarcely dared hope the plan would succeed.*
| begin to | continue to
· PREP. **for** *We are hoping for good weather.*
· PHRASES **hope against hope** (= to continue to hope for sth even though it is very unlikely), **hope for the best** (= to hope that sth will happen successfully, especially ...
(*Oxford Collocations Dictionary for Students of English* 2002)

Die Wörterbucheinträge gliedern sich nach der Wortart der Lemmata und der Wortart der Kollokate sowie derer grammatischen Relation (VERB+HOPE, HOPE+VERB). Innerhalb der einzelnen Wortarten werden semantisch motivierte lexikalische Felder gebildet und mitunter Beispielsätze gegeben. Auf den ersten Blick zeigt sich, dass dies für den Fremdspracherwerb eine geeignete Form der Kollokationsdarstellung ist.
Hausmann wehrt sich gegen einen Kollokationsbegriff[2], der ab den 50er Jahren vom Britischen Kontextualismus geprägt wurde: "Collocation is the syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur, at n removes (a distance of n items) from an item x, the items a,b,c ... " (Halliday 1961: 276). Eines der Grundkonzepte des Britischen Kontextualismus betrifft die Trennung von Grammatik und Lexik, wie sie von J.R. Firth formuliert wurde. Firth unterscheidet dementsprechend zwischen *colligations ("*the interrelation of grammatical categories in syntactical structure") und *collocations ("*Collocations are actual words in habitual company") (Firth 1957: 14-15). Während die Grammatik die paradigmatischen und syntagmatischen Relationen grammatischer Kategorien darstellt, beschreibt die Lexik die paradigmatischen und syntagmatischen Beziehungen, die zwischen einzelnen Lexemen bestehen. Die lexikalische Beziehung zwischen Lexemen erscheint rein linear: "... lexis seems to require the recognition merely of linear co-occurence together with some measure of significant proximity, either a scale or at least a cut off point" (Halliday 1966: 152).
Die Identifikation von Kollokationen geschieht also syntagmatisch ungeachtet morphologischer und grammatischer Aspekte mittels Frequenzdaten und darauf angewandte mathematische Verfahren. Im Gegensatz zum Saussureschen und Amerikanischen Strukturalismus, die die *Langue* bzw. *Kompetenz* (im Gegensatz zu *Parole* und *Performanz*) als die relevanten Ebenen der Sprachbeschreibung ansehen, betrachtet Firth Sprache wieder unter dem Aspekt ihres sprachlichen, situativen und kulturellen Kontextes. Unterstützt wird dieser Ansatz durch die technischen und algorithmischen Fortschritte in der Elektronischen Datenverarbeitung, die es seit den 70-er Jahren ermöglichen, größere Textcorpora in elektronischer Form anzulegen und zu bearbeiten.
Die Bedeutung eines Wortes wird nicht mehr mit semantisch-inhaltlichen Kriterien gegeben, sondern über dessen Kontext, die Beschreibung der "formalen" Unterschiede. Der Auffassung Firths, dass Wörter ihre Bedeutung durch ihr charakteristisches Vorkommen mit anderen Wörtern in ihrer unmittelbaren linguistischen Umgebung erlangen[3], folgen Halliday und Sinclair[4]. Sie referieren auf die untersuchte lexikalische Einheit mit *node*, die *collocates* sind diejenigen Einheiten, die sich innerhalb der Kollokationsspanne (*collocational span)* der untersuchten Einheit befinden. Ein c*luster* umfasst alle festgestellten Kollokate des *node*,

2 Einen genauen Überblick über die Entwicklung des Kollokationsbegriffs geben Bahns (1996: 6-25), Bartsch (2003: 27-64), Klotz (2000: 63-100) und Lehr (1996: 7-61).
3 Formuliert wurde dieser Gedanke von Firth bereits 1935 (Firth [1935] 1964: 19)
4 Dargestellt haben sie ihre Theorie in Halliday/McIntosh/Strevens (1964), Halliday (1966), Sinclair (1966).

Kollokationsreihen (*collocational range*) unterteilen ihn in Gruppen von Wörtern, die ihrerseits signifikant oft miteinander interkollokieren. Das *lexical set* entspricht einem paradigmatisch durch gleiches Kollokationsverhalten seiner Teilnehmer bestimmten Feld. Halliday nennt als Beispiel eines lexikalischen Sets *bright, shine* und *light,* die mit den Lexemen *sun* und *moon* kollokieren (1966: 158); Sinclair verdeutlicht es anhand des Tripels *tome, paperback, cruelty*, in dem nur *tome* und *paperback* mit *edition, bookshop* und *print* kollokieren (1966: 410-411). Die erste ausführliche Sprachanalyse mittels Computer im Umfeld des Britischen Kontextualismus erfolgte 1973 auf einem Corpus von ca. 150.000 Wörtern (Jones/Sinclair 1973). Die Ausgabe der Daten erfolgte in Rankinglisten und statistischen Tafeln.

Vor diesem Hintergrund ist die Entstehung des *Dictionary of English Collocations* (1994) von Göran Kjellmer auf der Grundlage des *Brown Corpus* mit 1 Millionen Wörtern zu verstehen. Die Kollokate des als Lemma verzeichneten *node* werden hier sortiert nach ihrer Position relativ zum *node* in alphabetischer Reihenfolge ungeachtet ihrer Wortartenzugehörigkeit verzeichnet. Eine unter syntaktischen Kriterien vorgenommene Auswahl der Wortverbindungen findet aber auch hier statt: "..., collocations have been defined as such recurring sequences of items as are grammatically well formed" (Introduction: XIV). Wortkombinationen wie *but too, day but, Editor Sir*, die keine "organische Wechselbeziehung" zeigen, werden somit ausgeschlossen. Die Zuordnung jeder Wortverbindung zu einer Klasse des auf grammatischen Relationen beruhenden Klassifikationsschemas wird manuell vorgenommen und durch Kleinbuchstaben bei der jeweiligen Wortkombination dargestellt. Recht viel Platz nimmt das Aufführen der Frequenzdaten ein.

Das *Oxford Collocations Dictionary* ist ebenfalls corpusbasiert entstanden. Grundlage ist das *British National Corpus* mit 100 Millionen Wörtern. Doch scheint hier ein anderes Verständnis von Kollokationen vorzuliegen. In der Mikrostruktur der Lemmata wird auf die Wortartenzugehörigkeit der Kollokate explizit Bezug genommen, die Kollokate sind primär nach ihrer Wortart sortiert. Die extrahierten Daten erscheinen im Wörterbuchartikel wohl geordnet, gerade durch das Einschließen grammatischer Information in die Kollokationskonzeption. Ausschlaggebend für den Aufbau der Mikrostruktur ist nicht mehr der einfache lineare Abstand des Kollokators vom Bezugswort, sondern das Vorkommen der Kollokation in einer durch die Syntax vorgegebenen Struktur.

Das lexikografische Interesse an Kollokationen, das sie dem Wörterbuch konsultierenden Menschen näher bringen will, regte sich in den 70-er Jahren. Wörterbücher, die explizit corpusbasiert entstehen, treten in den Vordergrund. Eine präzise Kollokationsdefinition und ein corpusbasiertes Vorgehen prägt das neue Konzept von Wörterbüchern, die Wortkombinationen verzeichnen. Der Kollokationsbegriff wird im *Oxford Dictionary of Current Idiomatic English* (1975) von Cowie und im *BBI Combinatory Dictionary of English: A Guide to Word Combinations* (1986) von Benson in den Vorworten ausführlich erläutert. In allgemeinsprachlichen Wörterbüchern wie dem *Collins COBUILD English Language Dictionary* (1987) von Sinclair werden aktuelle Corpusbelege zu Illustrationszwecken eingebunden, um typische Kontext-, Kollokations- und Grammatikstrukturen zu zeigen.

Es rückt eine neue Auffassung von Kollokationen in den Vordergrund, die lexikografisch-didaktischer Natur ist. Sie stellt mehrere Konzepte des Britischen Kontextualismus in Frage: das Frequenzkriterium, den Ausschluss der inhaltlichen Bedeutung und das Verhältnis von Kollokationen zur syntaktischen Struktur. Übernommen wurde der Terminus, die Theorie der Kollokation wurde different definiert: "Eine Theorie der Kollokation wird vor allem

zweierlei zu leisten haben. Sie muss einerseits die Kollokation als charakteristische Zweierkombination abgrenzen gegen unspezifische, banale Zweierkombinationen, die der *parole* und nicht der *langue* angehören. Zum zweiten muß sie den Status der beiden Kombinationspartner in dieser Zweierkombination zueinander untersuchen" (Hausmann 1985: 118).
Neu sind zum einen die *Ko-Kreationen*, freie Verbindungen von Wörtern, deren Kombinationsfähigkeit nur durch semantische Bedingungen begrenzt ist. Sie werden "entsprechend den Regeln des Sprachsystems" kreativ zusammengestellt (Hausmann 1984: 398). Kollokationen hingegen sind, "wenn nicht Fertigprodukte, so wenigstens Halbfertigprodukte der Sprache, zwar nicht der Sprache als System, aber im Sinne Coserius der Sprache als Norm" (Hausmann 1985: 118), sie werden als Kombinationen aus dem Gedächtnis abgerufen. Nicht alle Elemente eines *clusters* bilden somit automatisch eine Kollokation. Hingegen gibt es Kollokationen, die im Sinne der Disponibilität des Wortschatzes verfügbar sind, aber nicht frequent (Hausmann 1985: 124).
Neu ist zum anderen die explizite Begrenzung von Kollokationen auf bestimmte Wortarten in bestimmten syntaktischen Mustern. Kollokationen wurden 1978 von Cowie, der sich mit deren Aufnahmekriterien und Darstellungsweise in Wörterbüchern beschäftigte, ganz allgemein definiert als: "co-occurrence of two or more lexical items as realizations of structural elements within a given syntactic pattern" (Cowie 1978: 132). Kollokationen müssen nicht direkt nebeneinander stehen und werden mitunter von unterschiedlichen grammatischen Strukturen realisiert (*canvass the theory* vs. *the theory is canvassed*). Hausmann (1989: 1010) engt die Anzahl der möglichen Kollokationstypen weiter ein: V+**N**, **N**+V, **N**+N, **N**+Adj, **V**+Adv, **Adj**+Adv. In der doppelten Nennung der Nomen-Verb Kombination spiegelt sich die mögliche Realisierung des Substantivs als Subjekt oder Objekt wider. Steht das Substantiv an erster Stelle fungiert es als Subjekt, hinter dem Verb als Objekt.
Hausmann führt das Konzept der internen hierarchischen Ordnung der Kollokation ein. Die *Basis* (oben hervorgehoben) bestimmt den hinzutretenden *Kollokator*: "Die wichtigste Basiswortart ist das Substantiv, weil es die Substantive sind, welche die Dinge und Phänomene dieser Welt ausdrücken, über die es etwas zu sagen gilt. Adjektive und Verben kommen als Basiswörter nur insoweit in Frage, als sie durch Adverbien weiter determiniert werden können" (Hausmann 1985: 119). Unter der Basis würde man eine Kollokation bei der Textproduktion im Wörterbuch suchen, diese hat man vor Augen, auch wenn einem der Kollokator fehlt (*schütteres **Haar**, commit **suicide**, **hurt** seriously)*. Mit diesem Konzept bricht Hausmann mit einer langen, valenztheoretisch bedingten Tradition, die das Verb in das Zentrum des Satzes und damit auch der lexikografischen Beschreibung setzte.
Weitgehend außer Acht gelassen wurde bisher das Kriterium, das sie in den Fokus der Lexikografie stellt. Kollokationen sind zwar semantisch transparent, aber aufgrund ihres idiosynkratischen Charakters beim Femdsprachenlernen nicht vorhersehbar und daher als ganze Syntagmen zu memorieren. Welche das genau sind, kann nur sprachkontrastiv ermittelt werden. Im Bereich des zweisprachigen Wörterbuchs ergeben sich somit wesentlich klarere Kriterien in Hinblick auf die Frage, was für lexikografische Zwecke als relevante Kollokation zu gelten hat (Herbst/Klotz 2003: 138). Es wird gefordert den Kollokationsbegriff auszudehnen und " ... auf alle nicht-lexikalisierten Kombinationen anzuwenden, die aus der Perspektive der L1 in der L2 als nicht vorhersehbar erscheinen" (Herbst/Klotz 2003: 84).

Nachdem der Begriff 'Kollokation' aus dem Schatten des Britischen Kontextualismus hinausgetreten ist, sind sich Grammatik, Lexik und Semantik wieder näher gekommen. Diese Einsicht zeigt sich auch in der Maschinellen Sprachverarbeitung: "It is probably best to restrict collocations to the narrower sense of grammatically bound elements that occur in a particular order and use the terms *association* and *co-occurence* for the more general phenomenon of words that are likely to be used in the same context" (Manning/Schütze 2002: 185). Dies befriedigt eine von Hausmann gestellte Forderung nur teilweise: "Ich nehme für mich in Anspruch, den Terminus von Firth fruchtbar weiterverarbeitet zu haben und betrachte mit dem Erscheinen des *Oxford Collocations Dictionary* den Krieg als gewonnen. Den Computerlinguisten dürfte es leicht fallen, ihren weiten Kollokationsbegriff mit einem anderen Terminus zu besetzen" (Hausmann 2004: 321). Doch neben die grammatische Verbindung der Elemente setzen auch Manning und Schütze weitere linguistische Kriterien für die Identifizierung von Kollokationen: "non-compositionality, non-substitutability, non-modifiability" (Manning/Schütze 2002: 184).
Diese Terminologie hat sich in der Computerlinguistik durchgesetzt: "In order to make a clear distinction between the two approaches to collocations, I refer to the distributional notion as cooccurrences, which encompasses both the observable (cooccurrence) frequency information and its interpretation as an indicator of statistical association" (Evert 2005a: 9). Den Kookkurrenzen gegenübergestellt wird der Ausdruck 'Kollokation' für ein "... intensionally defined concept that does not depend on corpus frequency information. ... A collocation is a word combination whose semantic and/or syntactic properties cannot be fully predicted from those of its components, and which therefore has to be listed in a lexicon " (Evert 2005a: 9).
Doch auch Stefan Evert gibt zu bedenken, dass Kollokationen aufgrund ihrer unvorhersehbaren Kombinatorik einen Grad der Lexikalisierung aufweisen müssen, es handelt sich bei Kollokationen um rekurrente Kombinationen der Sprache. Sie sollten somit bei ausreichender Corpusgröße im Corpus manifestiert, und damit in den Kookkurrenzen inkludiert sein (Evert 2005a: 10). In diesem Sinne spricht Heid bei der maschinellen Extraktion grammatisch restringierter Wortpaare von Kollokationskandidaten (*collocation candidates*), deren Abgrenzung in triviale Kombinationen und Kollokationen nur vom Menschen (*human filter*) vorgenommen werden kann (Heid 1994, 2000). Somit ist der computerlinguistische Kollokationsbegriff durch die Einbettung der beiden (ehemals) konträren Standpunkte der reichste:

	syntaktische Relation	Frequenzkriterium	humane Introspektion
Britischer Kontextualismus	−	+	−
Hausmann	+	−	+
Computerlinguistik	+	+	+

Der linguistische Kollokationsbegriff spielte ohne Corpora, Computer und Frequenzen noch keine Rolle. So stellt der "fleißige Vollidiot Computer" (Hausmann 2002: 320) zwar beliebige und banale Kookkurrenzen zur Verfügung, Freund Computer ist aber auch in der Lage, diese über das Corpus bei Bedarf mit POS-Tags (Part-of-Speech Tags) und präziser syntaktischer Information durch Parser anzureichern und uns in Form von Frequenzlisten und Werten statistischer Assoziationsmaße auf den Usus in unserer Sprache hinzuweisen. Er ruft die verfügbaren (aber nicht immer paraten) Kollokationen in unser Bewusstsein, er macht die reale Disponibilität erst möglich.

Auch der anscheinenden Diskreditierung des Begriffs der Kookkurrenz soll vorgebeugt werden. Kookkurrenzdaten können vollautomatisch von adjazenten Wörtern ohne syntaktische Relation genauso wie von Wörtern, die in syntaktischer Verbindung miteinander aber mitunter weiter auseinander stehen, gewonnen werden. Kookkurrenzdaten sind für viele Anwendungsgebiete der Maschinellen Sprachverarbeitung von großem Wert: in Sprachgenerierung und Maschineller Übersetzung begrenzen sie die lexikalische Auswahl, sie lösen Ambiguitäten in PP-Attachment, syntaktischen Parsewäldern und Nomenkomposita, helfen bei der Disambiguierung von Wortbedeutungen und dienen nicht zuletzt als Grundlage der Kollokationsidentifikation (vgl. Evert 2005a: Kapitel 1.1 "Applications of cooccurrence data").
Ein weiterer Kritikpunkt am Konzept des Britischen Kontextualismus bezieht sich auf die Gliederung der Kollokate in Kollokationsreihen. Diese setzen sich aus Kollokaten zusammen, die signifikant oft untereinander kollokieren (Sinclair 1966: 426). Dass dies nicht der Fall ist, will Klotz anhand einer Corpusanalyse, statistischer Assoziationsmaße und der Wörter *bus, train, plane, ferry* und *boat* zeigen, die alle mit dem Verb *catch* vor-kommen. Zurück führt er die mangelnde Interkollokation darauf, dass diese Lexeme als Kollokate gerade in einer paradigmatischen Beziehung zueinander stehen, so dass im Normalfall das eine oder das andere Lexem gewählt wird, nicht jedoch beide gemeinsam (Klotz 2000: 72). Sinclair waren 1966 bei der Bildung von Kollokationsreihen Phrasen wie "I went by *bus* and *train*" (Sinclair 1966: 425) vor Augen. Klotz unternimmt zunächst eine "intuitive" Einteilung des *clusters* von *catch*, die oben genannten Substantive werden neben die Reihen *eye/attention* und *sight/glim* gestellt. Die von Klotz gewählten Elemente des öffentlichen Transports entsprechen daher auch einem *lexikalischen Set*, intuitiv kollokieren sie noch mit einer ganzen Reihe weiterer Verben, sie weisen also das gleiche Kollokations-verhalten auf.
Eine paradigmatische Bestimmung kann nur über das syntagmatische Verhalten der einzelnen Mitglieder des Sets vorgenommen werden. Sie kann eindimensional angelegt sein, wenn nur das Verhältnis zu einem bestimmten Wort interessiert, oder aber über das gesamte *cluster* verlaufen. Kollokationsreihen sind daher nur Ausschnitte aus einem lexikalischen Set. Die Sprache anhand der lexikalischen Sets zu strukturieren war eines der Anliegen des Britischen Kontextualismus: "It was hoped to reveal associations between words which were part of the regularly recurring structure of the language, and to find some indication, however rudimentary, that meaninfgul "lexical sets" ... could be produced from significant collocations" (Jones/Sinclair 1973: 39). Als Beispiel eines lexikalischen Sets wird bei Halliday (1966: 158) einmal die Schnittmenge der Kollokate von *sun* und *moon*, nämlich *bright*, *shine* und *light* vorgestellt. Jones/Sinclair (1973: 40) sehen in Pluralnomina, die Zeitspannen benennen wie *minutes*, *hours* und *weeks* den Nukleus eines lexikalischen Set mit der Tendenz in der gleichen Umgebung vorzukommen. "Collocational and lexical set are mutually defining as are structure and system: the set is the grouping of members with like privilege of occurrence in collocation" (Halliday 1966: 153).
Zu jener Zeit beliefen sich die Beispiele und Gedanken eher auf theoretische Überlegungen oder Untersuchungen sehr kleiner Corpora. Mit der Zunahme der prozessierbaren Corpora wurde die Komplexität natürlicher Sprache immer deutlicher und auf die Idee, durch Kollokationsverhalten sprachstrukturierende lexikalische Sets zu bilden, werden verschiedene Clusterverfahren angewandt. Der Sinn bleibt der gleiche: Elemente zu finden, die in homogener Weise mit den gleichen Kookkurrenzdaten vorkommen, die lexikalischen Sets werden jetzt Cluster genannt. Die paradigmatische und syntagmatische Beschreibung weicht

einer Darstellung im multidimensionalen Raum, in dem es für jedes untersuchte Wort einen Vektor gibt, der aus den Häufigkeiten des miteinander Vorkommens des betreffenden Wortes mit den Kollokaten innerhalb eines gewissen Suchraums besteht.
Die Zuordnung des betreffenden Wortes zu einem bestimmten Cluster erfolgt aufgrund der Ähnlichkeit oder Differenz im Kookkurrenzverhalten: "First, the left and right neighbors of tokens of each word in the Brown corpus were tallied. These distributions give a fairly true implementation of Firth's idea that one can categorize a word by the words that occur around it. But now, rather than looking for distinctive collocations, as in chapter 5, we are capturing and using the whole distributional pattern of the word. Word similarity was then measured as the degree of overlap in the distributions of these neighbors for the two words in question" (Manning/Schütze 2002: 495-496). Dass ein mit Frequenzangaben gewonnenes Cluster nicht immer intuitiv zu erfassen ist oder einem semantisch bestimmten lexikalischen Wortfeld entspricht, zeigt das letzte Kapitel dieser Arbeit. Die Elemente eines Clusters, einzelne Substantive der Gefühle, kollokieren mit denselben Verben, doch die Ergebnisse entziehen sich mitunter der corpusunabhängigen menschlichen Introspektion, sie scheinen sich nicht unbedingt mit semantischen Kriterien zu überschneiden. Ein großer Teil der Gefühlssubstantive wird jedoch mit weiteren (quasi)synonymen Gefühlssubstantiven im gleichen Cluster zusammengefasst.
Um einer weiteren terminologischen Konfusion entgegenzuwirken, soll der *cluster* des Britischen Kontextualismus 'Kookkurrenzbereich' heißen. Die Kollokationsspanne (oder das 'Fenster') wird im Perl-Programm auch 'Suchraum' eines Nomens genannt. Als 'Kollokationspotenzial' wird die durch ein statistisches Assoziationsmaß ermittelte numerische Relation zwischen zwei Wörtern bezeichnet.[5] Ein 'Kollokationsbereich' wird aus dem Kookkurrenzbereich durch den Lexikografen gewonnen, er umfasst alle Kollokate (oder Kollokatoren bei Hausmann) einer Basis, die bestimmte linguistische Kriterien erfüllen, die Ko-Kreationen sind in ihm nicht mehr enthalten. Die interne hierarchische Unterscheidung der Kollokation in Basis und Kollokator wird übernommen, doch kann je nach Aufgabe auch einmal der Kookkurrenzbereich eines Verbs mit den Nomina von Interesse sein. Diese Möglichkeit wird in PECCI ebenso berücksichtigt wie die Bildung von Cluster, die von Verben initiiert sind.
Ist im Folgenden doch häufig von Kollokationen die Rede, wenn auch freie Wortkombinationen und daher eigentlich Kookkurrenzen gemeint sind, hat dies verschiedene Gründe. In der zitierten Literatur wird der Terminus 'Kollokation' je nach Autor und Sichtweise unterschiedlich verwandt. Bei der Beschreibung der verschiedenen Kollokationstheorien wird die dortige Terminologie übernommen. Erfolgt die Klassifikation der Kollokationen beispielsweise mit lexikalischen Funktionen, bestimmt die Teilnahme an einer lexikalischen Funktion über den Kollokationsstatus und nicht umgekehrt. Dies führt dazu, dass auch "semantisch transparente, weniger stark fixierte und vorhersehbare Kombinationen" als Kollokationen gelten (Mel'čuk 1998: 42). Im Bereich der lexikalischen Akquisition spricht man häufig von Kollokationsextraktion, auch wenn damit zunächst die Gewinnung von Kookkurrenzdaten und deren Weiterverarbeitung mit statistischen Assoziationsmaßen gemeint ist. Die Klärung des Kollokationsstatus einer Wortkombination unter linguistischen

5 Um weitere terminologische Verwechslungen auszuschließen, sei hier darauf hingewiesen, dass die Kollokationsreihen ("lexical range") bei Klotz (2000: 67) die "Kollokationsbereiche" sind. Die Cluster (lexikalischen Sets) bezeichnet Lehr (1996: 40) als "Kollokationspotential", Klotz (2000: 67) nennt sie "Kollokationsfelder".

Kriterien wäre zusätzlich vorzunehmen. Insofern extrahiert das Programm PECCI auch keine fertige Liste mit Kollokationen, sondern eine Rankingliste mit Kollokationskandidaten.
In den Verfahren der Maschinellen Sprachverarbeitung, die die Kookkurrenzdaten zur Disambiguierung weiterverwerten, ist der Status einer Wortkombination als Kollokation oder Ko-Kreation irrelevant. Relevant bleibt die Frage, ob man die Kookkurrenzdaten allein von positionellen Parametern abhängig macht oder sie unter dem Gesichtspunkt der grammatischen Korrelation gewinnt. Der Status einer Wortkombination als Kollokation oder Ko-Kreation spielt vor allem eine Rolle in der lexikografischen Theorie und Praxis. Die Entscheidung über die Aufnahme bestimmter Kookkurrenzdaten in ein Wörterbuch steht in Abhängigkeit der zugrunde liegenden Kollokationstheorie. Anhand welcher Kriterien der Linguist die Grenze zwischen Kollokationen und verwandten Kombinationen zieht und wie die lexikografische Praxis aussieht, wird Kapitel 3 zeigen. Auch in allgemeinsprachliche Wörterbücher wird das Kollokationskonzept zunehmend integriert (vgl. Kapitel 3.2.2). Grundlage ist immer die corpusbasierte lexikalische Akquisition von Kollokationen. In Kapitel 2 werden zunächst die Akquisitionsmethoden der Computerlinguistik vorgestellt. Sie basieren zum einen auf großen, elektronisch verfügbaren Corpora, die in ganz unterschiedlichem Maße linguistisch annotiert sein können, und zum anderen auf mathematischen Verfahren, die eine Sortierung der Kookkurrenzdaten nach deren statistischer Kollokationswahrscheinlichkeit vornehmen.
Im Bereich der computergestützten Lexikografie dient die lexikalische Akquisition der Beschaffung von Daten, auf denen die lexikografische Beschreibungsarbeit aufsetzen kann. Neue Wörterbücher werden erstellt mit Hilfe elektronischer Versionen früherer Ausgaben, den extrahierten Informationen aus den Corpora und der Intuition des Lexikografen. Dabei beschränkt sich die corpusbasierte lexikalische Akquisition nicht auf die Angabe der Distribution einzelner Wortformen und Lexeme oder die Extraktion von Kollokationen. Aus geparsten Texten können detaillierte Subkategorisierungsrahmen gewonnen werden. Gegenstand der lexikalischen Akquisition sind auch lexikalisch-semantische (z.B. taxonomische) Relationen sowie lexikalisch-semantische Klassen (Heid 2001: 419).
Die statistischen Verfahren mit denen semantische Informationen für die automatische Sprachprozessierung ermittelt werden, findet man bei Manning und Schütze im Kapitel "Lexical Acquisition" (2001: 265-314) erklärt. Semantische Ähnlichkeiten von Wörtern lassen sich über ihre spezifischen Kookkurrenzen beschreiben, im multidimensionalen Vektorraum berechnen und entsprechend speichern. Mögliche Anwendungsgebiete sind Textverstehen und Information Retrieval. Selektionspräferenzen oder Selektionsrestriktionen können helfen, um auf die Bedeutung eines Wortes zu schließen, das nicht im Wörterbuch verzeichnet ist. Kookkurrenzdaten dienen auch zur Disambiguierung des PP-Attachments bei Parsingverfahren. Maschinenlesbare Wörterbücher können und müssen all diese Information enthalten, die sich in Print-Medien, die für menschliche Benutzer mit kognitiven Fähigkeiten konzipiert sind, als redundant erweisen. Die Ergebnisse dieser Verfahren dienen dem Lexikografen beim Erkennen, Auswählen und Klassifizieren relevanter Worteigenschaften, die in gezielten Einträgen und dem deskriptiven Programm entsprechend in das Wörterbuch einfließen.

2. Computerlinguistische Methoden der corpusbasierten Kollokationsakquisition

2.1. Statistische Assoziationsmaße für Kookkurrenzen

Statistische Assoziationsmaße für Kookkurrenzen bilden numerische Verhältnisse zwischen (zwei) Wörtern ab, die sich aus ihrem Frequenzverhalten zueinander und gegenüber anderen Wörtern ergeben. Die Ergebnisse werden üblicherweise nach fallender statistischer Signifikanz sortiert in Kookkurrenzlisten ausgegeben. Ein mit statistischen Assoziationsmaßen ermittelter Wert gibt eine andere Art der Information an, als die einfache Frequenz der Kookkurrenz: er favorisiert die Paare, die häufiger sind, als aufgrund der Vorkommen der Einzelwörter zu erwarten ist. Bei Kollokationen wird von der Unabhängigkeitsannahme, die besagt, dass es keine Beziehung zwischen den Wörtern gibt, eine signifikante Abweichung erwartet.
Die tatsächlich beobachteten Frequenzen lassen sich auf anschauliche Weise in einer Kontingenztabelle darstellen (*acalentar* - 'erwärmen' und *esperança* - 'Hoffnung' = *acalentar esperança* - 'Hoffnung hegen' im Corpus *Cetempúblico*):

	w1 = esperança	w1 ≠ esperança	
w2 = acalentar	70 $O11$	488 $O12$	O11 + O12 = R1
w2 ≠ acalentar	11043 $O21$	174173053 $O22$	O21 + O22 = R2
	C1 = O11 + O21	C2 = O12 + O22	C1 + C2 = N

Im Gegenstück zu den tatsächlichen Okkurrenzen kann man die erwarteten Häufigkeiten für die Felder der Kontingenztabelle berechnen:

	w1 = esperança	w1 ≠ esperança
w2 = acalentar	$E11 = \frac{R1 * C1}{N}$	$E12 = \frac{R1 * C2}{N}$
w2 ≠ acalentar	$E21 = \frac{R2 * C1}{N}$	$E22 = \frac{R2 * C2}{N}$

In der Literatur zur statistischen Sprachverarbeitung[6] sowie in den konkreten Akquisitionsverfahren findet man am häufigsten folgende vier Assoziationsmaße, die den Grad der Assoziiertheit (das Kollokationspotenzial) für ein Wortpaar angeben: t-score, log-likelihood, χ^2 und Mutual Information. Berechnet werden sie mit dem entsprechenden statistischen Test.

6 Die folgenden Ausführungen basieren auf Manning/Schütze (2002, Kapitel 5): *Collocations*, Evert (2005a): *The Statistics of Word Cooccurrences,* der vom selben Autor verfassten Webseite http://www.collocations.de sowie dem Skript *Statistische Methoden in der Maschinellen Sprachverarbeitung* von Helmut Schmid (2005). Statistische Methoden werden seit den 60er Jahren in der Maschinellen Sprachverarbeitung systematisch angewandt.

Der t-score gibt die Signifikanz der Abweichung der tatsächlichen Häufigkeit eines Wortpaars von der Häufigkeit wieder, die zu erwarten ist, wenn beide Wörter zufällig verteilt sind. Der **t-Test** misst die Differenz zwischen dem Mittelwert $\bar{x}$ der Daten und dem Erwartungswert μ einer gegebenen Normalverteilung und skaliert sie mit der Varianz s^2 der Daten. Der Maximum-Likelihood-Schätzer berechnet den Erwartungswert ebenfalls aus den Daten:

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

$$\bar{x} = P(acalentar\ esperança) = \frac{f(acalentar\ esperança)}{Corpusgröße} = \frac{O11}{N}$$

$$\mu = P(acalentar)P(esperança) = \frac{f(acalentar)}{Corpusgröße} * \frac{f(esperança)}{Corpusgröße} = \frac{R1}{N} * \frac{C1}{N}$$

Die Nullhypothese H_0 besagt, dass die Okkurrenzen von *acalentar* und *esperança* unabhängig voneinander sind. Der t-Test ermittelt einen Wert, der es erlaubt mittels einer Konfidenztabelle[7] den Prozentsatz zu ermitteln, den wir der Glaubwürdigkeit der Ablehnung der Nullhypothese zugestehen. Die Varianz kann bei Annahme einer Bernoulli-Verteilung durch einen Näherungswert angegeben werden, wodurch man eine Formel erhält, die auf die Kontingenztabelle angewandt werden kann:

$$t \approx \frac{\frac{O11}{N} - \frac{R1}{N} * \frac{C1}{N}}{\sqrt{\frac{O11}{N^2}}} \approx \frac{O11 - E11}{\sqrt{O11}} \approx 8.36$$

Bei einem t-score von 8.36 und der in der Kontingenztabelle angegebenen Frequenz von *w1* und *w2* kann man der Ablehnung der Nullhypothese zu 99,9% vertrauen. Wie auch die Werte der weiteren Tests kann der t-score zum Ranking herangezogen werden. Als theoretisch problematisch erweist sich beim t-Test die Annahme der Normalverteilung (die in natürlicher Sprache üblicherweise ohnehin nicht gegeben ist) in Abhängigkeit zur binären Bernoulli-Verteilung.

Der **Likelihood-Ratio-Test** vergleicht die Wahrscheinlichkeit zweier Hypothesen, die log-likelihood zeigt, wie viel wahrscheinlicher die eine Hypothese ist als die andere. Hypothese 1 formalisiert die Unabhängigkeit der Ereignisse (die Okkurrenz von *w2* ist unabhängig von der vorherigen Okkurrenz von *w1*), Hypothese 2 formalisiert deren Abhängigkeit:

Hypothese 1: $P(w2 \mid w1) = p = P(w2 \mid \neg w1)$

Hypothese 2: $P(w2 \mid w1) = p_1 \neq p_2 = P(w2 \mid \neg w1)$

Mit dem Maximum-Likelihood-Schätzer lassen sich die Wahrscheinlichkeiten p, p_1 und p_2 errechnen. Für das tatsächliche Vorkommen von *w1*, *w2* und *w1w2* wird c_1, c_2 und c_{12} verwandt:

7 Konfidenztabellen für den t-score und χ^2 - findet man z.B. in Schickinger/Steger (2001: 241-242).

$$p = \frac{c_1}{N} \qquad p_1 = \frac{c_{12}}{c_1} \qquad p_2 = \frac{c_2 - c_{12}}{N - c_1}$$

Unter der Annahme der Binomialverteilung

$$b(k; n,x) = \binom{n}{k} x^k (1-x)^{(n-k)}$$

ist die log-likelihood definiert als $-2\log\lambda$:

$$\begin{aligned} -2\log\lambda &= -2\log\frac{L(Hypothese\,1)}{L(Hypothese\,2)} \\ &= -2\log\frac{b(c_{12,}c_{1,}p)\,b(c_2 - c_{12,}N - c_{1,}p)}{b(c_{12,}c_{1,}p_1)\,b(c_2 - c_{12,}N - c_{1,}p_2)} \\ &= -2\log\frac{L(c_{12,}c_{1,}p)\,L(c_2 - c_{12,}N - c_{1,}p)}{L(c_{12,}c_{1,}p_1)\,L(c_2 - c_{12,}N - c_{1,}p_2)} \end{aligned}$$

mit $L(k,n,x) = x^k(1-x)^{(n-k)}$.

Gelegentlich wird ein Unterschied in der Literatur gemacht zwischen den Tests, die direkt auf die Kontingenztabelle angewandt werden, und denen, die mit den gezählten Frequenzen arbeiten. Vergegenwärtigt man sich, dass die Felder der Kontingenztabellen in einem Fall die tatsächlichen Okkurrenzen der beiden Wörter *w1* und *w2* wiedergeben und im anderen Fall deren Erwartungswert, kann man auch in die Formel für log-likelihood die Felder der Kontingenztabelle mit den tatsächlich beobachteten Werten einsetzen und erhält:

$$\begin{aligned} &= -2\log\frac{L(O11\,,C1\,,p)\,L(O12\,,C2\,,p)}{L(O11\,,C1\,,p_1)\,L(O12\,,C2\,,p_2)} \\ &= 931.42 \end{aligned}$$

Doch bringt diese Formel für die Berechnung der log-likelihood mit Zahlenwerten in den Größenordnungen der untersuchten Corpora gewisse Schwierigkeiten mit sich. Berechnet man *L(O11,C1,p)* mit den Werten aus dem obigen Beispiel erhält man eine sehr kleine Zahl, die erst 389 Stellen nach dem Komma beginnt. Dies liegt daran, dass man für *p* durch die immense Corpusgröße schon einen kleinen Wert erhält (0.00006380), dieser wird dann mit *O11* (= 70) potenziert. Die exakte Berechnung von *L(O12,C1,p)* ist nur noch mit erheblichem Rechenaufwand durchzuführen, *p* wird diesmal mit *O12* (= 488) potenziert. Im Falle der denotationellen Berechnung würde dies einen sehr großen Zeitaufwand bedeuten. Helmut Schmid schlägt zur Berechnung der log-likelihood eine weitere Formel vor, die durch Umformung aus der obigen entsteht (Schmid 2005: 42):

$$\begin{aligned} -2\log\lambda = 2\,(&O11\log O11 + O12\log O12 + O21\log O21 + O22\log O22 \\ &- (O11+O12)\log(O11+O12) - (O11+O21)\log(O11+O21) \\ &- (O12+O21)\log(O12+O21) - (O21+O22)\log(O21+O22) \\ &+ (O11+O12+O21+O22)\log(O11+O12+O21+O22)\,) \end{aligned}$$

Sie wird bei der Berechnung der log-likelihood von PECCI verwendet.

Der Test, der ursprünglich auf Kontingenztabellen angewandt wurde, ist der **χ^2-Test**. Wie der t-Test vergleicht er die tatsächlichen Okkurrenzen mit den zu erwartenden Frequenzen bei Unabhängigkeit. Anders als der t-Test setzt er aber keine Normalverteilung voraus. Die χ^2-Statistik summiert die Differenzen zwischen tatsächlichen und erwarteten Werten in den Feldern der Kookkurrenzen und dividiert sie durch die Größe der Erwartungswerte:

$$\begin{aligned}\chi^2 &= \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \\ &= \frac{N(O11\ O22 - O12\ O21)^2}{(O11+O12)(O11+O21)(O12+O22)(O21+O22)} \\ &= \frac{N(O11-E11)^2}{E11\ E22} \\ &= 137507.92\end{aligned}$$

Ist die Differenz zwischen erwartetem und beobachtetem Wert signifikant, kann die Nullhypothese der Unabhängigkeit zurückgewiesen werden. Dies geschah ursprünglich wie beim t-score über das Signifikanzniveau einer Konfidenztabelle. In computerlinguistischen Anwendungen werden die Ergebnisse des Tests aber genauso zum Ranking verwendet wie die intuitiv leichter zu interpretierende Ergebnisse der log-likelihood, welche die Wahrscheinlichkeit zeigen, mit der Hypothese 1 der Hypothese 2 vorzuziehen ist. Die Werte der log-likelihood ihrerseits verhalten sich asymptotisch zur χ^2-Verteilung. Dadurch kann man die Nullhypothese gleich Hypothese1 setzen und gegen die alternative Hypothese 2 testen und die Konfidenzen der Tabelle der χ^2-Verteilung entnehmen.

Die punktweise **Mutual Information** ist ein Maß aus der Informationstheorie, das Aufschluss gibt wie viel Information *w1* in Position *i* enthält, dass *w2* in der Folgeposition *i*+1 erscheint und umgekehrt. Die Mutual Information ist ein symmetrisches Maß der gemeinsamen Information von zwei Zufallsvariablen.

$$I(x;y) = \log\frac{p(x,y)}{p(x)\,p(y)} = \log\frac{O11}{E11} = 7.58$$

Im Gegensatz zu den drei ersten Assoziationsmaßen gibt sie eher den Grad der Assoziiertheit wieder und nicht deren Signifikanz. Kookkurrenzen werden mit ihr bei gleich bleibenden Verhältnissen zu den einzelnen Okkurrenzen umso höher bewertet, je seltener sie im Corpus erscheinen. Mit der Mutual Information wird die Korrelation seltener Paare im Corpus stark überbewertet, sie nehmen im Vergleich zu anderen Tests einen höheren Platz ein. So führt ihre Art der Berechnung der Assoziiertheit zu dem Paradox, dass ein Wortpaar *w1w2*, das nur einmal im Corpus und zwar gemeinsam vorkommt, höher bewertet wird, als ein 2-faches ebenfalls nur gemeinsames Auftreten eines anderen Wortpaars.

Beim t-Test erhalten die häufigen Kookkurrenzen in der Korrelation zu deren absoluten Frequenzen den höchsten Wert. Der Ausschnitt aus einer der Dateien, die von PECCI erzeugt werden, soll das Ranking der Kookkurrenzen und die Werte der unterschiedlichen Assoziationsmaße anhand des Nomens *esperança* im Corpus *Cetempúblico* zeigen. Man kann beobachten wie die weniger häufigen Substantiv-Verb Paare vom t-Test über den Likelihood-Ratio-Test und den χ^2-Test bis zur Mutual Information langsam nach oben steigen (beispielsweise *acalentar* und *depositar*) (vgl. auch Anhang C3).

esperança: 11113

t-score			log-likelihood			chi-square			MI		
ter	26.98	870	manifestar	3637.25	411	acalentar	137507.92	70	acalentar	7.58	70
manifestar	20.18	411	ter	2750.62	870	manifestar	89724.96	411	infundir	6.50	2
ser	17.41	640	alimentar	1520.52	169	alimentar	40169.36	169	renascer	6.11	41
haver	16.64	328	perder	1296.69	223	restar	22146.39	149	alimentar	5.48	169
perder	14.62	223	restar	1201.74	149	renascer	18409.19	41	manifestar	5.39	411
alimentar	12.95	169	haver	1052.39	328	depositar	14129.83	68	depositar	5.35	68
restar	12.13	149	acalentar	931.42	70	perder	10313.27	223	restar	5.01	149
manter	10.15	110	ser	628.42	640	ter	8589.55	870	exprimir	4.93	46
dar	9.67	114	depositar	593.02	68	exprimir	6249.22	46	restaurar	4.60	15
acalentar	8.36	70	manter	543.62	110	haver	3415.30	328	nutrir	4.44	3
depositar	8.21	68	renascer	420.65	41	manter	3191.85	110	crivar	4.43	1

...

(Pecci: SentimentoCetemp/AusgabeNomina4Scores)

Eine weitere Möglichkeit zum Vergleich des Rankingverhaltens der vier Assoziationsmaße geben die Ausgabedateien in Anhang C1, C2 und C4.
Folgende Tabelle soll die Ergebnisse der statistischen Test noch einmal in Abhängigkeit vom Vorkommen der Wörter *w1* und *w2* verdeutlichen:

	O11	O12	O21	O22	t-score	log-like	chi-square	MI
Versuch 1	25	10000	300	10000000	4.93	169.70	1872.52	4.34
Versuch 2	25	300	10000	10000000	4.93	169.70	1872.52	4.34
Versuch 3	1	1	1	10000000	1.00	28.69	2499999.75	14.73
Versuch 4	2	2	2	10000000	1.41	54.61	2499999.50	14.04
Versuch 5	100	100	100	10000000	10.00	1948.07	2499975.00	10.13
Versuch 6	10	100	100	10000000	3.16	162.27	82628.18	9.02
Versuch 7	100	1000	1000	10000000	9.99	1162.43	82480.25	6.72
Versuch 8	100	100	1900	10000000	9.96	976.16	26078.89	5.57

Der Vergleich von Versuch 1 und Versuch 2 zeigt, dass die vier Tests wie zu erwarten keine Differenz in den Ergebnissen, je nach Verteilung der Okkurrenzen auf *w1* und *w2,* zeigen. Versuch 3, 4 und 5 beschreiben die extremen Verhältnisse, in denen beiden Wörter im Corpus nur gemeinsam auftreten. Man kann beobachten, dass die Werte der Mutual Information, wie oben erwähnt, bei zunehmendem Vorkommen sinken - χ^2 bietet das gleiche Verhalten, wenn auch in geminderter Weise. Für χ^2 scheint weiterhin die tatsächliche Frequenz der Kookkurrenz kaum eine Rolle zu spielen, der Test beurteilt das Vorkommen der Wörter fast ausschließlich nach dem Verhältnis von Okkurrenz zu Kookkurrenz wie auch die Werte für Versuch 6 und 7 zeigen. Der t-Test hingegen scheint dieses Verhältnis kaum zu beachten (vergleiche Versuch 5 und 7), er bezieht sich in verstärktem Maße auf die Frequenz der Kookkurrenz. Dass eine asymmetrische Verteilung der Okkurrenzen von den Tests unterschiedlich bewertet wird, veranschaulichen Versuch 7 und 8. Die log-likelihood scheint im Vergleich aller Versuche die intuitiv nachvollziehbarsten Werte zu liefern. Dass die Ergebnisse des t-Tests nicht nur mit der Frequenz der Kookkurrenz korrelieren, kann man anhand des Verbs *ser* ('sein') verdeutlichen. Anders als *acalentar* kommt es im Corpus über 3,5 Millionen mal vor, davon 123 mal in der Kombination mit *esperança<u>s</u>*. Trotz der 123 Vorkommen platziert der t-Test es an die 51. Stelle, weit hinter Verben, die nur 4 oder 5

mal in der Kombination mit *esperanças* erscheinen. Beim Likelihood-Ratio-Test steht es an Stelle 68, beim χ^2 -Text an Stelle 75 und bei der Mutual Information steht es auf Platz 92.
Darüber, welches der geeignete statistische Test zur automatischen Identifikation von Kollokationskandidaten ist, gehen die Meinungen auseinander. Die Mutual Information wird aufgrund ihrer Ergebnisse, die mitunter im umgekehrten Verhältnis zur tatsächlichen Relevanz der Kookkurrenzen stehen, meist ausgeschlossen. Beim t-Test erweist sich die Annahme der Normalverteilung gerade für große Stichproben als problematisch. Beim χ^2 -Test ist die Berechnung von kleinen Werten für Okkurrenzen und Kookkurrenzen, die in einer, für die natürliche Sprache üblichen, sehr asymmetrischen Kontingenztabelle stehen, unzureichend wegen der Annäherung an eine diskrete Binomialverteilung durch eine kontinuierliche Normalverteilung. Wie Evert berichtet, hat sich die log-likelihood in der Computerlinguistik als statistisches Maß der Assoziiertheit zwischen Wörtern als de facto Standard durchgesetzt (2005a: 113), auch Schmid (2005: 45) und Manning/ Schütze (2002: 170, 175) favorisieren sie als Assoziationsmaß. Vom mathematischen Standpunkt her gibt sie eine leicht zu handhabende Annäherung an den Exakten Test von Fischer, der in der Statistik als das geeignete Maß der Signifikanz von Assoziiertheit auch für asymmetrische Kontingenztabellen gilt, der aber wegen seines beträchtlichen Rechenaufwands meist keine Anwendung findet.[8]
Betrachtet man die für diese Arbeit gesichtete Literatur zu Kollokationen, so bietet sich ein differenzierteres Bild. Bartsch (2004) zieht zur Beurteilung möglicher Adverb-Verb Kollokationen die Werte von Mutual Information, t-score und χ^2 zu Rate; Klotz (2000) bildet Listen mit signifikanten Kollokatoren zu Substantiven des öffentlichen Transports, deren Ranking nach Mutual Information oder t-score erfolgt; auch Sardinha (1999) bedient sich, um zu einer Unterscheidung zwischen Kookkurrenz und Kollokationen zu kommen, dieser beiden Werte. Die log-likelihood spielte in den ersten Tagen der automatisierten Kollokationsextraktion noch keine Rolle, sondern wurde erst 1993 von Dunning eingeführt. Verwendet wird die log-likelihood bei Zinsmeister/ Heid (2002) bei der Extraktion von Substantiv-Verb Paaren aus einem mit einer statistischen Grammatik geparsten Text sowie von Seretan/Nerima/Wehrli (2004) zur Bewertung verschiedener Kookkurrenzdaten.
Die Güte der Ergebnisse der statistischen Tests scheint auch in Abhängigkeit zu der zu extrahierenden Kollokationsart zu stehen. Eine Bewertung der statistischen Tests gegenüber einer manuellen Kollokationsidentifikation nehmen z.B. Evert/Krenn (2001) und Krenn/ Evert (2001) vor. Sie folgen bei der Bestimmung von Substantiv-Verb Kollokationen den von Krenn (2000) vorgeschlagenen Kriterien und teilen eine n-Besten Liste in wahre und falsche Positive, um Precision und Recall zu berechnen. Bei der Bewertung der Kombinationen in den Rankinglisten muss eine Kombination drei Kriterien erfüllen um als Positive zu gelten: ihre Bestandteile müssen in einer grammatischen Relation zueinander stehen und in Kombination ein Funktionsverbgefüge oder einen figurativen Ausdruck bilden. Doch handelt es sich bei den nach diesen Kriterien ermittelten Positiven eigentlich nur um einen Teil der als Kollokationen zu bezeichnenden Kombinationen. Kollokationen wie *Zähne putzen*, *Hände waschen* und *Fahrrad fahren* sind weder Funktionsverbgefüge noch figurative Ausdrücke. Was mit ihnen geschieht wird nicht näher erläutert. Die Daten beziehen sich ausschließlich auf Präposition-Nomen-Verb (PNV) Tripel.

8 Angewandt wurde der Exakte Test von Fischer beispielsweise von Jones/Sinclair (1973), die noch sehr kleine Corpora bearbeiteten (150.000 Wörter).

Zwei der Diagramme, die Precision und Recall der verschiedenen Assoziationsmaße im Vergleich zur manuellen Evaluierung der Substantiv-Verb Kookkurrenzen zeigen, sind unten dargestellt. Wie in den beiden Schaubildern zu sehen ist, schneidet für Substantiv-Verb Paare der t-score am besten ab, er wird von den obigen Autoren zur Extraktion von Substantiv-Verb Kollokationen empfohlen.

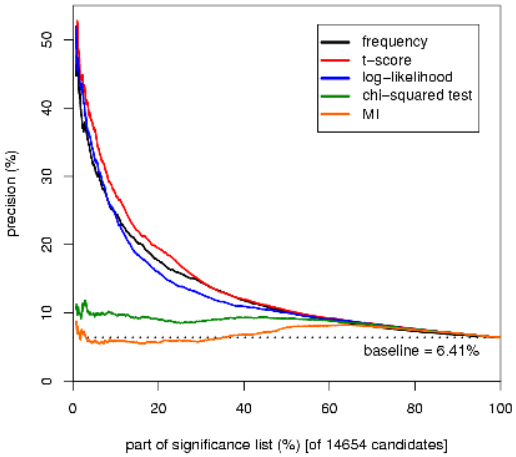


Abb. 1: Precision-Kurven für PNV-Daten (Evert/Krenn 2001: 191)

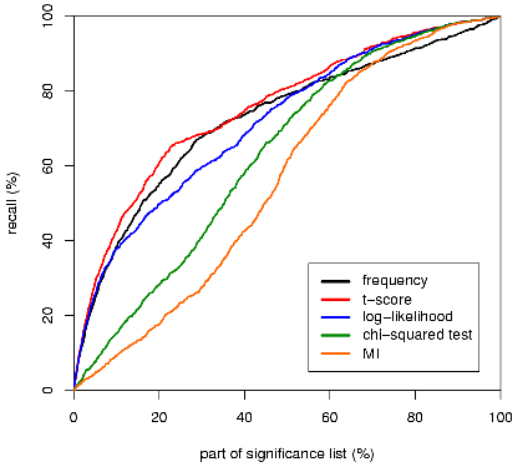


Abb. 2: Recall-Kurven für PNV-Daten(Evert/Krenn 2001: 191)

Betrachtet man hingegen eine Evaluierung zu Adjektiv-Nomen Kookkurrenzen, überrundet der χ^2 die anderen Assoziationsmaße bei weitem, er ist das zu präferierende Maß für diese Kollokationsart (Evert 2005a:134).
In PECCI wird als Default-Einstellung der t-Test und die Mutual Information gewählt, da sich der Lexikograf bei der manuellen Durchsicht der Ergebnisse meiner Meinung nach anhand der Kombination dieser beiden Assoziationsmaße am besten einen Einblick in das Kollokationsverhalten verschaffen kann (vgl. Kapitel 6). Es können je nach Präferenz auch die beiden anderen Assoziationsmaße log-likelihood und χ^2 in einer beliebigen Mischung ausgewählt werden, wünscht man ein anderes Bild. Die Information, die mit nur einem der statistischen Maße vermittelt werden kann, wird nicht als ausreichend empfunden. Auch von der Darstellung der Werte der vier Tests, wie oben, wird abgesehen und statt dessen in übersichtlicher Form weitere Daten angegeben zu Suchraumeinstellung, Okkurrenz des Verbs und des Nomens im Corpus sowie im Falle des Verbs innerhalb des untersuchten Samples (vgl. Anhang: B1 Programmdokumentation, C1, C2 AusgabeNomina.).

2.2. Linguistische Corpusaufbereitung

Die beschriebenen statistischen Tests zur Berechnung der Assoziationsmaße beziehen ihre Daten aus den ermittelten Zahlen zu Okkurrenzen und Kookkurrenzen der untersuchten Wörter sowie der Corpusgröße. Dabei wird unter Kookkurrenz nicht nur das Vorkommen der Wörter innerhalb eines aus einer bestimmten Anzahl von Wörtern bestehenden Fensters verstanden. Normalerweise wird vorausgesetzt, dass die beiden Ausdrücke innerhalb eines Satzes stehen, auch wenn dabei, vor allem über den pronominalen Gebrauch, ein satzübergreifender Bezug verloren geht (dieser ist aber auch satzintern nur über aufwendige Parsingverfahren nachzuvollziehen). Das Corpus kann unterschiedliche Stadien der linguistischen

Präprozessierung durchlaufen, und je nach der Art seiner Aufbereitung ergeben sich zur Kollokationsextraktion spezifische Anfragemöglichkeiten.
Die Corpora, die als Grundlage der Kollokationsextraktion dienen, haben in der Regel zunächst eine linguistische Vorverarbeitung erfahren, in der die Satzgrenzenerkennung und Tokenisierung stattfindet. Metadaten wie Autorennamen, Themengebiete, regionale und zeitliche Angaben werden mit XML- oder SGML-Tags integriert und sind so über Auszeichnungskonventionen zu identifizieren. Auch die Textstruktur (Überschriften, Absätze, Kapitel, usw.) wird mit Hilfe spezifischer Tags gekennzeichnet. Wie ein Corpus nach der linguistischen Vorverarbeitung und mit Metadaten aussieht wird in Kapitel 5.1 anhand der untersuchten portugiesischen Corpora verdeutlicht. Es handelt sich hierbei um Corpora, die ausschließlich aus Zeitungstexten bestehen. Grundsätzlich kann sich ein Corpus aus verschiedenen Quellen zusammensetzen: transkribierte gesprochene Sprache, Belletristik, wissenschaftliche Literatur und Gebrauchstexte können ebenso enthalten sein wie journalistische Texte. Soll das Corpus als Repräsentation einer Sprache dienen, gilt es, eine möglichst große Diversität zu erzielen.[9] Die Corpuszusammensetzung sollte immer separat beschrieben werden und die Zuordnung der Exzerptionsdaten zum entsprechenden Corpusteil und zur genauen Fundstelle nachvollziehbar bleiben.
Darauf folgt die linguistische Annotation. Mit Hilfe eines Part-of-Speech Taggers werden die Wörter des Corpus mit POS-Tags versehen. Die Klassifikation des Tagsets umfasst Informationen zu Wortart und Funktion eine Wortes im Satz, z.B. VAINF für ein Auxiliar im Infinitiv. Das Lemma der Wortform kann während des Lexikonvergleichs des Taggers zusätzlich notiert werden, auch präzise morphosyntaktische Angaben können hinzugefügt werden. Ein Problem für Tagger sind die Homographen, zu deren Disambiguierung sie stochastische oder regelbasierte Methoden oder Mischformen aus beiden einsetzen. Einen Überblick über verschiedene Part-of-Speech Tagger, deren Tagsets und Arbeitsweisen für das Englische findet man in Jurafsky (2000, Kapitel 8): "Word Classes and Part-of-Speech Tagging". Ein Tagger, der sprachübergreifend mit verschiedenen Tagsets arbeitet, ist der TreeTagger[10], speziell für das Deutsche wurde das Stuttgart-Tübinger Tagset (STTS)[11] entwickelt.
Die POS-Tags sind ihrerseits nur ein Schritt der Präprozessierung auf dem Weg zu einer syntaktischen Annotation. Diese kann in Form des Chunking[12] (auch als Partial Parsing oder Shallow Parsing bezeichnet) oder eines voll ausgebildeten Grammatikmodells erfolgen. Beim Chunking werden die Wörter gemäß ihrer Wortart und Satzstellung mit einer Grammatik aus regulären Ausdrücken zu größeren Konstituenten zusammengefasst. Strukturen wie Nominal-, Adjektiv- oder Verbalgruppen werden identifiziert und die entsprechenden Chunks mit ihrer syntaktischen Kategorie ausgezeichnet (<nx> </nx>, <ax> </ax>, <vx> </vx>). Über rekursive Regeln können die einzelnen Chunks in größere Einheiten integriert werden (Chunk linking), z.B. eine Nominal- in eine Präpositionalgruppe, bis man zur Satzebene gelangt. Eine Auszeichnung mit präzisen grammatischen Funktionen wird aus Mangel an Valenz- oder Subkategorisierunsinformationen jedoch unterbleiben.

9 Genaue Kriterien für den Corpusaufbau kann man z.B. Biber/Conrad/Reppen (1998) entnehmen.
10 http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
11 http://www.sfs.uni-tuebingen.de/Elwis/stts/stts.html
12 In "Parsing by Chunks" (1991) beschreibt Steven Abney das Standardverfahren. Einen frei verfügbaren Chunker für das Deutsche kann man unter http://www.ims.uni-stuttgart.de/tcl/SOFTWARE/German-Chunker.html finden.

Eine vollständige syntaktische Anlayse, die für einen Satz einen Syntaxbaum erzeugt, ist nur über komplizierte Grammatikformalismen zu erreichen. Die so annotierten Corpora werden als Baumbanken bezeichnet. Da es bislang nicht möglich ist jedem Satz einen eindeutigen Strukturbaum zuzuweisen, werden die vom Parser erhaltenen Möglichkeiten manuell disambiguiert. Die Penn Treebank[13], deren Entstehung auf das Jahr 1989 datiert, ist das erste Baumbanken Projekt, die Parsingalgorithmen und die Syntaxpräsentation orientieren sich an der Phrasenstrukturgrammatik. Heute umfasst sie über 2 Millionen Wörter. Das TIGER-Corpus[14] ist das entsprechende deutsche Konzept, geparst wird hier mit der Lexikalisch Funktionalen Grammatik. Durch die Satzstellung des Deutschen bedingt, erfolgt die Darstellung anhand eines Grafen mit kreuzenden Kanten und keiner einfachen Baumstruktur. Das TIGER-Corpus umfasst 1 Millionen Wörter.
Ausschnitte aus einem "normal" annotierten Corpus (worunter hier ein Corpus verstanden wird, der mit POS-Tags, Lemmaangaben und morphosyntaktischen Informationen versehen ist) und einer Baumbank zeigt Kapitel 4.2, wo die computationell verfügbaren linguistischen Ressourcen des Portugiesischen erläutert sind. Eine auf die syntaktischen Angaben aufbauende semantische Aufbereitung der Corpora auf der Grundlage bestehender lexikalischer Wissensnetze wie FrameNet oder WordNet befindet sich noch in den Kinderschuhen. Das SALSA-Projekt[15] ist für das Deutsche ein aktuelles Vorhaben mit dem Ziel, den TIGER-Corpus nach dem Konzept von FrameNet mit semantischen Relationen zu versehen.
Eine Kollokationssuche auf vollständig geparste Texte findet in der Regel nicht statt, da diese zur Kollokationsextraktion keine geeignete Größe haben. Die "geeignete" Corpusgröße hängt natürlich stark von der Aufgabenstellung ab. Interessiert man sich für das Verhalten zweier sehr frequenter Lexeme, genügt möglicherweise schon ein Corpus von hunderttausend Wörtern. Dies mag auch der Fall sein, bezieht man Funktionswörter in die Kollokationsdaten mit ein. Sind jedoch alle möglichen Kollokate zu einer bestimmten Basis im Hausmannschen Sinne gesucht, und übernimmt die Kollokationsextraktion primär die Funktion, den Lexikografen mit numerisch möglichst gut vorsortieren Kollokationskandidaten zu versorgen, sollte das Corpus mindestens 100 Millionen Wörter umfassen.

2.3. Akquisitionsmethoden

Liegt ein mit Lemmainformationen und POS-Tags annotiertes Corpus vor, so bringt dies den Vorteil, dass zum einen eine komplizierte Mustersuche der morphologisch verschieden-en Formen eines Lexems durch die Angabe des Lemmas ersetzt werden kann. Zum anderen besteht die Möglichkeit, die gesamten gesuchten Kollokate einer bestimmten Wortart zu einer Basis zu extrahieren (z.B. alle Verben zu einem bestimmten Nomen) oder ganze Kollokationsparadigmen auszusortieren (z.B. V+N). Wie in Kapitel 1 beschrieben, wurde die Ansicht des Britischen Kontextualismus einer gegenüber der Grammatik autonomen Lexik zugunsten eines Kollokationsbegriffs aufgegeben, der auf Wortarten und somit auch auf grammatischen Relationen basiert. Die Suche nach Wörtern in bestimmten Sequenzen passiert mittels einer Anfragesprache, die reguläre Ausdrücke integriert. Mit dem Suchbegriff '[word="acalentar" & temcagr="INF"] []{0,3} [pos="N.*"]' (in CQP kodiert, vgl. Kapitel 2.3.3) wird z.B. das Verb *acalentar* im Infinitiv ein bis drei Wörter links von einem beliebigen Nomen stehend extrahiert und als typische Verb-Objekt Stellung interpretiert.

13 http://www.cis.upenn.edu/~treebank/
14 http://www.ims.uni-stuttgart.de/projekte/TIGER/
15 http://www.coli.uni-saarland.de/projects/salsa/

Die Extraktion der Kollokationsparnter muss durch den Mangel an syntaktischen Strukturangaben, die zeigen, auf welche mögliche Basis sich ein bestimmter Kollokator bezieht, innerhalb eines gewissen Suchraums (auch Kollokationsspanne oder Fenster genannt) geschehen. Mit der bloßen Einschränkung auf ein satzinternes Erscheinen beider Wörter würde die Fehlerquote stark steigen, da es bei zunehmender Satzgröße auch eine zunehmende Anzahl anderer möglicher Kollokationspartner der gesuchten Wortart gibt. Die Größe des Suchraums hängt von der gesuchten Kollokationsart, der untersuchten Sprache und von der gewünschten Qualität der Ergebnisse ab - ist der Recall wichtiger als die Precision, wird man ein größeres Fenster, mitunter den ganzen Satz wählen. In der Regel wird eine Kollokationsspanne von ±5 bis ±3 Wörtern um die Basis herum selegiert. Die Auswirkungen der Ausweitung der Kollokationsspanne von 4 auf 10 Wörter, zeigt Klotz in einem ausführlichen Vergleich, in dem das Verhältnis gewertet wird von nicht gefundenen und falsch extrahierten Substantiv-Verb Kollokationen, die nicht in der gewünschten grammatischen Beziehung vorliegen (Klotz 2000: 76-84).
Behandelt man die Kollokationsextraktion ungeachtet der Wortartinformation und der damit auch immer gegebenen grammatischen Relation, spricht man von positionellen Verfahren (Evert 2005a: 57). Diese können je nach Aufgabenstellung durchaus sinnreich sein, interessiert man sich für das gesamte kombinatorische Verhalten eines Wortes in einem rein statistischen oder "pattern and priming" Modell. Die auf grammatischen Relationen basierenden Verfahren heißen entsprechend relationale Verfahren. Mit der Einschränkung der gesuchten Lexeme auf bestimmte Wortarten innerhalb eines Suchraums erhält man auch für einen Corpus ohne linguistische Annotation ein relationales Verfahren, präziser werden die Ergebnisse, wenn die syntaktischen Strukturen vereinfacht über die Satzstellung simulierbar sind.
In Abhängigkeit zur Corpusaufbereitung ergeben sich bei der Kollokationsextraktion verschiedene Anfragemöglichkeiten, deren Komplexität mit der Annotationsstufe zunimmt:

1. linguistische Vorverarbeitung + Tokenisierung/Satzgrenzen/Textstruktur/Metadaten
2. "normale" Annotation + Lemma/POS/morphosyntaktische Information
3. syntaktische Annotation + syntaktische Kategorie/grammatische Funktion

Für die Extraktion aus Corpora mit Chunking- oder Parsinginformationen stehen meist spezielle Anfragetools zur Verfügung, deren Suchfunktionen über eine grafische Benutzeroberfläche zu steuern sind, denn die Formulierung der Suche in einer Anfragesprache ist aufgrund der umfangreichen Annotation schon recht kompliziert.
Eine syntaktische Annotation erlaubt die Aufhebung des Suchraumkriteriums bei angemessenen Precision-Ergebnissen, da nun über die Beziehung zweier Wörter satzintern entschieden werden kann. Entgegen der ursprünglichen Annahme des Britischen Kontextualismus zeigt sich in einem Modell, das auf grammatischen Relationen basiert, dass Kollokationen nicht unbedingt ein lokales Phänomen innerhalb einer Kollokationsspanne sind (vgl. Kapitel 5.3.2). Zwar findet eine enge semantische Beziehung zwischen Wörtern oft Ausdruck in einer adjazenten Stellung oder einer positionellen Nähe der Wörter im Satz, doch können Kollokationen auch weit auseinander stehend auftreten: "*16106698: A* ***<esperança>*** *de um Estado palestiniano independente que Yitzhak Rabin e Shimon Peres* ***alimentaram*** *e que Netanyahu quase* ***enterrou*** *ao congelar indefinidamente a aplicação dos acordos de Oslo* ." (Cetempúblico*)* ('Hoffnung - nähren - begraben'). Trotzdem kann man die Fenstergröße nicht beliebig erhöhen, wie folgendes Beispiel zeigt, in dem sich *alimentar* auf

emoção ('Gefühl') bezieht und *esperança* ('Hoffnung') die Funktion des Prädikativum zusammen mit *é* ('sein') erfüllt: "*115039354: Alan Prost é o patrão , Panis* ***é*** *a* ***<esperança>*** *na pista e o motor* ***alimentará*** *a* ***emoção*** *com o ruído Peugeot ."* (Cetempúblico). Eine kleine Fenstergröße hingegen wirkt sich immer negativ auf die Recall-Ergebnisse aus.

2.3.1. Smadjas *Xtract*

Eines der frühen und gut dokumentierten Kollokationsextraktionstools ist das 1993 von Frank Smadja beschriebene *Xtract*, das mit einem 10 Millionen Wörter großen Corpus aus Aktienmarktberichten arbeitet. Ausgegangen wird von einer Kollokationsdefinition, die Kollokationen als arbiträr und rekurrent subsumiert, was als ersten Schritt das Auffinden von Wortpaaren impliziert, die häufiger als erwartet im Text erscheinen. Eine Kollokationsspanne von ±5 Wörtern wird satzintern um ein bestimmtes Wort festgelegt und alle Kollokate mit ihrer Position zum untersuchten Wort, ihren POS-Tags, den positionellen Frequenzen und der Gesamtzahl der Okkurrenzen gespeichert.
Im Gegensatz zu den in Kapitel 2.1 beschriebenen statistischen Assoziationsmaßen, bei deren Berechnung die vorkommenden Kookkurrenzen über einen angegebenen Suchraum generalisieren, kommt hier zusätzlich eine positionelle Berechnung zum Tragen. Zu der Annahme, dass sich Kollokationen signifikant häufiger als erwartet zeigen, tritt die These, dass sie durch die syntaktischen Relationen bedingt in einer relativ rigiden Wortstellung stehen. Als Kollokationen werden hier die Gipfel in den Histogrammen der positionellen Verteilung betrachtet. Kollokationales Verhalten beinhaltet die Bildung einer ungleichen Verteilung des betreffenden Kollokats in den verschiedenen Positionen. Über die Varianz s^2 wird üblicherweise die Sampleabweichung mit $s = \sqrt{s^2}$ berechnet. Erläutert wird das Standardverfahren von Manning und Schütze (2001: 157-162). Die Varianz misst wie groß die Differenz der Position d_i der einzelnen Kookkurrenzen (1 bis *n*) vom Mittelwert $\bar{d}$ ist:

$$s^2 = \frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n - 1}$$

Smadja bezeichnet die Varianz als *spread*, und nimmt eine hohe Zahl als Indiz dafür, dass die Kookkurrenzverteilung mindestens einen Gipfel aufweist. Ist dieses Maß entsprechend hoch, erfüllt die Kookkurrenz schon eine der Bedingungen, um als Kollokation zu gelten. Begründet wird dies mit der Annahme " ... that, if the two words are repeatedly used together within a single syntactic construct, then they will have a marked pattern of co-appearence, i.e., they will not appear in all the possible positions with an equal probability" (Smadja 1993: 156). Dementsprechend kann es passieren, dass man in der Ausgabe der Kollokationen dieselbe Kombination mehrmals vorfindet, mit dem Kollokat in verschiedenen Positionen. Die Ausgabe der Kollokationen erfolgt getrennt nach den Wortarten der Kollokate, die anhand der POS-Tags ermittelt werden. Mittels des positionellen Verfahrens sollen Kombinationen wie *telephone-television, bomb-soldier, trouble-problem* aussortiert werden, die den oben beschriebenen Kollokationsreihen entsprechen, und deren Kookkurrenz nicht auf einer strukturellen Konsistenz, sondern der Zuordnung zum gleichen Kontext basiert.
Die Kollokationsidentifikation soll rein automatisch erfolgen. Neben der numerisch festgelegten Signifikanzschwelle des positionellen Verfahrens muss eine Kookkurrenz auch

eine Grenze für ein anderes statistisches Maß überschreiten. Mit dem ermittelten Wert des z-scores (ein dem in Kapitel 2.1 vorgestellten t-score ähnliches Assoziationsmaß), der *strength*, soll sichergestellt werden, dass die Kombinationen rekurrent im Sinne von signifikant häufig sind. Eine dritte Bedingung wählt aus den möglichen Positionen einer Kombination die relevante(n) aus, die in der Kollokationstabelle erscheinen.
Die Bedingungen, die an die Kookkurrenz gestellt werden, um als Kollokation zu gelten, werden als ein Tupel von drei Gleichungen dargestellt. Umso tiefer für die drei zu überschreitenden Werte die jeweilig Signifikanzgrenze gesetzt wird, umso mehr Kollokationen werden aufgenommen, was wiederum zu Lasten der Precision geht. Setzt man den Schwellenwert der s*trength*, die minimale Standardabweichung auf 1, werden z.B. für das Wort *takeover* 95% möglicher Kollokate zurückgewiesen. Für die aufgenommenen Kollokationen werden in einem zweiten Schritt die Konkordanzen erzeugt und " ... n-word collocations from two-word associations ... " (Smadja 1993: 160) ermittelt. Die Wörter im Umfeld einer Kollokation werden in die Berechnung mit einbezogen und aus Eingabedaten wie *average-industrial* die feste Verbindung *the Dow Jones industrial average* als einzige Vorkommensumgebung gewonnen. In einem dritten Schritt werden die Konkordanzen der Kollokationen mit syntaktischen binären Informationen anhand der POS-Tags versehen.[16] Auf diese Weise werden Verb-Objekt, Verb-Subjekt, Nomen-Adjektiv und Nomen-Nomen Kollokationen gefunden. Eine Kollokation wird nur akzeptiert, wenn sie mit einer der syntaktischen Strukturen kongruiert. Untersucht werden nur lexikalische Kollokationen, eine Ausdehnung auf grammatische Kollokationen ist mit *Xtract* jederzeit möglich.
Zur Evaluierung des Systems wurden aus den durch die ersten beiden Schritten bestimmten Kollokationen 4000 zufällig ausgewählt, die einerseits von Schritt drei weiterverarbeitet werden, und andererseits von einem Lexikografen in Bezug auf ihre Gültigkeit bewertet werden. Das Kriterium für die menschliche Introspektion ist die Aufnahme in ein domänen spezifisches Wörterbuch, wobei zwischen Kollokationen unterschieden wird, die zwar gültig, aber zu spezifisch für die Aufnahme oder nicht vollständig sind, und den Kollokationen, die ohne Einschränkung aufgenommen werden. Auf diese beiden Kategorien entfallen jeweils 20%, das heißt, dass 60% der nach Schritt zwei vorhandenen Kollokationen keine sind, die Precision beträgt somit 40%.
Ein Vergleich mit einer rein manuell vorgenommen Extraktion neuer und interessanter Ausdrücke aus Dokumenten zum Wörterbuch-Update zeigt, dass die Rate der dort vorgeschlagenen Ausdrücken zu den tatsächlich aufgenommenen bei nur 4% liegt. Hinzu kommt, dass das manuelle Filtern vorgeschlagener Kollokationen schneller verläuft, als die manuelle Durchsicht von Texten. Als größten Mangel der maschinellen Extraktion sieht Smadja die Nicht-Identifizierung von niedrigfrequenten Wörtern, die von *Xtract* nicht als Kollokationen erfasst werden. Zu deren Auffinden in Texten ist auch weiterhin ein humaner Leser von Nutzen.
Weitaus bessere Ergebnisse werden erzielt, haben die Kollokationen in *Xtract* auch den dritten Schritt durchlaufen. Hier werden 60% der Kombinationen aus Schritt zwei abgelehnt und 40% mit einem syntaktischen Label versehen. Diese Kollokationen überschneiden sich zu 94% mit den vom Lexikografen bestimmten Kollokationen, woraus Smadja auf einen Recall von 94% schließt. Überprüft der Lexikograf die gültigen 40% der Kombinationen aus

16 Die syntaktische Annotation erfolgt mit dem robusten Parser "Cass", der von Abney entwickelt wurde und dessen Arbeitsweise auch in dem 1991 erschienenen Artikel "Parsing by Chunks" beschrieben wird.

Schritt drei, so akzeptiert er davon 80% als tatsächliche Kollokationen, was eine Verdopplung der Precision-Ergebnisse von Schritt zwei zu Schritt drei bringt. Das Recall-Ergebnis von 94% ist sicherlich erheblich zu mindern, schließt man in den Vergleich nicht identifizierte Kollokationen aus dem gesamten Textbereich mit ein.
In dem Kollokationsextraktionstool *Xtract* werden somit verschiedene statistische und linguistische Ansätze vereint. Die Extraktion erfolgt zunächst unabhängig von POS-Tags und syntaktischen Strukturen über die Varianz der einzelnen Positionen der Kollokate und deren *strenght*. Linguistische Informationen fließen erst in die gefilterte Ausgabe mit ein, die nach bestimmten POS-Tags sortiert erfolgen kann. Erzielen die so gewonnenen einzelnen Bigramme in einem zweiten Schritt gleiche Umgebungsdaten werden aus ihnen N-Gramme maximaler Länge. Feststehende Wendungen wie *The NYSE´s composite index of all its listed common stocks* (Smadja 1993: 159) werden unter dem Namen Kollokationen extrahiert. Weitere linguistische Informationen kommen auch im dritten Schritt zum Tragen, nach dem nur noch Kombinationen mit bestimmten syntaktischen Strukturen als Kollokationen erscheinen.

2.3.2. Seretan/Nerima/Wehrli

Einige Jahre später ist eine differenziertere und modularisiertere Auffassung der corpusbasierten lexikalischen Akquisition von Kollokationen anzutreffen, bei der positionelle Verfahren in der Regel keine Rolle mehr spielen und linguistische Informationen nicht nur im Schritt der Corpusaufbereitung sondern auch in dessen Weiterverarbeitung in prominenter Stelle einfließen, womit eine Extraktion überflüssiger Daten vermieden wird. Auch wird die Definition von Kollokationen restriktiver gezogen, wodurch Wendungen mit mehreren Lexemen zu einem gesonderten Paradigma zählen. Diese werden komplementär zu binären Kollokationen behandelt, wobei der Begriff Kollokation heute üblicherweise nur noch bei Bi- und Trigrammen Verwendung findet. Dabei gibt es verschiedene Vorgehensweisen.
Übereinstimmung findet man im Grad der linguistischen Corpusaufbereitung und der Filterung von Kollokationskandidaten mit syntaktischen Kriterien. Anders als bei *Xtract* werden die syntaktischen Informationen aber schon im ersten Schritt der Kollokationsextraktion genutzt. Seretan, Nerima und Wehrli stellen in ihrem Artikel "A Tool for Muli-Word Collocation Extraction and Visualization in Multilingual Corpora" (2004) ein System vor, das nur syntaktisch wohlgeformte Kollokationskandidaten unab-hängig von ihrer relativen Position, Entfernung oder Morphologie extrahiert. Möglich ist dies durch robuste Parser, die auf großen Textmengen laufen und auch weit entfernte Abhängigkeiten erkennen.
Die Kriterien der syntaktischen Wohlgeformtheit werden für die Extraktion der Zweiwort Kollokationen festgelegt, aus deren Resultaten Trigramme gewonnen werden, deren drittes Wort einer beliebigen Wortart angehört. Dadurch können syntaktische Präferenzen von Trigrammen erkannt und nach ihrer Frequenz aufgelistet werden (vgl. Seretan/ Nerima/Wehrli 2003: 430). Rein frequenzorientierte Ergebnisse sind für den englischen Text: "weapon of mass destruction, have impact on, got out of, pull out of, make difference to, ... " (Seretan/Nerima/Wehrli 2004: 765). Für die binären Kollokationen werden über die Konkordanzen und Alignement Verfahren die Textstellen mit den betreffenden Kollokationen auch im Parallelcorpus der anderen Sprache automatisch angezeigt.

2.3.3. Heid et al.

Eine weitere Möglichkeit der Tripel-Kollokationsextraktion zeigen Zinsmeister und Heid (2003): "Significant Triples: Adjective+Noun+Verb Combinations". Hier wird die Kollokationsdefinition im Sinne von Hausmann auf offene Wortklassen festgelegt, wodurch ein Großteil der bei Seretan et al. extrahierten Kombinationen ihre Gültigkeit verlören. Als Reaktion auf diesen Artikel erweiterte Hausmann seinen binären Kollokationsbegriff und schließt Tripel-Kollokationen als Verbindung von zwei Kollokationen mit ein: "dringlichen Appell richten (an), konkrete Hilfe leisten, reißenden Absatz finden" werden von ihm als Beispiele genannt (Hausmann 2004: 316).
Zinsmeister und Heid schlüsseln die Adjektiv-Nomen-Verb Kombinationen, in denen das Substantiv als Akkusativobjekt fungiert, in fünf Untergruppen auf. Die erste Gruppe umfasst idiomatische Phrasen, in denen A+N+V lexikalisch fixiert erscheinen (*sich einen schönen Lenz machen*), in der zweiten Gruppe sind A+N lexikalisch fixiert und V kompositionell (*schwarze Zahlen schreiben*), in der dritten Gruppe ist A kompositionell und V+N lexikalisch fixiert (*einen neuen Haftbefehl erlassen*), die vierte Gruppe besteht aus einer Kombination von zwei Kollokationen, die die gleiche Basis teilen, N+V und A+N sind jeweils lexikalisch fixiert (*ein biblisches Alter erreichen*), in der fünften Gruppe finden sich die trivialen Kombinationen (*neue Politik fordern*) (Zinsmeister/Heid 2003: 93-94).
Die Extraktion der Tripel-Kollokationen ist durch die relativ freie Wortstellung des Deutschen nur sinnreich mit einem vollständig geparsten Text. Das Parsing erfolgt hier mit einer lexikalisierten probabilistischen Grammatik (siehe unten), die die geschätzten Häufigkeiten von Wortpaaren innerhalb ihres Trainings gleich mit aufbereitet. Aus einem Corpus mit 5 Millionen Sätzen wurden über 440.000 A+N+V Kombinationen extrahiert, von denen ca. 70% Hapax Legomena sind. Für die statistische Weiterverarbeitung sind auch die Substantiv-Verb Kombinationen relevant, die nicht von einem Adjektiv modifiziert werden und deren Vorkommen bei über 1,2 Millionen liegt, hier wird das Adjektiv über das Merkmal 'NoAdj' simuliert. Die beobachteten Frequenzen werden den erwarteten Frequenzen gegenübergestellt und die log-likelihood für jedes Tripel viermal berechnet - für jede der möglichen Zweierkombinationen (A+N, N+V, A+V) und für das Tripel (A+N+V), zu dessen Berechnung auch die Kombinationen ohne Adjektivmodifikation zum Paradigma zählen.
Das Ziel der mehrfachen Berechnung sind nicht einfache Rankinglisten, die man auch erhält, sondern der Versuch, die Dreiwort Kombinationen über die verschiedenen möglichen Konstellationen der Höhe der log-likelihood automatisch einer der fünf beschriebenen Gruppen zuzuschreiben. Dies geschieht über einen Entscheidungsbaum, der dem maschinellen Lernverfahren C4.5 entspricht, und der manuell klassifizierte Tripel als Grundlage erhält. Ist zum Beispiel die log-likelihood des Tripels hoch, die log-likelihood der binären Kombinationen A+N und N+V hingegen niedrig handelt es sich um ein Idiom und wird der ersten Gruppe zugeordnet (*offene Türen einrennen*). Ist die log-likelihood der drei Konstellationen niedrig, gilt das Tripel als triviale Kombination und gehört zur fünften Gruppe (*andere Probleme haben*).
Ziel der Arbeit ist es nicht, eine vollständige automatische Kollokationsklassifikation zu erreichen, vielmehr sollen die maschinell erzielten Ergebnisse einem Lexikografen als Grundlage dienen, um diese mit möglichst wenig Aufwand manuell auszuwerten.

Damit steht dieser Ansatz der Tripel-Kollokationsextraktion in einer Tradition der corpusbasierten lexikalischen Akquisition, die

1. Kollokationen auf lexikalische Wortklassen beschränkt,
2. Kollokationen anhand syntaktischer Strukturen bestimmt,
3. Kollokationsidentifikation mit statistischen Assoziationsmaßen als Rohmaterial für den Lexikografen benennt und
4. möglichst effiziente Methoden der Corpusaufbereitung und Extraktionstools bereitstellt.

Eine direkte Weiterverarbeitung des Kollokationsextraktionsmaterials zur Wortbedeutungsdisambiguierung, Sprachgenerierung oder Maschinellen Übersetzung, wie sie unter anderem in "Collocations" von McKeown und Radev (2000) beschrieben ist, wird hier nicht angestrebt, die Kollokationsextraktion konzentriert sich auf die Bedürfnisse des Lexikografen zum Wörterbuch-Update.
Die Notwendigkeit eines "menschlichen Filters" bei der Wörterbuchproduktion wird betont damit unter anderem corpusbedingte Besonderheiten in den Einträgen unterbleiben. Heid et al. (2000: 193) geben als Beispiel für das Nomen *deadline* das Erscheinen des Kollokators *15th* an dritter Stelle der Frequenzliste auf der CD-Rom mit den *Cobuild English Collocations*, was darauf zurückzuführen ist, dass ein großer Teil des Corpus aus der Zeit vor dem Ablaufen des Ultimatums (der *deadline*) am 15. Januar 1991 an Saddam Hussein stammt. Die Wörterbuchaufbereitung wird als ein letztendlich von der menschlichen Introspektion geleiteter Prozess aufgefasst, der durch das maschinell bereitgestellte Material erheblich vereinfacht werden kann.
Ein speziell auf diese Bedürfnisse zugeschnittenes Tool ist "LexiView"[17], dessen Architektur in Heid et al. "Tools for upgrading printed dictionaries by means of corpus-based lexical acquisition" (2004) beschrieben wird. Als Input werden die Corpusdaten (ca. 350 Millionen Wörter) und das zu bearbeitende Wörterbuch gewählt und zum Vergleich in einem internen XML-basierten Format abgelegt. Da die Wörterbücher in elektronischer Form in verschiedenen Formaten vorliegen, müssen diese individuell analysiert und konvertiert werden. Das gemeinsame Repräsentationsformat enthält alle relevanten Fakten zu einem gegebenen Wort, die maschinell aus Corpusdaten gewonnen werden können:

- Lemma und Wortklasse (zur Identifikation),
- Corpusfrequenzen eines Lemma-Wortklassen Paars,
- linguistische Eigenschaften eines Lemma-Wortklassen Paars wie syntaktische Subkategorisierung oder Morphosyntax und deren Frequenzen,
- Kollokationen und andere signifikante Wortpaare, ihre linguistischen Eigenschaften und Frequenzen.

Der Vergleich der beiden Quellen erfolgt mit einer interaktiven grafischen Oberfläche (GUI), in der die Wörter nach den vom Benutzer bestimmten Kriterien sortiert erscheinen. Für das markierte Wort werden die obigen Informationen in übersichtlicher Form gegeben, zusätzlich werden die Corpusbelege und, falls vorhanden, Kollokations- und Lexikonbelege gezeigt. Aus ihnen kann der Lexikograf beliebige Beispiele wählen. Die Entscheidung über Aufnahme oder Streichung eines Wortes aufgrund der Corpusdaten kann nicht automatisiert werden, denn sie basiert auf verschiedenen nicht zu formalisierenden komplexen Kriterien

17 Die Entwicklung von LexiView lief innerhalb des Projekts Transferbereichs 32 'Automatische Exzerption', einer Kooperation der Textcorpora- und Lexikongruppe des IMS mit dem Langenscheidt Verlag und Brockhaus. Projektziele, Systemarchitektur und computerlinguistische Module werden unter http://www.ims.uni-stuttgart.de/projekte/TFB/projekt.shtml näher erläutert.

wie der Größe des gedruckten Wörterbuches und der individuellen Zielgruppe. Die Kollokationsextraktion ist hier nur Teil eines Prozesses, der als Automatische Exzerption bezeichnet wird, und der vor allem auf die Bedürfnisse des Wörterbuchredakteurs zugeschnitten ist.
Ein grundlegender Teil diese Ansatzes ist die Corpusaufbereitung und die Speicherung bzw. Weiterverarbeitung der daraus resultierenden Daten. Während die Tools der linguistischen Präprozessierung sprachabhängig arbeiten, abstrahiert das interne Format von LexiView über sprachindividuelle Eigenschaften und kann daher für verschiedene Sprachen eingesetzt werden. Die gleichen linguistischen Tools der Corpusaufbereitung können auch mit dem alleinigen Ziel der Kollokationsextraktion verwendet werden, wobei auch hier eine möglichst präzise Corpusannotation angestrebt ist.
In den neuen Arbeiten zur semi-automatischen Kollokationsextraktion werden die deutschen Textdaten mittels stochastischem Corpus Parsing aufbereitet (Zinsmeister/Heid 2002, 2003, 2004). Besonders für Nomen-Verb Paare ist die Adjazenz im Deutschen nur selten gegeben, mit der Anwendung des statistischen Grammatikmodells werden nicht nur die Verb-argumente unabhängig von ihrer Linearisierung oder Aktiv-Passiv Alternierung erkannt, auch die weit entfernten Partikel eines Verbs werden identifiziert und entsprechend verarbeitet. Das probabilistische Grammatikmodell und seine Nutzung für die Extraktion lexikalischer Information wird in Schulte im Walde et al. (2001) vorgestellt. Es basiert auf einer manuell erstellten kontextfreien Grammatik mit Merkmalsstruktur Annotationen (wie zum Beispiel der Spezifikation von Subkategorisierungsrahmen) und lernt seine Regelwahr-scheinlichkeiten mit dem statistischen "LoPar"[18] Parser. Während des Trainings werden die Grammatikregeln mit Informationen über ihre lexikalischen Köpfe angereichert. Die Lexikalisierung erlaubt später anhand des trainierten Grammatikmodells, Kookkurrenzdaten lexikalischer Köpfe (sowie deren geschätzte Häufigkeiten) gemäß ihrer grammatischen Struktur direkt abzulesen und sie gegebenenfalls zur Berechnung statistischer Assoziations-maße an den nächsten Algorithmus weiterzureichen.
In LexiView wird die volle syntaktische Analyse nur für bestimmte Kollokationstypen (wie den Nomen-Verb Kombinationen) benutzt, deren Verteilung im Satz regelmäßig große Entfernungen überschreitet, ansonsten greift man auf die Daten der "normalen" Corpus-annotation zu. Neben der Tokenisierung, der Lemmatisierung und dem Part-of-Speech Tagging zählt hierzu auch das Chunking mit dem rekursiven partiellen Parser YAC, der versucht, möglichst große Phrasen zu bilden und wenn möglich Satzbausteine wie Subjekt oder Prädikat zu erkennen. Mit der Annotation der lexikalischen Köpfe, deren morphosyntaktischer Information und der lexikalisch-semantischen und strukturellen Eigenschaften für die einzelnen Chunks geht YAC über andere Chunking Verfahren hinaus.

Eine Verbesserung der Recall-Ergebnisse durch die Bereitstellung von Kookkurrenzdaten in Form von lexikalischen Köpfen verschiedener Chunks wird in dem Artikel von Evert und Kermes "Experiments on Candidate Data for Collocation Extraction" (2003) aufgezeigt. Wird ein perfektes Tagging als Grundlage genommen steigt der Recall für syntaktisch tatsächlich zusammengehörige Adjektiv-Nomen Kombinationen von 90.58% für adjazente Wortpaare (basierend auf POS-Tags) auf 96,74% für ein fensterbasiertes Verfahren (Fenster = 10 Wörter) und auf 97,94% mit einer Chunk-Annotation.

18 "LoPar" (**l**eft c**o**rner **par**ser for head-lexicalised probabilistic context-free grammars) wurde am Lehrstuhl für Theoretische Linguistik am IMS entwickelt und stehen unter http://www.ims.uni-stuttgart.de/tcl/SOFTWARE/LoPar.html für die nicht kommerzielle Nutzung zur freien Verfügung.

Die Arbeitsweise von YAC wird in dem Artikel von Kermes und Heid "Using chunked corpora for the acquisition of collocations and idiomatic expressions" (2003) näher erläutert, wo das eher selten untersuchte Kollokationspaar Adjektiv-Verb und dessen syntaktischer Kontext sowie Lexikalisierungsgrad von Interesse ist. YAC basiert auf einer symbolischen Grammatik regulärer Ausdrücke, die in der Anfragesprache für CQP kodiert ist. Der "**C**orpus **Q**uery **P**rocessor" ist ein spezielles Anfragemodul für die Corpora, die im Datenmodell der IMS Corpus Workbench[19] verwaltet sind.
Die Ablage der Corpora in einem einheitlichen Datenformat erlaubt den Gebrauch einer Anfragesprache, die mit den Informationen verschiedener Annotationsstufen arbeitet und auch Corpora bis zu 300 Millionen Wörtern in kurzer Zeit prozessierbar macht. Die Ergebnisse der Corpusaufbereitung werden in einer gemeinsamen Datenbank festgehalten, die sich an der Corpusposition eines Wortes orientiert. Die Anzahl der positionellen Attribute ist unbeschränkt: Lemmata, Wortklasseninformationen (auch aus mehreren Tagging-Ver-fahren) und Frequenzdaten sind zum Beispiel enthalten, neue Informationen wie aus dem Chunking können dem annotierten Corpus hinzugefügt werden, sie bestimmen eine Region im Satz und werden Positionen umschließend gespeichert. Bestandteil der Datenbank sind auch Frequenzangaben zu Kookkurrenzen oder morphosyntaktischer Distribution.

Bevor die Parsing-Verfahren für große deutschsprachige Corpora verfügbar waren, konnten mit CQP fein granulierte Kollokationspaare extrahiert werden (vgl. Heid 1998: "Towards a corpus-based dictionary of German noun-verb collocations"). Bei der Bereitstellung von Substantiv-Verb Kollokationen beinhaltet dies eine mögliche Separation gleicher Basis-Kollokator Paare nach dem Determinantentyp oder dem Numerus der Basis. Grammatische Relationen werden über die Wortstellung und Morphosyntax simuliert, eine Beschränkung auf einen Kollokationstyp bestehend aus den offenen Wortklassen ist auch hier schon impliziert. Das Kollokationsverständnis der heutigen Zeit ist weitgehend an das Vorkommen der Kollokation in einer bestimmten syntaktischen Struktur gebunden, die positionellen Verfahren aus der Anfangszeit werden bei der Kollokationsextraktion nur noch selten gefunden.

2.4. Automatische semantische Klassifikation von Kollokationen

Neben die generelle Unterteilung der Kollokationen nach der Wortartenzugehörigkeit ihrer Teilnehmer tritt bei den Substantiv-Verb Kollokationen eine weitere Differenzierung nach der grammatischen Funktion des Substantivs im Satz. Von Hausmann vorgesehen ist eine Zweiteilung in Subjekt-Verb und Objekt-Verb Kollokationen. Mit statistischen Assoziationsmaßen und einer linguistischen Corpusaufbereitung lassen sich die Kandidaten bestimmen, die als Kollokationen in Frage kommen. Eine syntaktisch präzisere Einteilung und die Beschreibung der Substantiv-Verb Kollokationen unter semantischen Kriterien bieten die lexikalischen Funktionen von Igor Mel'čuk. Die Granularität der lexikalischen Funktionen übersteigt die Möglichkeiten der systematischen Kollokationsbeschreibung der anderen Kollokationstheorien und bietet ein adäquates Kollokationsmodell in einer Metasprache.

19 Weitere Informationen unter http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/. CQP ist als Anfragesprache für Corpora aus mehreren Ländern (englische, dänische, italienische, katalanische und portugiesische), die im Format der Corpus Workbench vorliegen, unter verschiedenen Adressen im Internet zu finden (vgl. Kapitel 4.1). Über die Möglichkeiten von CQP informiert auch Kapitel 5.1.

2.4.1. Lexikalische Funktionen und Kollokationen

Die lexikalischen Funktionen (LF) wurden im Rahmen der *Meaning-Text Theory* von Mel'čuk in die Linguistik eingeführt.[20] Diese Theorie beschreibt natürliche Sprachen als eine Art logischen Apparat, der eine beliebige gegebene Bedeutung *M* mit der Menge der Texte T_i assoziiert, die diese Bedeutung adäquat wiedergeben. Eine Äußerung wird gleichzeitig auf 7 Ebenen repräsentiert: semantisch - tiefensyntaktisch - oberflächensyntaktisch- tiefenmorphologisch - oberflächenmorphologisch - tiefenphonetisch - oberflächenphonetisch.
Bei der Textproduktion hat der Sprecher oder Computer eine lexikalische Auswahl zu treffen, wenn er von der semantischen Repräsentation zu der entsprechenden tiefensyntaktischen Repräsentation wechselt. Werden die lexikalischen Einheiten in Abhängigkeit von anderen lexikalischen Einheiten selektiert, kann man sie mit lexikalischen Funktionen bezeichnen. Formal ist eine LF *f* eine Funktion im mathematischen Sinne (f(X) = Y), die für einen gegebenen lexikalischen Ausdruck *L* (das Argument oder *Keyword*) von *f*, ein Feld $\{L_i\}$ lexikalischer Ausdrücke - die Werte von *f* - ermittelt, die in Abhängigkeit von *L* eine spezifische Bedeutung ausdrücken, die *f* zugeordnet wird: $f(L) = \{L_i\}$. Eine LF ist eine spezielle Bedeutung (oder semantisch-syntaktische Rolle), deren Ausdruck abhängig ist von der lexikalischen Einheit, auf die sich die Bedeutung bezieht (Mel'čuk 1996: 40).
Das Modell der lexikalischen Funktionen bietet semantisch universale Relationen, deren Werte einzelsprachlich variieren. Das Lexikon nimmt bei der Beschreibung der lexikalischen Kombinierbarkeit eine zentrale Rolle ein. Der lexikografische Teil der *Meaning-Text Theory* wurde in in den *Dictionnaire explicatif et combinatoire du français contemporain: recherches lexico-sémantiques I-IV (DEC) (*1984, 1988, 1992, 1999) praktisch umgesetzt. Die Lexeme in den Wörterbüchern zum zeitgenössischen Französisch werden jeweils in fünf Zonen beschrieben: die Einleitung enthält das Lemma mit spezifischen morphologischen und syntaktischen Informationen, die semantische Zone die propositionale Form, Definitionen und Konnotationen, die Zone der syntaktischen Kombi-natorik Subkategorisierungs- und Restriktionseigenschaften, die lexikalische Kombinatorik wird mit LF und illustrativen Beispielen in Zone vier belegt, die fünfte Zone zeigt Phraseo-logismen. Im theoretischen Teil der explikativen und kombinatorischen Wörterbüchern wird die Repräsentation der Ebenen und deren Beziehung zur *Meaning-Text Theory* näher erläutert.
In dem in der Einleitung erwähnten Artikel von Mel'čuk und Wanner (1994)[21], der die verbalen Kollokate von Gefühlssubstantiven des Deutschen als Werte der LF repräsentiert, sind die Lexikoneinträge nach den Vorgaben im *DEC* konzipiert, enthalten aber nur einen Teil der dort gebotenen Angaben:

HOFFNUNG, fem
Hoffnung von X auf Y 'X's hope for Y' = X's pleasant, mental, permanent Gefühl [caused by X's belief and desire that Y takes place]

X = I	Y = II
1. $\mathbf{N}_{gen}$ **2.** ***von*** $\mathbf{N}_{dat}$ **3.** $\mathbf{Adj}_{poss}$	1. *auf* N_{acc} 2. *daß* PROP

20 Veröffentlicht wurde die *Meaning-Text Theory* 1974 von Igor Mel'čuk. Eine aktuelle Zusammenfassung gibt er 1997: *Vers une linguistique Sens-Texte.* Einen ausführlichen Überblick über lexikalische Funktionen zeigt Mel'čuk (1996).

21 "Lexical Co-occurrence and Lexical Inheritance. Emotion Lexemes in German: A Lexicographic Case Study". Die 40 untersuchten Gefühlssubstantive werden in Abbildung 9 aufgeführt.

IncepPredMinus : nachlassen
$Oper_1$: **empfinden, haben**, hegen [~$_{acc}$], ?fühlen
$IncepOper_1$: bekommen [~$_{acc}$]
fast $FinFunc_0$: verfliegen
$IncepFunc_1$: **aufkommen [in N_{dat}]**
$CausFunc_1$: einflößen, machen [N_{dat} ~$_{acc}$], wecken [in N_{dat} ~$_{acc}$]
(Mel'čuk/Wanner 1994: 114)

Die syntaktischen Realisierungsmöglichkeiten der semantischen Aktanten werden wie im *DEC* in einer Tabelle dargestellt. Die syntagmatischen Beziehungen zwischen Gefühlssubstantiven und Verben werden mit den LF gegliedert. Für die Beschreibung der Kollokationen sind nur die syntagmatischen LF relevant. Die paradigmatischen LF benennen Substitutions- oder Kontrastbeziehungen, in denen das *Keyword* zu anderen Lexemen steht. Dazu zählen zum einen semantische Beziehungen wie **Syn**(onymie), **Ant**(onymie) oder **Conv**(ersion) (**Ant**(*hope*) = *dispair*) und zum anderen syntaktische Derivation (Nominalisierung: $\mathbf{S_0}$(*to analyze*) = *analysis*) und semantische Derivation (Aktantenbelegung: $\mathbf{S_1}$(*letter*) = *author; sender*), die formal wiedergegeben werden.
Die syntagmatischen LF bezeichnen oder bestimmen im Satz die Werte der Kollokate in Abhängigkeit vom *Keyword*, sie beschreiben kombinatorische Beziehungen. Eine LF hat eine generelle, abstrakte Bedeutung verbunden mit einer tiefensyntaktischen Rolle, die lexikalisch in Abhängigkeit vom *Keyword* durch variierende Lexeme wiedergegeben wird. Verschieden Werte für eine LF von einem *Keyword* sind (quasi)synonym. Es sind ca. 60 einfache lexikalische Standardfunktionen bekannt (Mel'čuk 1996: 47-72), die auf 10 semantisch/syntaktische Gruppen verteilt sind, von denen folgende vier Substantiv-Verb Kollokationen beschreiben. Benannt werden die LF mit Abkürzungen lateinischer Wörter:

- PHASALS: verbs denoting the three phases of an event - the beginning (**Incep**), the end (**Fin**), and the continuation (**Cont**). These LFs are often used in combination with other verbal LFs.
- CAUSATIVES: verbs denoting the three possible types of causation: causation of existence (**Caus**), causation of non-existence (**Liqu**), and non-causation of non-existence (**Perm**). ...
- AUXILIARIES (= support, or 'light', verbs): these are semantically empty verbs linking a DSynt-actant [= A] of L to L; $\mathbf{Oper_{1,2}}$ takes L as its DSyntA **II**, $\mathbf{Func_{0,1,2}}$ as its DSyntA **I**, and $\mathbf{Labor_{12,21}}$ as its DSyntA **III**. ...
- REALIZATIONS: $\mathbf{Real_{1,2}}$, $\mathbf{Fact_{0,1,2}}$, $\mathbf{Labreal_{12,21}}$.
(Mel'čuk 1998: 35-36)

Die einfachen LF können miteinander zu komplexen LF kombinieren, was beispielhaft oben im Eintrag von 'Hoffnung' zu sehen ist. Die LF **Func**, **Oper** und **Labor** für die semantisch entleerten Funktionsverben erlauben die Identifizierung der grammatischen Funktion, die das *Keyword* einnimmt - respektive ist es Subjekt, erstes Komplement oder zweites Komplement. Die Indizes am Namen der LF spezifizieren wie das Argument der LF und die semantische Aktantenstruktur der LF mit der syntaktischen Struktur der verbalen Werte der LF korrespondieren. Das Schaubild auf der folgenden Seite verdeutlicht den Zusammenhang zwischen Funktionsverben und ihrer tiefensyntaktischen Beziehung zu den Argumenten. Die LF der *fulfilment Verbs* **Fact**, **Real** und **Labreal** funktionieren syntaktisch analog. Im Kontrast zu den Funktionsverben können nicht nur abstrakte, sondern auch konkrete Nomina als *Keyword* fungieren, die LF bedeuten "'[to] fulfil the requirement of L' [= '[to] do with L what you are supposed to do with L']" (Mel'čuk 1996: 68).

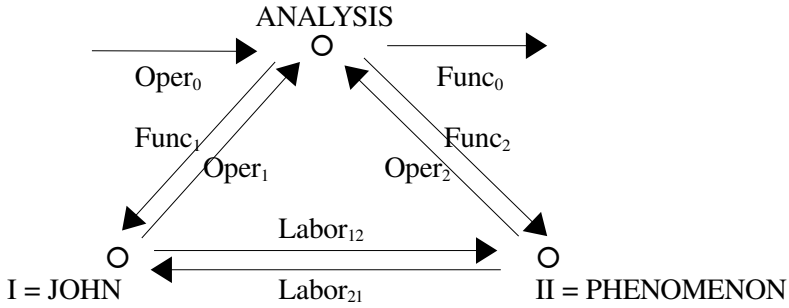


Abb. 3: Tiefensyntaktische Beziehung der Funktionsverben zu ihren Argumenten (Mel'čuk 1998: 39)

$Oper_1$(*analysis*) = *[to] carry out [John carries out the analysis of the phenomenon];*
$Oper_2$(*analysis*) = *[to] undergo [The phenomenon underwent (careful) analysis (by John)];*
$Func_1$(*analysis*) = *is due [The analysis of this phenomenon is due to John];*
$Func_2$(*analysis*) = *covers, concerns [John's analysis concerns this phenomenon];*
$Labor_{12}$(*analysis*) = *[to] submit [John submits this phenomenon to a (careful) analysis];*
$Labor_{21}$(*analysis*) = *leads [The phenomenon leads John to a (specific) analysis];*
$Func_0$(*analysis*) = *is in progress [John's analysis of the phenomenon is in progress];*
$Oper_0$(*analysis*) = *[one] sees [One sees an analysis of the phenomenon by John].*

In den folgenden Beispielen (Mel'čuk 1996, 1998) zeigen die Ausdrücke in den eckigen Klammern, die die Werte der LF begleiten, das reduzierte Subkategorisierungsmuster des lexikalischen Subeintrags:

$Oper_2$(*exam*)	= [*to*] *take* [ART ~]	$Real_2$(*exam*)	= [*to*] *pass* [ART ~]
$Labreal_{12}$(*saw*)	= [*to*] *cut* [N *with* ART ~]	$Real_1$(*bus*)	= [*to*] *drive* [ART ~]
$Labor_{12}$(*interrogation*)	= [*to*] *subject* [N *to an* ~]	$Fact_0$($film_N$)	= *is playing, is on*
$LiquFunc_2$(*attention*)	= [*to*] *divert* [N'*s* ~ *from* N]	$Fact_0$($hope_N$)	= *comes true*

Die LF erlauben für Substantiv-Verb Kollokationen eine detaillierte Analyse der verschiedenen Konstruktionstypen aufgrund der Realisierung der Argumente des prädikativen Nomens innerhalb des syntaktischen Rahmen des Verbs. In Verbindung mit den LF, die Aktionsart und Kausativität ausdrücken, ergibt sich aus den lexikalischen Standardfunktionen ein formaler Apparat, mit dem viele Substantiv-Verb Kollokationen präzise beschrieben werden. Für die spezifischen Kollokationen, in denen für eine kleine Anzahl von Argumenten nur eine kleine Anzahl von Werten eine bestimmte Bedeutung einnimmt, stehen die Nicht-Standard Funktionen zur Verfügung (ohne Milchprodukt(*Kaffee)* = schwarz).

Mit den LF können alle Kollokationen charakterisiert werden (Mel'čuk 1998: 41).[22] Neben den fixierten, idiosynkratischen Phrasen, die als typische Kollokationen gelten, werden auch Phrasen erfasst, die sich regulär verhalten. So wird die LF $\mathbf{Oper}_1$ von Körperteilen immer durch das Verb *have* ausgedrückt. In Analogie zu der Mehrzahl der Fälle, in denen $\mathbf{Oper}_1$ durch phraseologisch verbundene lexikalische Einheiten zum Ausdruck kommt (wie in *pay attention*), werden alle Werte der LF als Phraseme und somit als Kollokationen betrachtet (Mel'čuk 1998: 42). Trotz des primär idiosynkratischen Charakters der LF, können die Werte für verschiedene *Keywords* identisch sein, was meistens bedingt ist durch eine semantische Ähnlichkeit der Argumente.

Entgegen dem arbiträren Charakter der Kollokation korrelieren laut Mel'čuk und Wanner die Bedeutung eines Lexems und dessen spezifische lexikalische Kookkurrenz: "lexemes with

22 Eine Ausnahme bilden die Kollokationen, in denen das Kollokat ein Aktant des *Keywords* ist. Verbindungen mit einem idiosynkratisch gewählten Aktanten (wie *auto-école, driving school, assurance maladie, assurance vie*) werden nicht durch die LF beim *Keyword* beschrieben, sondern unter dem Rektionsmuster einzeln aufgelistet.

common restricted lexical co-occurrence also share semantic features" (1994: 88). Die semantischen Merkmale der von Mel'čuk und Wanner untersuchten Gefühlssubstantive werden mit semantischen Dimensionen[23] bezeichnet, deren Werte positiv, negativ oder neutral sind. Die Korrelation zwischen den Werten der lexikalischen Funktionen und den semantischen Merkmalen bietet die Möglichkeit, diese Beziehung zu systematisieren. Die verschiedenen Substantive gehören idealerweise zu bestimmten Klassen, mit deren Teilnehmern sie Ähnlichkeiten im Kollokationsverhalten aufweisen. Die beiden Extreme bilden zum einen ganze semantische Klassen von Lexemen, die ein identisches restringiertes Kookkurrenzverhalten zeigen (*empfinden, fühlen* mit 'Gefühl'-Lexemen) und zum anderen individuelle Lexeme, die in idiosynkratischer, nicht generalisierbare Weise kookkurrieren (*machen* mit *Angst, Freude, Hoffnung*). Dazwischen liegt das Feld, das es zu klassifizieren gilt: in einer Domäne, die per Definition irregulär ist, gilt es Regelmäßigkeiten zu finden (Mel'čuk/Wanner 1994: 109-110).

Mel'čuk und Wanner formulieren 1994 das Prinzip der lexikalischen Vererbbarkeit: "All lexicographic data shared by a family of semantically related LUs should be stored just once - under one LU of the corresponding vocable or under the generic LU of the corresponding semantic field, from where these data are 'inherited' in each particular case" (zusammengefasst nach Mel'čuk 1998: 42). Der Lexikoneintrag für das Lexem 'Gefühl' besteht aus einem individuellen "privaten" Subeintrag für das spezifische Nomen und aus einem Subeintrag für das semantische Feld der Gefühlslexeme, der aus den rekurrenten Werten der LF der gesamten Teilnehmer besteht. Die Gemeinsamkeiten werden nur noch im "öffentlichen" Subeintrag unter Gefühl gespeichert. Hinter den Werten der LF werden die semantischen Bedingungen verzeichnet, die das Nomen als Argument erfüllen muss, um mit diesem Verb zu kombinieren:

> An emotion lexeme governs an NP denoting the Experiencer (X = I) and - if it has the SemA Z - an NP denoting the Reason for the emotion (Z = III).
>
> | X = I | Y = II |
> |---|---|
> | 1. N_{gen}
2. *von* N_{dat}
3. $Pron_{poss}$ | 1. *wegen* N_{gen} |
>
> | | | | |
> |---|---|---|---|
> | IncepPredMinus | : | nachlassen | \| 'excited-state' |
> | $Oper_1$ | : | empfinden, fühlen [~$_{acc}$]; | |
> | | | entgegenbringen [N_{dat} DET ~$_{acc}$] | \| 'attitudinal' |
> | | | haben [~$_{acc}$] | \| 'permanent' |
> | Magn + $IncepOper_1$ | : | geraten [in ~$_{acc}$] | \| 'manifested' |
> | | | ausbrechen [in ~$_{acc}$] | \| 'intense' ∧ 'manifested' |
> | $FinFunc_0$ | : | sich legen | \| 'excited-state' |
> | fast $FinFunc_0$ | : | verfliegen | \| 'excited-state' |
> | $Liqu_1Func_0$ | : | überwinden [$PRON_{poss}$ / DET ~$_{acc}$] | \| ¬ 'moderate' |
> | $IncepFunc_1$ | : | aufkommen [in N_{dat}] | |
> | Magn + $IncepFunc_1$ | : | erfassen [N_{acc}] | \| ¬ 'moderate' |
> | Magn + fast $IncepFunc_1$ | : | packen [N_{acc}] | \|'self-control-loss-inflicting' |
> | $Caus_2Func_1$ | : | hervorrufen [bei N_{dat} ~$_{acc}$], | |
> | | | erregen [in N_{dat} ~$_{acc}$] | \| 'reactive' |
> | $Liqu_1Fact_0$ | : | unterdrücken [$PRON_{poss}$ / DET ~$_{acc}$] | |
> | Magn + $IncepFact_1$ | : | überkommen [N_{acc}] | \| ¬ 'moderate' |
>
> (Mel'čuk/Wanner 1994: 118-119)

23 Die 11 semantischen Dimensionen sind: "intensity, polarity, manifestability, directionality, mentality, reactivity, attitudinality, activity, excitation, self-control, permanence" (Mel'čuk/Wanner 1994: 97).

Dadurch können für die einzelnen Gefühlssubstantive komprimierte Lexikoneinträge erstellt werden. Die fett hervorgehobenen Werte im Eintrag von 'Hoffnung' oben erweisen sich durch Vererbung der Eigenschaften des übergeordneten Eintrags als redundant. Auf der anderen Seite übersteigt bei etlichen neuen Einträgen der Gefühlssubstantive, die durch Vererbung an den öffentlichen Eintrag von 'Gefühl' angebunden sind, die Anzahl der als negativ zu verzeichnenden Werte die redundanten Werte (vgl. Mel'čuk/Wanner 1994: 143-154), es findet daher nicht immer eine effiziente Reduktion der Einträge statt. Auf der klassifikatorischen Ebene bildet die Vererbung der Kookkurrenz stark überlappende Klassen, die keine klare Hierarchie aufweisen. Daher zeigen die Angaben von Mel'čuk und Wanner keine "reine" Klassifikation von Lexemen anhand ihrer Kookkurrenzdaten, sondern die Korrelationen zwischen Werten von LF und semantischen Merkmalen der Argumente (Mel'čuk/Wanner 1994: 90).
Die geringe Anzahl der Werte der LF entsteht durch die Beschränkung auf 25 untersuchte Verben. Ein realistisches Szenario mit allen verbalen Kookkurrenzdaten würde die Klassifikationsproblematik zusätzlich steigern. Neben den komprimierten Lexikoneinträgen für die 40 Gefühlssubstantive bieten Mel'čuk und Wanner Abhängigkeiten von LF und semantischen Merkmalen, die auf der manuellen Auswertung corpusbasierter Kollokationsdaten und der Akzeptanz möglicher Kombinationen durch Muttersprachler basieren. Für jede der 11 semantischen Dimensionen werden Verben genannt, die analog zum semantischen Merkmal kombinieren: "'Permanent' emotion lexemes do not co-occur with the verb *geraten* '[to] get into' ($IncepOper_1$); they tend to co-occur (although with many exceptions) with the verb *haben* '[to] have' ($Oper_1$)" (Mel'čuk/Wanner 1994: 134). Die ebenfalls vorgenommene Systematisierung der substantivischen Kookkurrenten der Verben führt zu sehr heterogenen Daten, und ist nur bedingt mit semantischen Merkmalen zu beschreiben (Mel'čuk/Wanner 1994: 135-142). Trotz einer signifikanten Korrelation zwischen restringierter lexikalischer Kookkurrenz und semantischen Merkmalen ist ein großer Teil der Kollokationen idiosynkratisch und nur durch eine Auflistung zu erfassen (Mel'čuk/Wanner 1994: 91).
Ungeachtet der Schwierigkeiten, die sich bei der Klassifikation von Lexemen oder ihrer semantischen Merkmale in Korrelation zum Kookkurrenzverhalten ergeben, bieten die LF ein Modell, das eine adäquate Systematisierung der Substantiv-Verb Kollokationen bereitstellt. In Heid (1996) "Using Lexical Functions for the Extraction of Collocations from Dictionaries and Corpora" werden Möglichkeiten der Extraktion von Funktionsverben und deren automatische Einteilung nach LF auf syntaktischen Grundlagen diskutiert. Auch bei Heid wird der Zusammenhang zwischen semantischen und kollokationalen Eigenschaften von Substantiven (aus dem Wirtschaftsbereich) analysiert, die Ergebnisse werden in Kapitel 3.1.5 unter dem Stichwort konzeptuelle Kollokationen erklärt.
Die explikativen und kombinatorischen Wörterbücher *(DEC* 1984, 1988, 1992, 1999) bilden eines der zentralen Module der *Meaning-Text Theory.* In ihnen sind sämtliche lexikalischen Informationen enthalten, die in einem Modell der *Meaning-Text Theory* relevant sein können. Die lexikalischen Einheiten werden exhaustiv beschrieben und der Eintrag eines *Keywords* kann sich über mehrere Seiten erstrecken. Das *DEC* ist ein formales Wörterbuch innerhalb einer kohärenten linguistischen Theorie, das sich als lexikalische Datenbasis für die automatische Sprachgenerierung eignet (Mel'čuk 1998: 50). Der Einsatz von LF in der Maschinellen Übersetzung und Textgenerierung wird von Mel'čuk verdeutlicht. Dabei stehen die LF als "*tool*" in den computerlinguistischen Anwendungen unter folgenden Aspekten im Vordergrund:

1. LF garantieren die korrekte lexikalische Wahl, nur das *Keyword* wird übersetzt und die Werte der Kollokate durch die LF bestimmt;
2. ändert sich durch die lexikalische Wahl die syntaktische Struktur des Satzes, kann die Ersetzung einer LF durch eine andere LF in der Zielsprache durch Standardtransformationen beim *Keyword* verzeichnet werden;
3. LF präzisieren die kommunikative Struktur durch paraphrasierende Gleichungen (V ⇔ S_0(V) + $Oper_2(S_0(V))$ *'X analyzes Y ⇔ Y undergoes an analysis by X');*
4. LF ermöglichen eine maximale Textkohäsion und vermeiden überflüssige Wiederholungen durch anaphorische Verbindungen und die Wahl von semantischen und syntaktischen Derivaten.

   (Mel'čuk 1996: 91-96, 1998 43-49).

Konkrete Anwendungen der LF in der Textgenerierung werden von Heid/Raab (1989) und in der Maschinellen Übersetzung von Heylen/Maxwell/Verhagen (1994) aufgezeigt, Apresjan et al. (2002) geben einen Überblick über die Verwendung von LF in weiteren NLP-Bereichen.
Die Wörterbücher Mel'čuks bestimmen lexikalische Bezüge in umfassender und systematischer Weise und repräsentieren anhand bestimmter *Keywords* einen Ausschnitt aus einer lexikalischen Datenstruktur, der eher die Funktion einer universalen Sprachbeschreibung oder Interlingua zukommt, als die, Sprache in einer für den Menschen optimal interpretierbaren Weise zu zeigen. Eine formale Datenbasis, die als Grundlage von NLP-Systemen dient und an den Einträgen im *DEC* orientiert ist, sowie ein Wörterbuch für den Sprachen-lerner, das die Namen der LF in einer einfach verständlichen Form mit popularisierten Ausdrücke wiedergibt und zur Gliederung seiner Einträge nutzt, werden in Kapitel 3.2.3 vorgestellt. Das *DiCouèbe (Dictionnaire en Ligne de Combinatoire du Français)*[24] ist das entsprechende Wörterbuch im Internet, das als Interface die 520 Vokabeln zugänglich macht, die in der formalen Datenbasis spezifiziert werden. Im Internet findet man auch ein Wörterbuch zu Kollokationen des Spanischen, das *DiCE* (*Diccionario de Colocaciones del Español*)[25], welches zu ausgesuchten Substantiven der Gefühle Kollokationsinformationen im Format des *DEC* bietet. Die beiden elektronischen Wörterbücher werden in Kapitel 6.1.2 genauer erläutert.
Kritik an der Darstellung von Kollokationen mit LF wird von Grossmann/Tutin (2003: 13) geübt. In ihrem Artikel "Quelques pistes pur le traitement des collocations" sehen auch sie einen fundamentalen Beitrag im Modell der LF zu einer adäquaten Kollokationsbeschreibung. Doch ist zum einen die syntaktische Annotation nicht ausreichend - im Falle der Funktionsverben bleibt man über die Möglichkeiten der Passivierbarkeit, der Modifizierbarkeit des Nomens und eines Wechsels der Determinanten im unklaren. Zum anderen werden unter den synonymen Werten einer LF keine Angaben zu Frequenz, situativem Kontext oder Intensität gemacht. Das größte Problem stellt die Kodierung der Einträge dar: "... although the notion of lexical function in itself is very appealing ... , using the very LFs proposed by Mel'čuk is impractical without the collaboration of an MTT guru"[26]. In dem Artikel "Critères heuristiques pour l'encodage des collocations au moyen de fonctions lexicales" schlägt Alonso Ramos (2000) einen Kriterienkatalog vor, der die Bestimmung der LF

24 http://olst.ling.umontreal.ca/dicouebe/
25 http://www.dicesp.com
26 Robin, J. (1990): *Lexical Choice in Natural Language Generation, CUCS-040-90*. New York, Columbia University: 26. Zitiert nach Wanner (1996: 2).

erleichtern soll. Ein automatisches Klassifikationsverfahren, das für Substantiv-Verb Kollokationen die entsprechenden LF bestimmt, beruht auf semantischer Dekomposition der beteiligten Lexeme und wird im folgenden Kapitel vorgestellt.

2.4.2. Das Klassifikationsverfahren von Wanner

Ein Verfahren zur automatischen semantischen Klassifikation von Kollokationen, das die LF als Klassifikationsgrundlage wählt, wurde von Leo Wanner entwickelt.[27] Verdeutlicht wird die Methode für Nomen-Adjektiv - und Nomen-Verb Kollokationen anhand spanischer Textcorpora und dem spanischen Teil des EuroWordNet[28]. Die semantischen Informationen, die die externe lexikalische Datenbasis bereitstellt, ermöglichen es, für Wortkombinationen eine signifikante Anzahl der gebräuchlichsten syntagmatischen LF zu finden. Techniken aus dem Maschinellen Lernen werden angewandt, um die Instanzen der LF zu bestimmen. In drei Schritten geschieht die Akquisition und Klassifikation der Kollokationen:

1. Corpusprozessierung: Extraktion möglicher zu klassifizierender Kandidaten, deren syntaktisches Muster zur syntaktischen Struktur mindestens einer LF passt.
2. Lernen: Für jede LF werden Trainingsbeispiele manuell zusammengestellt. Die charakteristischen semantischen Profile werden anhand der semantischen Merkmale der Beispiele gelernt.
3. Klassifikation: Die semantischen Merkmale der Kandidaten werden mit den gelernten semantischen Angaben der LF verglichen. Diejenige LF, deren semantisches Profil dem Kandidaten am ähnlichsten ist, wird als Zielfunktion gewählt. Ist die Ähnlichkeit eng genug, wird der Kandidat als Instanz der LF klassifiziert.
   (Wanner et al. 2005b: 150-151).

Dem Schritt der Corpusprozessierung folgt zunächst ein manueller Schritt, in dem alle irrtümlich extrahierten Bigramme entfernt werden.[29] Als Trainigsbeispiele dienen Kollokationen aus Wörterbüchern oder ein Teil der extrahierten Kollokationen aus den Corpora. Auch die Trainingsbeispiele müssen manuell zusammengestellt und den einzelnen LF zugeordnet werden. Ebenso geschieht die Disambiguierung der einzelnen Lexeme von Hand, mit der eventuelle polyseme Lesarten auszuschließen sind. Die semantische Beschreibung der betreffenden Lexeme stammt aus den Einträgen des spanischen Teils der multilingualen lexikalischen Datenbasis EuroWordNet, die an dem amerikanischen Prototyp Princeton WordNet[30] für das Englische orientiert ist, und in die 8 europäische Sprachen integriert sind.

Die Darstellung von Bedeutungen erfolgt mit Hilfe semantischer Relationen. Grundlegende Einheit ist das Synset, eine Menge von Wörtern, die in einem bestimmten Kontext austauschbar sind und dann als Synonyme gelten. Nomina und Verben sind hierarchisch angeordnet und mittels der Relation Hyponymie und Hyperonymie zwischen den Synsets verbunden. Basiskonzepte können als Archilexeme für semantische Felder betrachtet werden, sie fungieren als zentrales Vokabular des polylingualen Wortnetzaufbaus und garantieren die Kompatibilität der einzelnen Sprachnetze. Die Basiskonzepte (ca. 1000 Nomen und 300 Verben) werden durch Merkmale oder Merkmalskombinationen aus der

27 Erläutert werden Verfahren und Ergebnisse in Wanner (2004) und Wanner et. al (2005a, 2005b).
28 http://www.illc.uva.nl/EuroWordNet/
29 Die Extraktion erfolgt einmal aus einem Text mit POS-Tags (Wanner et al. 2005b), das andere mal mit *Partial Parsing* (Wanner 2004), Wanner et al. (2005a) verwenden die Kollokationen aus dem *DiCE*.
30 http://wordnet.princeton.edu/ Einen kurzen Überblick über die Struktur der beiden lexikalisch-semantischen Netze gibt Kunze (2001), ausführlichere Literatur ist unter den Internetadressen zu finden.

Top-Ontologie charakterisiert. Die Synsets der Basiskonzepte variieren sprachindividuell, die 63 Topkonzepte sind sprachunabhängige Komponenten. Die Hyperonym-Hierarchie für *admiración* ('Bewunderung') verdeutlicht die Struktur eines Eintrags (Lexeme in Großbuchstaben mit dem Index für die Lesart, Basiskonzepte sind klein geschrieben, Topkonzepte beginnen mit einem Großbuchstaben):

(4. feeling ADMIRACIÓN3
3. feeling AFICIÓN2 GUSTO5
2. Tops Dynamic | Experience | Mental SENTIMIENTO1
1. Tops Mental | Property RASGO-PSICOLÓGICO1)

(nach Wanner et al. 2005a: 262)

Nur durch die Vereinigung der drei Komponententypen (Hyperonyme, Basiskonzepte, Topkonzepte) entspricht die semantische Beschreibung den diskriminatorischen Anforderungen einer automatischen Klassifikation. In Wanner (2004) werden zwei Experimente für Substantiv-Verb Kollokationen durchgeführt. Das eine für Nomina desselben semantischen Feldes (Substantive der Gefühle), das andere Experiment für beliebige Substantive. Als LF werden zum einen diejenigen gewählt, die ähnliche Typen von Kollokationen beschreiben (darunter werden LF verstanden, die die semantische Aktantenstruktur des Nomens in gleicher Weise auf die syntaktische Struktur des Verbs projizieren) und als Kontrast mindestens eine LF, die deutlich von den anderen abweicht. Für Experiment 1 und 2 werden unterschiedliche LF gewählt.

Die Klassifikation basiert auf einer Implementierung der *Nearest Neighbour* Technik des instanzbasierten Lernens. Die Assoziationsstärke eines Wortpaars der Form $\langle b_i,c_j \rangle$ wird anhand der Bedeutungskomponenten $\{b_1,b_2,...,b_i\}$ und $\{c_1,c_2,...,c_j\}$ entsprechend der semantischen Angaben in EuroWordNet mittels der paarweisen Kookkurrenzwahrscheinlichkeit der Bedeutungskonzepte der beiden Lexeme nach einer Formel ähnlich der punktweisen Mutual Information berechnet.[31] Der Durchschnitt aller Werte der Instanzen einer LF ergibt den Wert der "prototypischen Kollokation" dieser LF, eine künstliche ideale semantische Repräsentation, die den Zentroid der LF bildet. Die Differenz zwischen dem Wert W einer LF-Instanz und dem Zentroid $\bar{W}(LF)$ einer LF zeigt an, wie typisch diese LF-Instanz für eine bestimmte LF ist. In Experiment 1 und 2 wird für jede der LF der ideale maximale Wert der Abweichung Δ anhand der Precision- und Recall-Werte empirisch bestimmt, den eine LF-Instanz vom Zentroid höchstens entfernt liegen kann, um noch als Argument-Wert Paar für diese LF zu gelten. Für das homogene Wortfeld der Gefühlssubstantive werden die drei Komponententypen der semantischen Beschreibung gleich gewichtet, für die Klassifikation der heterogenen Substantive ist eine Gewichtung der Komponentensignifikanzfunktion τ von Hand für jede LF und für das Argument und den Wert separat anhand einer Testsuite zu leisten, um befriedigende Ergebnisse zu erzielen[32].

Sind diese Schritte durchgeführt, werden in Experiment 1, bei einem variierenden idealen Wert von Δ, für alle LF Ergebnisse mit hohen Precision- und Recall-Werten erreicht. Die Werte des *f*-scores, bei dem Precision und Recall gleich gewichtet sind, liegen zwischen 78% und 100%. Die Abweichung der Werte einiger Kandidaten vom Zentroid der jeweiligen LF und deren Klassifikation zeigt folgende Tabelle:

31 Die genauen Rechenverfahren und der Algorithmus werden in Wanner (2004: 102-107) wiedergegeben.
32 Die genauen Werte sind in Wanner (2004: 117) zu finden.

	$Caus_2Func_1$ (15.131)	$Oper_1$ (12.743)	$ContOper_1$ (3.481)	$FinFunc_0$ (10.863)	$IncepFunc_1$ (1.337)
[la] admiración cesa	1.170	1.684	5.626	**0.029**	4.599
[la] admiración se desvanece	0.985	1.512	6.153	**0.139**	6.730
guardar admiración	1.297	1.289	**0.225**	1.344	1.306
sentir admiración	1.027	**0.004**	3.909	1.925	8.964
experimentar alegria	1.021	**0.013**	3.935	1.941	8.984
tener alegria	0.980	**0.108**	8.679	1.282	6.940
producir [una ADJ] decepción	**0.133**	2.175	5.218	1.780	9.987
[la] desesperación se apodera [de N]	1.089	1.448	1.519	1.180	**0.539**
provocar [una ADJ] desesperación	**0.470**	2.596	6.994	2.144	15.403 ...

(Wanner 2004: 137)

In Experiment 2 hingegen ist die Homogenität von richtigen Kandidaten unterhalb eines bestimmten Wertes von Δ nicht gegeben. Die dekompositionale Beschreibung einiger positiv klassifizierter Kandidaten weicht erheblich von der Beschreibung der LF-Instanzen im Trainingsset ab. Der *f*-score liegt hier auch bei einem idealen Wert für Δ nur zwischen 58% und 76%. Die Instanzen des Trainingssets teilen selten ein gemeinsames Hyperonym und die zu klassifizierenden Kandidaten enthalten nur in wenigen Fällen ein Lexem, das auch im Trainingsset vorhanden ist. Die Verteilung von positiven und negativen Instanzen sortiert nach dem Wert der Abweichung Δ für die LF $Real_2$ zeigt folgendes Bild:

obedecer [la] regla	**0.053**	...	
estar [sobre] la agenda	0.103	echar [la] culpa	1.640
aceptar [un] argumento	**0.304**	seguir [la] etiqueta	**1.674**
reconocer [una] acusación	**0.326**	hacer tentativa	1.687
sellar [un] convenio	0.482	satisfacer [una] condición	**1.742**
tener [un] pasaporte	0.591	aceptar [una] excusa	**1.763**
seguir [un] consejo	**0.594**	recibir [un] homenaje	1.813

(Wanner 2004: 140)

Hier wird deutlich, dass der numerische Wert der Abweichung vom Zentroid nur bedingt Aufschluss über die Unterscheidung von positiven und negativen Kandidaten gibt.
Die Ergebnisse der Klassifikation werden hauptsächlich durch drei Faktoren beeinträchtigt: Eine Ähnlichkeit der untersuchten LF spiegelt sich in einer ähnlichen semantischen Beschreibung ihrer Instanzen wieder, wodurch die Fehlerrate der Klassifikation ansteigt. Die Zusammenstellung des Trainingssets ist entscheidend, repräsentative Basen mit charakteristischen semantischen Eigenschaften für jedes der Wortfelder, aus denen die Kandidaten stammen, müssen enthalten sein. Die semantische Beschreibung der Lexeme in EuroWordNet zeigt erhebliche Lücken, spezifische Bedeutungen von Substantiven und Verben, die diese in Kollokationen haben, sind häufig nicht als polyseme Lesart verzeichnet. Es werden ähnliche Lesarten als Ersatz gewählt, was zu unpräzisen Berechnungen führt. Probleme bereitet auch die unterschiedliche Beschreibung semantisch ähnlicher Wörter und die fehlerhafte Auszeichnung der Lexeme mit Bedeutungskomponenten (Wanner 2004: 129-132).
Ausschlaggebend scheinen aber letztendlich die semantischen Eigenschaften der Substantive und das dadurch determinierte Kookkurrenzverhalten zu sein. In Experiment 1, das ausschließlich mit Nomina aus dem Wortfeld der Gefühle arbeitet, werden akkurate Ergebnisse erzielt, während in Experiment 2, das heterogenen Daten verarbeitet, die Abweichung einer

Instanz vom Zentroid nur bedingt Auskunft über die korrekte Zuordnung zu einer LF gibt. In Wanner et al. (2005b) wird ein weiteres Paradigma des Maschinellen Lernens angewandt, das auf Bayes'schen Netzen aufbaut. Beide Methoden werden mit heterogenen Substantiv-Verb Kookkurrenzen getestet, die einmal freie Wortkombinationen enthalten, das andere mal nur aus Kollokationen bestehen. Die *Nearest Neighbour* Technik schneidet in beiden Experimenten besser ab, da mit den Netzwerken eine Abbildung des transitiven Verhältnisses zwischen den Bedeutungselementen des Nomens und des Verbs nicht zu simulieren ist. Beide Verfahren erzielen erheblich bessere Ergebnisse, wenn sie nur Kollokationen klassifizieren und sich damit auf eine Lesart der Verben konzentrieren. Wie in den Experimenten sichtbar wird, widerspricht zwar die Klassifikation aufgrund der Ähnlichkeit semantischer Merkmale dem idiosynkratischen Charakter der Kollokation, doch schließt die Arbitrarität der Kollokation eine partielle semantische Motivation nicht aus, was sich am deutlichsten innerhalb eines Wortfelds mit semantisch ähnlichen Merkmalen zeigt.

3. Kollokationen in lexikografischer Theorie und Praxis

Mit der Möglichkeit der automatischen Prozessierung größerer Corpora ab Mitte der 70er Jahre[33], regte sich das lexikografische Interesse an den Kollokationen. Die Vertreter des Britischen Kontextualismus befassten sich mit den Kollokationen, um zu einer formalen Beschreibung der lexikalischen Ebene der Sprache zu gelangen. Die Lexikografen interessieren die Darstellung von Kollokationsinformationen im Wörterbuch und die linguistische Interpretation der Kollokationen, um über eine Klassifikation der lexikalischen Verbindungen die Auswahlkriterien für die Aufnahme von Kookkurrenzdaten als Kollokationen in ein Wörterbuch zu motivieren.

3.1. Kollokationen und verwandte Wortkombinationen

3.1.1. Oxford Dictionary of Current Idiomatic English (Cowie)

Die Arbeit am *Oxford Dictionary of Current Idiomatic English 1: Verbs with Prepositions & Particles* (1975) (*ODCIE1*) veranlasste Cowie dazu, neu zu überdenken, welche Art von Wortkombinationen in ein idiomatisches Wörterbuch aufzunehmen sind. Idiome als Phrasen oder Sätze, deren "Gesamtbedeutung nicht aus der Bedeutung der Einzelelemente abgeleitet werden kann, vgl. *jemanden auf die Palme bringen* ‘jemanden wütend machen’" (Bußmann 1990), stellen nur den ersten Typ seiner Kategorisierung von Wortkombinationen dar. Cowie bezeichnet sie als *pure idioms.*[34] Cowie schlägt zwei Kriterien für die Kategorisie-rung von Wortkombinationen vor: a) die Art der Beziehung zwischen der Bedeutung der Wortkombination als Ganzes und den Bedeutungen ihrer Konstituenten, b) die Möglichkeiten der Ersetzung von Teilen der Wortkombination. Die Konstituenten der *pure idioms* sind in keiner Weise variierbar.

Die Ausdrücke in der Kategorie der *figurativen idioms* haben sowohl eine übertragene als auch (noch) eine wörtliche Bedeutung (*catch fire, close ranks, act the part/role*). Variationsmöglichkeiten in einer der Konstituenten sind selten. Bei den *restricted collocations* ist das Hauptkriterium darin zu sehen, dass ein Bestandteil im übertragenen Sinne gebraucht wird, während der andere Bestandteil seine 'Normalbedeutung' behält (*jog sb's memory, a blind alley, a cardinal error/sin/virtue/grace).* Mitunter ist ein gewisses Maß an Variation möglich. Der Grund für ihre Aufnahme in das Wörterbuch ist die Festlegung einer besonderen Bedeutung durch einen begrenzten Kontext. Das typische Merkmal der *open collocations* ist, dass beide Elemente in ihrer Normalbedeutung gebraucht werden, woraus Cowie schließt, dass beide Bestandteile frei kombinierbar sind.

Cowie spricht sich zwar dafür aus, in einem Wörterbuch der Idiome nur Wortkombinationen zu verzeichnen, die den drei Kategorien der *pure idioms, figurative idioms* und *restrictive collocations* angehören, dennoch werden auch die "collocates most often heard" (*ODCIE2*:

33 Das von Jones und Sinclair (1973: 17) bearbeitete Corpus unterscheidet sich bezüglich seiner Zusammensetzung von den heute zur Kollokationsakquisition üblicherweise verwendeten Corpora. Das Corpus (150.000 Wörter) bestand zu ca. 90% aus der Transkription spontaner Konversation, die restlichen 10% bildeten geschriebene wissenschaftliche Texte. Diese Auswahl wurde mit der Erwartung getroffen, Kollokationen nicht nur satzgrenzenübergreifend nachzuweisen, sondern auch verteilt auf verschiedene Redebeiträge.

34 Die Beschreibung der Kategorien und der Kategorisierungskriterien werden in ihrer knappen, übersichtlichen Form aus Bahns (1996: 15-16) übernommen. Cowie legt seine Gedanken ausführlich in den beiden Vorworten der idiomatischen Wörterbüchern des Englischen dar (*Oxford Dictionary of Current Idiomatic English 2: Phrase, Clause & Sentence Idioms* (1983) (*ODCIE2*)).

xiv) verzeichnet. Cowie möchte Englischlernern anhand des lexikalischen Kontexts Paradigmen für die Einsetzbarkeit weiterer Lexeme aufzeigen und die Möglichkeit bieten, mittels der Kollokate zwischen verschiedenen Bedeutungen eines Lemmas zu entscheiden (*ODCIE1*: xiii-xv).
Die verzeichneten Lemmata sind alle Verben, als Kollokate kommen Substantive und Adverbien bzw. adverbiale Phrasen in Frage. Die Verben und Nomina werden mit Valenzinformationen bzw. Angaben zur syntaktischen Kombinierbarkeit gekennzeichnet. Die Bedeutungen der Lettern und Abkürzungen beim Verb entsprechen verschiedenen "Satzmustern" (B1 = Verb + Particle + Transitive, B2 = Verb + Preposition + Transitive) und der für sie möglichen "grammatischen Transformationen" (pass = Passivtransforma-tion), die Lettern beim Nomen repräsentieren die wichtigsten Konstituenten der Satzmuster. Cowie widmet der Klassifikation der Satzmuster und grammatischen Transformationen 30 Seiten im Vorwort des Wörterbuchs (*ODCIE1*: xxviii-lvii). Zwischen dem Lemma und den Kollokaten steht die allgemeine Bedeutungsangabe:

> **hammer in/into**[1] [B1i pass B2 pass] force, drive, in by striking. **O**: post, stake; nail, rivet □ *First the pegs were **hammered into** the ground in large circle, then the tent ...*
>
> **hammer in/into**[2] [B1i pass B2 pass] force sb to learn sth (by tiresome repetition). **S**: teacher, preacher, pedant. **O**: lesson; moral, point; French, grammar. **o**: flock, class; head □ *Please don't advise me not to marry John; my parents have been **hammering** the lesson **in** long enough* □ *We had Latin **hammered into** our heads for five years ...*
>
> **hammer on/onto** [B1i pass B2 pass] fasten in position by beating or striking. **S**: Blacksmith, carpenter. **O**: lid, top, metal cap □ *A protective copper strip was ...*
>
> (*ODCIE1* 1975)

Im eigentlichen Sinne als idiomatisch zu bezeichnen ist nur die zweite Bedeutungsangabe von *hammer in/into* im Sinne eines *figurative idiom*. In den beiden anderen Artikeln handelt es sich um die Angabe der wichtigsten Kontextpartner des Verbs, um *open collocations*. Kontextpartner einer *restrictive collocation* werden in diesem Wörterbuchausschnitt nicht gezeigt, sie werden zusätzlichen mit einem Ausrufezeichen markiert und sind selten im Wörterbuch zu finden. Das Ausrufezeichen weist darauf hin, dass die Verwendung anderer als der genannten Kollokate im Sprachgebrauch nicht korrekt ist (*kick up* **!**△ *a row, fuss, dust, shindy*). Die Einträge entstanden noch anhand der manuellen Auswertung eines explizit aufgeführten Corpus geschriebener und gesprochener Sprache *(ODCIE1*: lviii-lix).

3.1.2. BBI Combinatory Dictionary of English (Benson)

Das *BBI Combinatory Dictionary of English* (1986) von Benson bietet "essential grammatical and lexical *recurrent word combinations*, often called *collocations*" (*BBI*: vii). Im Gegensatz zu den meisten Kollokationskonzepten, die sich auf die Kombinierbarkeit von Lexemen beschränken, behandelt Benson die *grammatical collocations* parallel zu den *lexical collocations*.
Die *grammatical collocations* (*BBI*: ix-xxiii) bestehen aus einem dominanten Wort (Nomen, Adjektiv, Verb) und einer Präposition oder grammatischen Struktur (Infinitiv, Nebensatz). Zum einen spielen bei der Klassifizierung der Kombinationsmöglichkeiten der grammatischen Kollokationen syntaktische Angaben eine Rolle wie *verb patterns.* Sie werden wie bei Cowie mit Lettern markiert und spezifizieren im Wörterbuch die Verben. Auch die Typen, die aus einem dominanten Wort und einem *to + infinitive* oder *that clause* bestehen, werden mit syntaktischen Parametern gebildet. Außer bei Benson werden diese Phänomene als relevante Informationen zu Kollokationen betrachtet, die auf den Valenzeigenschaften eines

der Kollokationspartner beruhen, sie bilden daher keinen klassifikatorischen Bezugsrahmen für Kollokationen. Zum anderen werden unter den *grammatical collocations* die Kombinationen der dominanten Wörter mit einer Präposition als Typen aufgeführt. Bezüglich der Aufnahme von Präpositionen in das Paradigma der Kollokationen gibt es heute verschiedene Ansichten. Das *Oxford Collocations Dictionary (OCD)* bezieht sie in seine Kollokationstypen mit ein, bei Hausmann hingegen gehört die Kombinierbarkeit mit einer Präposition nicht in den Bereich der Kollokationen (vgl. auch die unterschiedlichen Auffassungen zur Tripel-Kollokationsextraktion in Kapitel 2.3.2 und 2.3.3)
Die *lexical collocations* bestehen aus 7 Typen: V+N *creation/activation*, V+N *eradication/ nullification*, ADJ+N, N+V, N+N, ADV+ADJ, V+ADV (*BBI*: xxiv-xxviii). Der Unterschied zwischen den ersten beiden Typen ist semantisch motiviert, die Verbbedeutung fungiert als klassifikatorisches Element.[35] Auch der N+V Typ wird zusätzlich semantisch definiert: "the verb names an action" (*BBI* xxvii). Die anderen Typen unterscheiden sich aufgrund der Wortarten. Die verschiedenen Klassen der *grammatical* und *lexical collocations* werden im Wörterbuch (mit Ausnahme der *verb patterns*) nicht verzeichnet, sie bilden vielmehr die Grundlage für die Reihenfolge der aufgeführten Kollokate. Im *BBI* wird auf Corpusbelege in Form von ganzen Beispielsätzen verzichtet zugunsten illustrativer Phrasen:

> **hope** I *n*. 1. to arouse, inspire, stir up ~2. to raise smb.'s ~s 3. to express, voice a ~ 4. to cherish, entertain, nurse a ~ 5. to pin, place, put one's ~s on 6. to dash, deflate, dispel; thwart smb.'s ~s 7. to abandon, give up ~ 8. an ardent, fervent, fond; faint, slender, slight; false; high; idle, illusory, vain; real; realistic, reasonable; unrealistic, unreasonable ~ 9. ~s come true; fade 10. a flicker, glimmer, ray, spark of ~ 11. ~ for, in, of (~ of recovery; we had high ~s for her) 12. a ~ that + clause (it was our ~ that they would settle near us; there was little ~ that she would be elected) 13. in, with the ~ (we returned to the park in the ~ of finding her wallet) 14. beyond, past ~
>
> **hope** II *v*. 1. to ~ fervently, sincerely, very much 2. (D; intr.) to ~ for (to ~ for an improvement) 3. (E) she ~s to see them soon 4. (L) we ~ that you are comfortable 5. (misc.) I ~ so; I ~ not
>
> (*BBI* 1986)

Im Gegensatz zu Cowie nimmt Benson das Kriterium der Frequenz explizit in die Abgrenzung der *lexical collocations* gegenüber den *free lexical combinations* mit auf: "Free lexical combinations are those in which the two elements do not repeatedly co-occur; the elements are not bound specifically to each other; they occur with other lexical items freely" (*BBI*: xxiv). Es besteht eine Verbindung zwischen der Spezifizität (oder Restriktivität bei Cowie) einer Kollokation und ihrer Frequenz. Sind beide Kombinationspartner nicht spezifisch und kombinieren mit einer Vielzahl anderer Wörter, kommen sie mit diesen jeweils nur selten vor.[36] Da das Wörterbuch von Benson nicht auf einem bestimmten Corpus und dessen statistischer Bearbeitung basiert, sondern durch den Vergleich lexikografischer Quellen entsteht, entspringt diese Behauptung einer theoretischen Überlegung. Freie Kombinationen und Idiome sollen in dem Wörterbuch über die Kombinatorik der englischen

35 Die V+N/N+V Typen bei Benson entsprechen bestimmten lexikalischen Funktionen bei Mel'čuk. Benson beschreibt 1985 in dem Artikel "Lexical Combinability" den Zusammenhang zwischen lexikalischen Funktionen, Kollokationen und freien Kombinationen. Benson ist bestrebt durch die semantisch eingeschränkten Kollokationstypen die Aufnahme freier Kombinationen, die das Charakteristikum anderer lexikalischer Funktionen sind, zu vermeiden.

36 Dieser Gedanke wurde schon 1978 von Cowie in dem Artikel "The Place of Illustrative Material and Collocations in the Design of a Learner's Dictionary" formuliert. Cowie bezieht sich jedoch nur auf den positiven Zusammenhang zwischen figurativen Idiomen bzw. restriktiven Kollokationen und deren häufigem Auftreten.

Sprache nicht erscheinen. Ohne genauere Begründung wird festgelegt, dass die Eintragung der ersten fünf Kollokationstypen ausschließlich unter dem Nomen erfolgt, die Eintragung der letzten beiden unter dem Adjektiv bzw. Verb.

3.1.3. Hausmann

Bei Hausmann spiegelt der Ort der Eintragung einer Kollokation im Wörterbuch eine gerichtete Beziehung wider, die Unterscheidung von Basis und Kollokator (vgl. Kapitel 1 und 3.2.1). Fasst man die beiden ersten Typen der lexikalischen Kollokationen von Benson zu einem Typ zusammen, erhält man die von Hausmann vorgeschlagenen Strukturen (**V**+N, **N**+V, **N**+N, **N**+Adj, **V**+Adv, **Ad**j+Adv) für Kollokationen.[37] Bezüglich der Abgrenzung der Kollokationen von struktruverwandten Kombinationen findet man bei Hausmann folgende Gliederung, die sich nach der Fixiertheit und Affinität von Wortkombinationen richtet:

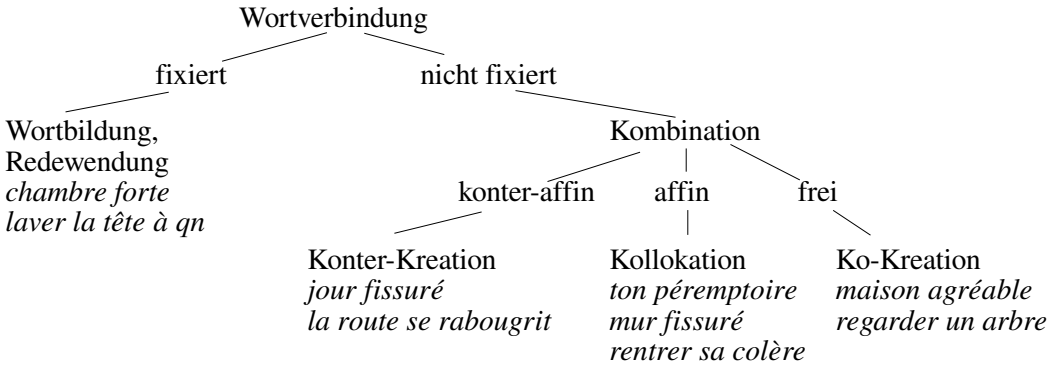


Abb. 4: Typologie der Wortkombinationen (Hausmann 1984: 399)

Die wenig begrenzte Kombinierbarkeit zeichnet den Kollokator der Ko-Kreationen aus, Ko-Kreationen werden entsprechend den Regeln des Sprachsystems kreativ zusammengestellt. Sie verbinden sich entsprechend gewisser semantischer Mindestregeln und sind von unauffälliger Üblichkeit. Die Affinität der Kollokationen wird definiert als die "Neigung zweier Wörter, kombiniert aufzutreten" (Hausmann 1984: 398). Affine Kombinationen werden nicht kreativ zusammengestellt, sondern als Ganzes aus der Erinnerung geholt. Die auffallende Üblichkeit der affinen Kombinationen sieht auch Hausmann im Zusammenhang mit einer begrenzten Kombinierbarkeit des Kollokators. Die Wörter mit begrenzter Kombinierbarkeit verbinden sich entsprechend differenzierter semantischer Regeln. Konter-Kreationen sind regeldurchbrechende Wortkombinationen, sie sind oft einmalig, unüblich und kennzeichnen literarischen Stil.

In anderen Arbeiten betont Hausmann den idiosynkratischen Charakter der Kollokation. Kollokationen werden nicht nur definiert über ihre Üblichkeit oder den Grad der Kombinierbarkeit, sondern über die Vorhersehbarkeit der Kombination in einer kontrastiven Sprachsituation:

> La collocation se distingue de la combinaison libre par la combinabilité restreinte (ou affinité) des mots combinés. La collocation se distingue d'autre part des locutions par son non-figement et par sa transparence. Or, cette transparence n'empêche nullement la collocation d'être imprédictible. L'apprenant étranger, tout en la comprenant (s'il comprend les mots combinés), ne saurait automatiquement la reproduire. Il doit l'apprendre, parce que les langues dans la totalité des combinaisons logiquement

37 Die Unterscheidung von Basis und Kollokator wurde von Hausmann 1979 eingeführt, die Präzisierung der Strukturtypen folgte 1989. Fett markiert werden in den Kollokationsstrukturen die Basen.

> possibles, font un choix idiosyncratique. La collocation est une unité, non de la parole, mais de la langue). (Hausmann 1989: 1010)

In der Definition Hausmanns wird die Abgrenzung zu den Idiomen erläutert und die Affinität der Kollokation noch einmal präzisiert. Der Terminus 'Transparenz' wird eingeführt um zu verdeutlichen, dass Kollokationen für den Fremdsprachenlerner zwar verstehbar, aber nicht unbedingt reproduzierbar sind aufgrund der 'idiosynkratischen' Wahl des Kollokators. Die Affinität wird nun definiert als eingeschränkte Kombinierbarkeit und nicht mehr als übliche Kombination. Damit ist aber offenbar keine Restriktivität im Sinne Cowies gemeint. Die Affinität ergibt sich eher aus der idiosynkratischen Wahl des Kollokators. Eine Kollokation wie *schwer verletzt*, die Hausmann als Beispiel gibt, würde bei Cowie zu den *open collocations* zählen, da der Kollokationsradius von *schwer* sehr groß ist. Für Hausmann ist sie keine Ko-Kreation, denn die Übersetzung der Kollokation ins Englische oder Französische ist nicht wörtlich äquivalent (*seriously injured*, *grièvement blessé*).

Hausmann führt zur Unterscheidung einer Kollokationen von einer Ko-Kreation in seinen frühen Arbeiten verschiedene Kriterien an: die "Üblichkeit" der Kombination, Restriktionen in der Kombinierbarkeit eines der Kollokationspartner, den sprachkontrastiven Gesichtspunkt, und die Differenziertheit der semantischen Regeln, die Basis und Kollokator verbinden. Unter dem Motto "Was sind eigentlich Kollokationen?" erweitert Hausmann 2004 die Kollokationsmöglichkeiten innerhalb des vorgegeben Kanons der Strukturtypen der lexikalischen Kollokationen durch die Aufnahme bestimmter Teilidiome und lässt neue Kollokationsstrukturen zu (313-317):

- in Vergleichsphrasemen wie ***dumm*** *wie Bohnenstroh* oder *wie die Faust aufs Auge* ***passen*** ist das Verglichene die Basis, der Kollokator ist ein Idiom, er ist von Sprache zu Sprache unvorhersehbar. Adjektive oder Verben werden in diesen Kombinationen zu Basen von Substantiven.
- auch unter dem Stichwort "ungewöhnliche Kollokationssyntax" wird eine abweichende Verteilung von Basis und Kollokator auf die Kollokationsbestandteile verzeichnet: a) eine deutsche Akkusativ-NP kann eine adverbiale Graduierungsbedeutung haben (*Bauklötze* ***staunen****, nur Bahnhof* ***verstehen***) - sie ist in diesem Fall nicht passivierbar und fungiert als Kollokator des Verbs - und b) ein Kopulaverb kann ein adjektivisches Prädikativ bedienen - das Adjektiv ist hier die unproblematische Basis, das Verb der unvorhersehbare Kollokator, der nach Art eines Funktionsverbs (in einem adjektivischen Funktionsverbgefüge) agiert (***verrückt*** *spielen,* ***kaputt*** *gehen*).
- feste Attribuierungen wie *krummer* ***Hund*** (in der Skizze von 1984 als 'Wortbildung' bezeichnet) gehören nun zu den Kollokationen, wenn einer der Kollokationspartner im weiteren Sinne seine Bedeutung behält (hier Hund = Mensch). Der unterschiedliche Transparenzgrad des Kollokators (im Vergleich zu *feiger* ***Hund***) ändert nichts am Status der Einheit. Bei der festen Attribuierung bleibt das für diese Kollokationsstruktur übliche Verhältnis von Basis und Kollokator erhalten.
- zwei Kollokationen können sich zu einer Tripel-Struktur verbinden (*scharfe* ***Kritik*** *üben*) (vgl. Kapitel 2.3.3).
- eine Teilmenge der deutschen Nominalkomposita ist nun als Kollokation interpretierbar. Handelt es sich um die Präzisierung eines bestimmten Typs oder einer Existenzform, sind die Kollokatoren im Falle der Übersetzung häufig unvorhersehbar (*Schiebe****dach*** *-* ***toit*** *ouvrant (öffnend)*, ***Wort****schwall - flot de* ***paroles*** (*Flut*)).

In einer aktuellen Gliederung Hausmanns (Abb. 5) aus dem Artikel "Lexicographie française et phraseologie" (2005) rückt die ganze Skala der Idiome näher an die Kollokationen als in der Darstellung von 1984 (Abb. 4). Die erste Ebene der Unterscheidung setzt wiederum die freie gegen die kodierte Kombinatorik der Sprache, doch sind die Kollokationen jetzt im Bereich der fixierten Wortkombinationen der Sprache enthalten. Kollokationen werden innerhalb eines differenzierten terminologischen Systems der phraseologischen Verbindungen dargestellt:

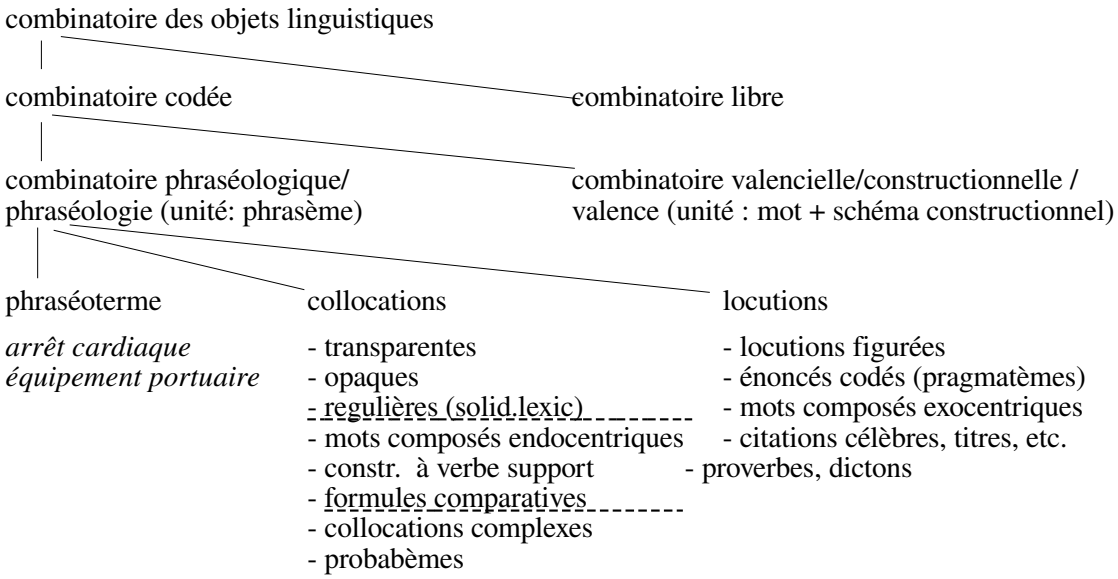


Abb. 5: Kombinatorik linguistischer Objekte (Hausmann 2005: 8)

Die Abgrenzung der freien von der kodierten Kombinatorik ist in Hausmann (2005) eine Frage der Semiotaxis. Wörter "qu'on peut définire, traduiere et apprendre sans contexte" sind semiotaktisch autonom, eine freie Wortkombination besteht aus zwei semiotaktisch autonomen Elementen. Hausmann legte 1997 seine Theorie der "Semiotaxis im Wörterbuch" dar. Die Probleme, die sich aus dem folgenden Semiotaxis-Konzept ergeben, werden erst in Kapitel 3.2.1 näher diskutiert:

> Semiotaxis ist eine semantische Dimension der Syntagmatik. Sie geht davon aus, daß die Wörter semantisch gesehen nicht alle gleichermaßen autonom sind. Sie trennt demnach den Wortschatz in Synsemantika und Autosemantika, setzt aber die Schnittstelle anders an als das üblicherweise geschieht. Üblicherweise betrachtet man alle Substantive, Verben und Adjektive als Autosemantika, und als Synsemantika lediglich inhaltslose oder schwer definierbare Funktionswörter vom Typ Konjunktion u.ä. In der Semiotaxis hingegen geht es um die Definierbarkeit der Wörter. Ist das Wort autonom definierbar oder bedarf es zur Definition eines Kontexpartners, der ihm recht eigentlich erst Identität gibt?
>
> (Hausmann 1997: 172-173)

Hausmann unterscheidet auf einer zweiten Ebene innerhalb der kodierten Kombinatorik zwischen phraseologischen Einheiten und den Kombinationen, die auf der Valenz eines Wortes beruhen. Valenzabhängig sind Satzstruktur und Präpositionen. Die Wahl der Präposition erfolgt idiomatisch und ist daher unvorhersehbar. Präpositionen geben bei Hausmann eine syntaktische Beziehung zwischen den Lexemen wieder, sowohl zwischen den freien als auch den fixierten Kombinationen.

Bei Burger (1998: 15-16) gehört jede feste Kombination von zwei Wörtern zur Phraseologie. Er bezeichnet auch Kombinationen aus zwei Synsemantika (*an sich, im Nu, so dass*) als Phraseme, da er keine plausiblen Kriterien für die eine oder andere Entscheidung sieht, "ob es sich dabei um "Autosemantika" (wie *Öl, geben*) und/oder "Synsemantika" (wie *an, und*) handeln soll". Doch auch Burger zählt Kombinationen mit valenzabhängigen Präpositionen nicht zum Gegenstandsbereich der Phraseologie. Das *Oxford Dictionary of Collocations* bietet eine weitere Lösung an: Präpositionen in Kombination mit mehreren Lexemen sind nur als struktureller Faktor relevant, zusätzlich gibt es aber die Strukturtypen P+N, N+P, V+P, A+P und entsprechende Einträge.
Auf einer dritten Ebene findet bei Hausmann die Klassifizierung der Phraseme statt: Phraseoterme stehen neben Kollokationen und Lokutionen. Phraseoterme sind Phraseme der Fachsprache, sie sind an ganz bestimmte Situationen und Bereiche gebunden. Es sind die äußeren Umstände, die zum Gebrauch einer Formel zwingen, und institutionalisierte Kontexte sind der Grund für einen normativen Phrasemgebrauch als Beweis von Sprach- und Fachkompetenz des Sprechers. Üblicherweise werden sie nicht getrennt von den Kollokationen behandelt, sondern als fachsprachliche Kollokationen bezeichnet (Heid/Freibott 1990, Caro Cedillo 2004). Hausmann grenzt unter anderem die im Schaubild aufgeführten Beispiele der Phraseoterme von den "terminologischen Kollokationen" ab, denn das semantische Gewicht der Adjektive verhindert in den Phraseotermen eine Analyse derselben als semiotaktisch abhängiger Partner des Substantivs und damit eine Interpretation als Kollokator.
Die Lokutionen unterscheiden sich von den beiden anderen Phrasemtypen, weil die Bedeutung keines der beteiligten Wörter in die Bedeutungsdefinition der gesamten Kombination mit eingeht. Sie haben weder Basis noch Kollokator, sondern nur eine Gesamtbedeutung. Zu den Lokutionen gehören die Idiome (*locutions figurées*), die Routineformeln (*énoncés codés, pragmatèmes*), Sprichwörter und ähnliches (*citations célèbres, titres, proverbes, dictons*).
Unter den Lokutionen befinden sich auch die exozentrisch determinierten Wortkombinationen. Endozentrisch und exozentrisch determinierte Syntagmen werden von Rothkegel (1973: 24-31) näher erläutert. Sie unterscheidet syntagmatische Verbindungen nach dem Prinzip der partiellen und kompletten Inklusion in: 1. variable Syntagmen, die sowohl Teilklasse von A als auch von B sind ($AB \in A \wedge AB \in B$ - *blinder Vogel*), 2. endozentrische Syntagmen, bei denen ein Lexem A in einer speziellen Bedeutung ausschließlich Kontextpartner von einem bestimmten Lexem B ist und eine komplette Inklusion einer Klasse in eine andere gilt ($AB \in B$ - *blinder Passagier*), und 3. exozentrische Syntagmen, bei denen komplette Inklusion beider Klassen A und B in eine neue Klasse C vorliegt ($AB \in C$ - *den Kopf verlieren*).
Die endozentrischen syntagmatischen Verbindungen sind bei Hausmann in der Klasse der Kollokationen zu finden neben einer ganzen Reihe weiterer Begriffe. Die Termini transparente, opake und reguläre Kollokation und deren Definition übernimmt Hausmann von Grossmann/Tutin (vgl. unten). Beispiele für komplexe Kollokationen und komparative Formeln wurden schon weiter oben gegeben. Probabeme sind polylexikale nichtlexikalisierte Einheiten, die ein Sprecher mit einer gewissen Wahrscheinlichkeit verwendet, um standardisierte Konzepte auszudrücken. Sie sind weder ganz frei, noch komplexe Kollokationen, können zwei Wörter oder einen ganzen Satz umfassen: *en fin d'après midi* (**am Ende des Nachmittags / am späten Nachmittag*), oder *que ceci ne se reproduise pas je*

vous prie! und nicht *ceci ne doit plus se reproduire!*. Siepmann 2005 untersucht die scheinbar freien Wortkombinationen mit ungewöhnlicher Kollokationsstruktur genauer.
Im Gegensatz zu den Probabemen fällt die Struktur der Funktionsverbgefüge sehr einheitlich aus, es handelt sich um Substantiv-Verb Kombinationen der Typen V+N (*Angst haben*) oder V+Präp+N (*zur Verfügung stellen*). Die Funktionsverbgefüge wurden in der Linguistik ausführlich diskutiert[38], die unstrittigen Definitionskriterien fasst Detges (1994:4) zusammen: als FVG bezeichnet man formal komplexe, feste und halbfeste Prädikatsausdrücke im Grenzbereich zwischen Syntax und Lexikon. Ihre Verbalkonstituenten sind - verglichen mit den jeweils gleichlautenden Vollverben - semantisch weitgehend "inhaltsarm" in dem Sinne, dass sie nur mehr satzkonstitutive grammatische Funktionen wahrnehmen, im wesentlichen also die Markierung von Numerus, Tempus, Person, Modus und *Genus verbi*. Funktionsverb und Substantiv erfüllen gemeinsam die für lexikalische Vollverben oder Adjektivprädikate charakteristische Funktion des Satzprädikats.
Das Nomen ist der Hauptträger der prädikativen Semantik, die Funktionsverben steuern allgemeine semantische Kriterien wie Aspekt und Aktionsart bei. Es können semantische Nuancen ausgedrückt werden, die das entsprechende Verb nicht beinhaltet (*aufschreien* kann im Portugiesischen nur mit FVG übersetzt werden: *dar/soltar um grito* 'einen Schrei *geben/ausstoßen/loslassen'). Auch die Valenz- und Rektionseigenschaften werden durch die Substantive bestimmt und die Selektionsrestriktionen, die beide Mitspieler betreffen, gehen nicht vom Funktionsverb, sondern vom Nomen aus. Die Nomina sind keine Aktanten der Funktionsverben, sondern haben als Prädikatskerne selbst prädikative Funktion.[39] Zur Erkennung von FVG wird häufig ein Kriterienkatalog genannt (fehlende Anaphorisierbarkeit und Erfragbarkeit des Nomens, keine Passivierbarkeit, *Nomen actionis*, Restriktionen in der Singular- und Pluraltransformation, ...), doch sind zu jedem der Kriterien zahlreiche Gegenbeispiele zu finden.[40] Mit vielen der verbal-semantisch reicheren Substantiv-Verb Kollokation haben die FVG folgende Eigenschaften gemeinsam: FVG sind usuell fixiert, da sie syntaktisch-distributionelle Restriktionen aufweisen (wie die Artikelfähigkeit und Attribuierbarkeit), lexikalisiert sind sie auch insofern, "als viele Konstruktionen, die nach Maßgabe systematischer, semantischer und syntaktischer Gegebenheiten virtuell möglich scheinen, in der Norm des Französischen nicht attestiert sind" (Detges 1994: 61).
Die Situierung der Kollokationen innerhalb der Phraseologie in Hausmann (2005) unterstreicht den Charakter der Kombinationen als "Fertigprodukte" der Sprache, die als lexikalische Verbindungen zu lernen und zu systematisieren sind.

3.1.4. Grossmann/Tutin

Auch Grossmann und Tutin (2003: "Quelques pistes pour le traitement des collocations") situieren die Kollokationen in der Klasse der "vorgefertigten lexikalischen Einheiten". Die "phrastischen" und "propositionalen lexikalischen Sequenzen" stimmen in etwa mit den Lokutionen (ohne die Idiome) bei Hausmann überein, von ihnen sind die "syntagmatischen lexikalischen Sequenzen" zu unterscheiden. Grossmann/Tutin führen die Terminologie von

38 Einen ausführlichen Überblick über die Forschungsgeschichte geben Detges (1994) und Yuan (1986).
39 Über die syntaktischen Eigenschaften der FVG gibt es auch abweichende Ansichten. Die vorgestellte syntaktische Situierung nimmt Detges (1994: 14-18) vor.
40 Aufgrund der besonderen Eigenschaften der FVG wurden in der Maschinellen Sprachverarbeitung spezielle Akquisitionsverfahren entwickelt, die Kombinationen bestehend aus Vollverb und Substantiv von den Kombinationen automatisch trennen, in denen das Verb in seiner polysemen Lesart als Funktionsverb auftritt (vgl. Rothkegel (1969), Grefenstette/Teufel (1995), Dras/Johnson (1996)).

Charles Bally (1901) wieder ein, der innerhalb dieser Gruppe die "phraseologischen Einheiten" von den "usuellen Gruppierungen" trennt. Die phraseologischen Einheiten umfassen die *locutions figées opaques* (*cordon bleu*) und die *locutions figées imagées*, die den figurativen Idiomen bei Cowie entsprechen. Die usuellen Gruppierungen sind die heutigen Kollokationen, "halb-erstarrte binäre Ausdrücke" mit einer Basis und einem *collocatif*. Die Basen behalten ihre ursprüngliche Bedeutung, aber anhand der Idiosynkrasie und Idiomatizität des Kollokators kann man drei Typen von Kollokationen unterscheiden:

1. opak: unvorhersehbar, semantisch unmotiviert, schwer zu dekodieren (*peur bleue*, *nuit blanche*)
2. transparent: leicht verstehbar, unter lexikalischem und/oder syntaktischem Gesichtspunkt unvorhersehbar (*avoir faim*, *prendre peur*, *gravement/?grièvement malade*, ? *gravement/grièvement blessé*)
3. regulär: herleitbar, vorhersehbar: die Assoziationsregeln sind mitunter komplex

(Grossmann/Tutin 2003: 8)

Zum einen fällt auf, dass Grossmann/Tutin den idiosynkratischen Charakter einer Kollokation nicht sprachkontrastiv bestimmen, sondern von der Wahlmöglichkeit zwischen mehreren ähnlichen Lexemen oder syntaktischen Strukturen innerhalb einer Sprache sprechen. Als syntaktische Wahlmöglichkeit wird hier das Nicht-Erscheinen des Artikels im Funktionsverbgefüge bezeichnet. Usuell fixierte morphosyntaktische Präferenzen der Kollokationen bilden ein Abgrenzungskriterium zu den freien Wortkombinationen.

3.1.5. Konzeptuelle Kollokationen (Heid)

Unter den regulären Kollokationen werden von Hausmann (2005: 4) die lexikalischen Solidaritäten genannt. Der Ausdruck 'lexikalische Solidaritäten' stammt von Coseriu (1967), er behandelt die semantische Vereinbarkeit von Lexemen: Affinität liegt vor, wenn sich ein Lexem mit einer paradigmatischen Klasse verbindet, deren Einheiten durch einen gemeinsamen inhaltsunterscheidenden Zug zusammenhängen (*lachen* +human); bei der Selektion kombiniert das Lexem mit einem Wortfeld, welches durch ein (nicht immer sprachlich manifestiertes) Archilexem definiert ist, die betreffenden Substantive im Wortfeld unterscheiden sich bezüglich eines Merkmals (*Rind* = Archilexem von *Kuh, Ochse, Stier, Bulle, Kalb*); Implikation liegt vor, wenn es sich um die Verbindung von nur zwei Lexemen handelt, bei denen das Verb das Objekt impliziert (*bellen* +Hund). Bei Coseriu kann eine lexikalische Solidarität "als inhaltliche Bestimmung eines Wortes durch eine Klasse, ein Archilexem oder ein Lexem definiert werden" (Coseriu 1967: 299). Innerhalb dieser vielfältigen Möglichkeiten von lexikalischen Solidaritäten werden als Beispiele von Hausmann für Kollokationen nur Coserius Implikationen genannt.

Grossmann/Tutin beschreiben den Bereich der regulären Kollokationen genauer und situieren die regulären Kollokationen eigentlich an der Grenze zu den semantischen Selektionsrestriktionen.[41] Sie sind sich des Widerspruchs der semantischen Regelmäßigkeit

41 "Elles se situent donc à la frontière des restrictions de sélection sémantique" (Grossmann/Tutin 2003: 8).Chomsky nennt die semantische Beschreibbarkeit der Verbindung von Lexemen 'Selekektions-restriktionen'. Seine Theorie basiert wie die lexikalischen Solidaritäten auf der Annahme, dass sich die Bedeutung eines Lexems anhand von Bedeutungsmerkmalen analysieren lässt. Bei den Selektions-restriktionen fällt jedoch weniger die positive Implikation in Form einer möglichen paradigmatischen Reihenbildung ins Gewicht, sondern die Beschränktheit der Kombinierbarkeit verschiedener Lexeme aufgrund differenzierter semantischer Bedeutungsmerkmale. (Eine Zusammenfassung der Theorie der Selektionsrestriktionen ist bei Klotz (2000: 101-104) zu finden.)

und der Idiosynkrasie bewusst und beschreiben reguläre Kollokationen anhand von Beispielen, bei denen die regelhafte semantische Verbindung nicht sofort ersichtlich ist. So bezeichnen verschiedene Adjektivgruppen die Intensivierung abstrakter Substantive. Ist der Grad der Intensivierung quantitativ ausdrückbar (*désespoir* > *tristesse*) kommen *grand, profond, absolu, extrême* in Betracht. Variiert ein abstraktes Substantiv hingegen qualitativ (*opinion*) kombiniert es mit *bonne, haute, mauvaise, médiocre* (2003: 15).

Während Grossmann/Tutin die regulären Kollokationen explizit benennen, möchten L'Homme/Bertrand die "spezialisierten lexikalischen Kombinationen" von den Kollokationen trennen "since many lexemes defined as co-occurrents can combine with groups of semantically-related terms" (2000: 498). Als Beispiel werden die Kollokationen mit dem Verb *install g*enannt, die kombinierenden Substantive sind Terme, die Software bezeichnen.

Eine ausgeprägte Neigung der fachsprachlichen Kollokate mit semantisch motivierten Klassen von Substantiven zu kombinieren wurde auch in Heid (1994) festgestellt. Die Korrelation von semantischen Klassen und Kollokationsverhalten in der Untersuchung der Gefühlssubstantive bei Mel'čuk und Wanner (1994) (vgl. Kapitel 2.4.1) fasst Heid folgendermaßen zusammen: "The results are of two types: on the one hand indeed, a number of collocations appear with most or all of the elements of the field or of a given subset; on the other hand, a non-negligible amount of exceptions is noted as well" (1994: 238). Da das Domänenmodell von Mel'čuk und Wanner nicht hierarchisch ist, sind es auch die Ergebnisse nicht: "What comes out are rather implications between the presence of certain semantic properties and the collocation behaviour" (Heid 1994: 239). In einer Untersuchung von Meyer/Mackintosh (1994), in der Kollokationsverhalten in der Fachsprache technischer Dokumentationen zu CD-Rom Laufwerken systematisiert wird, stellen sich die Ergebnisse strukturierter dar. Einige Generalisierungen sind möglich, die in einer Hierarchie durch Vererbung ausgedrückt werden können. Diese Eigenschaft korreliert laut Heid mit der taxonomisch einfacheren Modellierbarkeit terminologischer Domänen (insbesondere wenn sie konkrete Objekte behandeln) im Vergleich zu abstrakten Vorstellungen, die in der Allgemeinsprache bestehen (1998: 239).

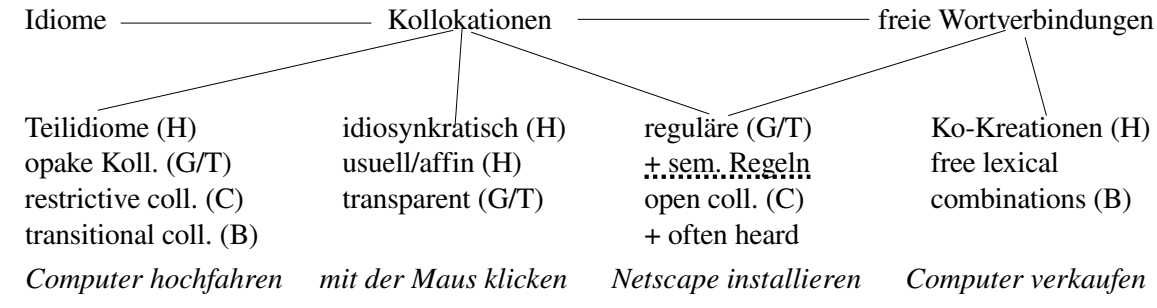
Durch eine Untersuchung von Kollokationen in einem fachsprachlichen Wörterbuch zur Aktienbörse, die sprachlich gesehen eine Mittelstellung zwischen Fach- und Allgemeinsprache einnehmen, und die die Zu- oder Abnahme des vom Nomen denotierten Prozesses ausdrücken, wird deutlich, dass es zum einen Verben gibt, die eine Passepartout-Funktion einnehmen und daher mit allen Nomen kollokieren wie *augmenter*. Andere Verben hingegen sind selektiv, so kombiniert *diminuir* nur mit Substantiven, die als Situation oder Zustand wahrgenommen werden *(charges, concurrence, emprunt, ...*), *freiner, réduire* nur mit Substantiven, die als Aktion aufgefasst werden (*activité, cours, demande, dividende, ...*). Die Substantive werden in ihrer Kombination mit weniger üblichen Verben (*bondir, s'effondrer*) untersucht: "and the sharing of collocations in such small groups of nouns is significant and most likely can be related with properties relevant for the semantic or conceptual description of the group of nouns" (Heid 1994: 241).

Im Gegensatz zu L'Homme/Bertrand, die die spezialisierten lexikalischen Kombinationen aufgrund der Realisierung des "Kookkurrenten" mit Ausdrücken aus derselben semantischen Klasse von den Kollokationen abgrenzen, stellen die Regelmäßigkeiten im Kombinationsverhalten für Heid eine Strukturierungs- und Beschreibungsmöglichkeit des Wortschatzes dar. Die Ergebnisse sind im lexikografischen Bereich pädagogisch verwertbar, und in der

Maschinellen Sprachverarbeitung können semantische und konzeptuelle Definition als Hintergrund für die Wahl eines entsprechenden Kollokats dienen.
L'Homme/Bertrand definieren 'konzeptuelle Kollokationen' " ... in which the co-occurrent combines with several terminological units" und 'lexikalische Kollokationen' " ... in which the co-occurrent combines with a single terminological unit" (2000: 499-500). Sie kommen bei der Untersuchung von technischen und philosophischen Texten zu dem Schluss, dass in beiden Gebieten ca. 14% lexikalische Kollokationen vorliegen. In den 86% konzeptuellen Kollokationen des Fachtextes beziehen sich die Kookkurrenten alle auf Terme derselben semantischen Klasse (über den philosophischen Bereich werden keine Angaben gemacht). Da sich bei L'Homme/Bertrand die "wahren" Kollokationen immer auf zwei Lexeme beziehen, die eine spezifische Bedeutung einnehmen, sind die konzeptuellen Kombinationen in Form von Selektionsrestriktionen zu beschreiben. Diese Vorgehensweise führt aber dazu, dass bei jedem Vorkommen des Kookkurrenten mit mehr als einem Lexem einer bestimmten Wortart, der Kollokationsbereich semantisch zu untersuchen ist, um zu entscheiden, ob die Kollokationspartner zur selben semantischen Klasse gehören. Bei einer entsprechenden Corpusgröße würden dann vermutlich unabhängig vom Register nur noch opake Kollokationen und die impliziten lexikalischen Solidaritäten verbleiben.
Der Terminus 'konzeptuelle Kollokationen' wurde ursprünglich nicht wie bei L'Homme/ Bertrand definiert. Martin (1992) führte bei einer Untersuchung von Kollokationen in Fachsprachen den Begriff 'concept-bound collocations' ein. "According to the author, modifying concepts (i.e. co-occurrents) are often conditioned by some sort of "definitional knowledge" held by the heads (i.e. terms) and are not strictly dictated by usage" (L'Homme/Bertrand 2000: 499). Heid führt den Begriff 'konzeptuelle Kollokationen' ein: "Martin observes a correlation between the semantic and conceptual description of items of a (technical) domain and the collocational behaviour" (1994: 239). Die Annahme über eine Beziehung zwischen Semantik und konzeptuellen Kollokationen bestätigt Heid in der Untersuchung zur Sprache der Aktienbörse.
Konzeptuelle Kollokationen entsprechen den regulären Kollokationen von Grossmann/ Tutin. Der Terminus konzeptuelle Kollokationen ist dem Terminus reguläre Kollokationen vorzuziehen, denn 'reguläre Kollokationen' stellen im Bezug auf das Kollokationskriterium der Idiosynkrasie ein offenes Paradox dar. Konzeptuelle Kollokationen sind semantisch herleitbar, doch sind ihre Relationen mitunter sehr komplex, so dass sie dem Sprecher idiosynkratisch erscheinen. Zu unterscheiden sind Kollokationen, die im Sinne der lexikalischen Solidaritäten eine orientierte Beziehung widerspiegeln. Laut Coseriu handelt es sich um Fälle, bei denen die Bedeutung des Substantivs im Inhalt des Verbs enthalten ist, nicht aber umgekehrt (Coseriu 1967: 296). Dies gilt für einzelne Lexeme oder Lexemklassen, für Ausdrücke der Allgemeinsprache (*fällen* +Baum, *beißen* +Zähne, *galoppieren* +Huftier) und der Fachsprache (*installieren* +Software (im Fachgebiet Informatik)). In den von Grossmann/Tutin und Heid beschriebenen Beispielen liegt eine andere Art der Relation vor. In den Kollokationen *absolut désespoir, bonne opinion, diminuir charges, réduire dividende* ist die Bedeutung der Substantive nicht mit den Kollokaten assoziiert. Die semantisch differenzierte Beschreibung der Substantive erleichtert vielmehr die Wahl der richtigen Kollokate.

3.1.6. Kollokationssystematik

Die folgende Systematik soll die Terminologie der vorgestellten Kollokationskonzepte noch einmal in vergleichender Form verdeutlichen. Darunter befinden sich die verschiedenen Kriterien, nach denen die Kollokationen unterschieden und von den freien Kombinationen abgegrenzt werden.



	opak	transparent		
1 idiomatisch	±	±	± (G/T) + (C)	±
2 usuell	±	±	±	-
3 idiosynkratisch interlingual	+	±	±	-
4 restriktiv (Kollokationsbereich)	-	-	+ (G/T) ± (C)	+
5 semantisch motiviert	+	+	- (G/T) ± (C)	-
6 idiosynkratisch intralingual				

(B) = Benson 1985, *BBI* 1986 (C) = Cowie 1978
(G/T) = Grossmann/Tutin 2003 (H) = Hausmann 1984, 1989, 2004

Abb. 6: Kollokationssystematik und Definitionskriterien

1. Der Faktor der Idiomatizität trennt die opaken Kollokationen von den anderen Kombinationen. Im Gegensatz zu den Idiomen bleibt die Bedeutung der Basis erhalten, der Kollokator gibt eine spezifische Bedeutung wieder, die erheblich von seiner Bedeutung in der Kombination mit anderen Substantiven abweicht. In dieser spezifischen Bedeutung kombiniert der Kollokator nur mit einem oder wenigen Substantiven. Liegt eine opake Kollokation vor, impliziert dies Polysemie beim Kollokator. Cowie unterscheidet außerdem die figurativen Idiome, Kombinationen, die eine Bedeutung als Kollokation (*Feuer fangen, Staub aufwirbeln*) und eine andere als Idiom (*Feuer fangen, Staub aufwirbeln* +fig.) haben.
2. Entgegen der Behauptung von Benson (s.o.) können alle Kollokationstypen und die freien Wortkombinationen usuell (frequent) sein oder auch nicht (vgl. Kapitel 3.2.4).
3. Betrachtet man Idiosynkrasie sprachkontrastiv, können sich alle Kollokationen arbiträr verhalten. Der Einzelfall hängt von den miteinander verglichenen Sprachen ab. Auch bei den opaken Kollokationen (oder Idiomen) kann zufälligerweise ein wörtliches Übersetzungsäquivalent vorliegen. Im Falle der regulär gebildeten Kollokationen ist sich der

Fremdsprachenlerner der Regelmäßigkeit im Kollokationsverhalten meist nicht bewusst, sie stellen sich für ihn daher idiosynkratisch dar - sowohl die restriktiven Kollokationen (*rotes Haar - cheveu roux/*rouge*) als auch die Kollokationen mit einem weiten Kollokationsbereich (*große Hitze - intense heat, große Kälte - severe cold, großer Verlust - heavy loss, großer Junge - big boy, große Person - tall person*).

4. Opake Kollokationen verhalten sich immer restriktiv. Bei den transparenten Kollokationen kann der Kollokator nur für eine Basis gebräuchlich sein (*acalentar esperança* (**Hoffnung aufwärmen*)) oder einen weiten Kollokationsbereich umfassen (*alimentar esperança/ódio/medo/...* (*Hoffnung/Hass/Angst nähren*)). Die regulären Kollokationen können sich ebenfalls sehr restriktiv verhalten (*Hunde bellen, blondes Haar*), oder einen sehr großen Kollokationsbereich haben (vgl. oben g*ravement*).
5. + 6. Semantische Motiviertheit und intralinguale Idionsynkrasie schließen sich aus. Diesbezüglich stellt sich viel mehr die Frage, ob nicht etliche der arbiträr erscheinenden Kollokationen auch auf semantischen Regelmäßigkeiten beruhen, die sich auf den ersten Blick nicht offenbaren, und die bisher noch nicht beschrieben sind. Hausmann definierte schon (1984: 398) die Ko-Kreationen als Kombinationen, die sich "entsprechend gewissen semantischen Mindestregeln" verbinden - die Kollokationen hingegen als Kombinationen, die sich "entsprechend differenzierten semantischen Regeln" verbinden, jedoch ohne diese These durch Beispiele zu belegen.

3.2. Präsentation von Kollokationen in Wörterbüchern

3.2.1. Der Ort der Kollokation im Wörterbuch

Die Frage der Eintragung der Kollokation im Wörterbuch unter einem der Kollokationsbestandteile steht in Abhängigkeit zur inneren Struktur der Kollokation und der Konzeption des Wörterbuchs. Im Britischen Kontextualsimus konnte ein beliebiges Wort als *node* gewählt werden, es bestimmte positionell die Kollokate[42] (vgl. Kapitel 1). In der Einleitung des *BBI* (1986) werden sechs verschiedene Strukturtypen lexikalischer Kollokationen bestimmt (V+N, N+V, N+N, N+Adj, V+Adv, Adj+Adv), neben acht grammatischen Kollokationstypen, die aus einem lexikalischen Element und einer Präposition bzw. einer grammatischen Struktur bestehen. Hausmann (1989) beschränkt Kollokationen auf die Kombinationen mit zwei Elementen aus den offenen lexikalischen Klassen, relevant wird die innere Beziehung der binären Strukturen, das Verhältnis von Basis und Kollokator.
Das Substantiv ist die wichtigste Basiswortart, weil es die Dinge und Phänomene dieser Welt benennt, über die es etwas zu sagen gilt, Adjektive und Verben kommen als Basiswörter nur insoweit in Frage, als sie durch Adverbien weiter determiniert werden können (Hausmann 1985: 119). Die idiosynkratische Wahl des Kollokators ist von der Basis abhängig. Die Kollokatoren werden von der Basis auch insofern determiniert, als dass für deren Bedeutungsdefinition auf die Basis zurückgegriffen werden muss: "la definition des collocatifs est incomplete sans la dimension syntagmatique des collocations" (Hausmann 1979: 192). Zur Definition der Basis braucht man die Kollokation nicht, die Basis ist "semantisch autonom und somit ko-kreativ" (Hausmann 1984: 401). In neueren Arbeiten bezeichnet Hausmann (1997, 2005) diese Beziehung als Semiotaxis, die Basis als autosemantisch und den Kollokator als synsemantisch.

42 Von Jones und Sinclair wurden 1973 je zwei Wörter links und rechts des Artikels *the* untersucht.

Hausmann (2005) unterscheidet Wörterbücher, die onomasiologisch konzipiert sind, und dem Produzenten bei der Wahl des Kollokators helfen - der Kollokator ist dann unter dem Artikel der Basis einzutragen -, und Wörterbücher, die den semasiologischen Gesichtspunkt unterstützen - sie erklären den Kollokator, daher ist die Kollokation beim Kollokator zu verzeichnen mit einem Verweis auf die Basis. Die früheren Überlegungen Hausmanns zum Ort der Eintragung der Kollokationen im Wörterbuch beziehen sich auf die mögliche Ausrichtung der Wörterbücher auf die Textproduktion oder Textrezeption. Zu Produktions- und Rezeptionsschwierigkeiten kann es auch beim Muttersprachler kommen, wenn die passenden Worte fehlen oder bestimmte (evtl. pragmatisch markierte) Ausdrücke nicht verstanden werden.

Hausmann formuliert mit Bezug auf die Verortung der Kollokation im Wörterbuch verschiedene sich mitunter widersprechende Thesen. Zum einen nimmt Hausmann den Standpunkt ein, dass in einem einsprachigen Wörterbuch generell eine Kollokation nur unter der Basis zu führen sei (2004, 1997). Impliziert wird hierbei offenbar, dass der Lerner das Wörterbuch nur zur freien Textproduktion verwendet, denn dann ist der Eintrag unter dem Kollokator nutzlos, der Sprachproduzent würde ihn dort nicht finden, er hat die Bedeutung der Basis vor Augen und parat (Hausmann 2004: 311- 312), "jegliches Formulieren geht von der Basis zum Kollokator und nicht umgekehrt" (Hausmann 1984: 401). Bei der Rezeption hingegen reicht der Eintrag der Kollokation im Artikel des Kollokators. Die Kollokation erweist sich für das Verständnis des Kollokators als grundlegend, dieser kann ohne Bezug auf die Basis gar nicht definiert werden (Hausmann 1985: 121). Bei der Rezeption ist der Gebrauch der Basis immer banal und daher transparent (1988: 150). Wenn ein Wörterbuch beide Funktionen erfüllen will, muss die Kollokation logischerweise zweimal eingetragen werden (Hausmann 1985: 122).

Die einsprachigen Wörterbücher sind oft allgemein als Lernerwörterbücher konzipiert und dienen daher dem Fremdsprachenlerner bei Rezeptions- und Produktionsproblemen. Hausmanns Kritik an verschiedenen Universal- , Stil- und Lernerwörterbüchern des Deutschen (2004, 1985) richtet sich gegen die Lücke, die sich bei der Textproduktion durch den Eintrag der Kollokation nur unter dem Kollokator ergibt. Bahns (1996: 79) zeigt anhand einer genauen Untersuchung der Kollokationspraxis in einsprachigen englischen Wörterbüchern, dass die Kollokationen fast doppelt so häufig unter dem Kollokator verzeichnet sind wie unter der Basis (die Zahl der Doppeleinträge variiert zwischen 30-50%). Wie Hausmann schon 1979 ausführte ist der Grund dafür, dass die Kollokation beim Kollokator zur Definition dazugehört (vgl. oben). Allgemeine einsprachige Wörterbücher legen offenbar die Betonung auf die Probleme der Sprachrezeption.

Kollokationswörterbücher sind als Nachschlagewerke bei der Textproduktion konzipiert. Spezielle Kollokationswörterbücher sollten sich auf "spezifische Kombinationen, d.h. auf wirkliche Kollokationen beschränken, sie sollten sich auch darauf beschränken, die Kollokationen ausschließlich unter dem Basiswort einzutragen" (Hausmann 1985: 123). Auch Herbst und Klotz (2003: 85) vertreten die Ansicht, dass eine Kollokation für Produktionszwecke im einsprachigen Wörterbuch unbedingt unter der Basis im Wörterbuch zu führen ist, denn man geht davon aus, dass die Basis den Benutzern bekannt ist, sie suchen im Wörterbuch nach dem Kollokator. In diesem Sinne sprechen sie von einer Verbesserung, wenn in neu aufgelegten einsprachigen Wörterbüchern mehr Kollokationen unter der Basis zu finden sind.

Im bilingualen Bereich unterscheidet man zwischen "aktivem" und "passivem" Gebrauch der Wörterbücher. Für den aktiven Gebrauch des Wörterbuchs - für das Übersetzen in die Fremdsprache - braucht der Benutzer mehr Informationen, hier fehlen ihm die adäquaten Wortkombinationen und Kenntnisse ihrer (morpho)syntaktischen Eigenschaften. Bei der Übersetzung in die Muttersprache fallen dem Produzenten die passenden Äquivalente wegen ihrer Rekurrenz und (häufigen) semantischen Transparenz eher ein. Zweisprachige Wörterbücher sind vor allem in Hinsicht auf einen fortgeschrittenen Fremdsprachenkundigen produktions- oder rezeptionsorientiert ausgerichtet. Viele der zweisprachigen Wörterbücher werden jedoch ungeachtet einer potentiell getrennten Zielgruppe verfasst, was am häufigsten die weniger ausführlichen und diejenigen, die beide Sprachrichtungen aufführen, betrifft.

Hausmann bemerkt, dass das passive zweisprachige Wörterbuch sich damit begnügen kann, die Kollokationen unter dem Kollokator einzutragen, denn nur dort kann es zu Verständnisschwierigkeiten kommen, das Basiswort ist banal. Bei der Produktion im zweisprachigen Wörterbuch ist der Eintrag der Kollokation unter Basis und Kollokator wünschenswert. Der Kollokator ist das schwierigere Wort, doch wird der Textproduzent für die aktive Hinübersetzung das Basiswort sehr viel öfter nachschlagen, und es begrüßen wenn ihm im Basisartikel sprachliches Material zur Kontextualisierung angeboten wird (Hausmann 1988: 150). Als generelle Regel für ein zweisprachiges Wörterbuch gibt Hausmann den Rat, die Sichtweise des Produzenten zu betonen, denn die Dekodierung der Kollokationen macht keine großen Probleme (2005: 3).

Im Hinblick auf den aktiven zweisprachigen Wörterbuchgebrauch scheidet bei Herbst und Klotz die Eintragung der Kollokation beim Kollokator in der Fremdsprache aus, denn wäre dieser bekannt, müsste man nicht nachschlagen. Für den fremdsprachigen aktiven Benutzer hingegen wäre genau diese Stelle durch die äquivalenzdifferenzierende Funktion von Kollokationen beim Kollokator relevant (2003: 141-142). Es gibt in einem zweisprachigen Wörterbuch, das beide Sprachrichtungen behandelt, für eine Kollokation vier verschiedene Stellen für mögliche Eintragungen. Um ein unmäßiges Anwachsen des Wörterbuchs durch das Aufführen der Kollokationen an den vier Stellen zu vermeiden, wäre auch ein Verweissystem denkbar, das sich aber durch das mehrmalige Nachschlagen nicht unbedingt benutzerfreundlich gestaltet.[43]

2004 erschien eine Dissertation im bilingualen Bereich von Ekatarina Butina-Koller: *Kollokationen im zweisprachigen Wörterbuch. Zur Behandlung lexikalischer Kollokationen in allgemeinsprachlichen Wörterbüchern des Sprachenpaares Französisch/Russisch.* Beschrieben werden drei umfangreiche zweisprachige Wörterbücher, die jeweils nur eine Sprachrichtung behandeln, und deren Bearbeitung der Kollokationen in Makro- und Mikrostruktur. Zwei der Wörterbücher sind produktionsorientiert konzipiert, eines für die Sprachrezeption bestimmt. Bei Butina-Koller (2004: 31-32) ergibt sich bezüglich des Orts der Kollokation im zweisprachigen Wörterbuch ein klares Bild. In einem rein rezeptionsorientierten Wörterbuch kann es genügen, wenn die Kollokationen nur im Artikel zum Kollokator erfasst werden. In produktionsorientierten zweisprachigen Wörterbüchern ist die Kollokation doppelt zu verzeichnen (zur Not mit Hilfe von Querverweisen), denn bei einem zu übersetzenden Text würde man im Wörterbuch unter dem fehlenden Kollokator nachschauen. Nur in der freien Textproduktion kann es passieren, dass man den Kollokator auch

43 Als Beispiele für Wörterbuchartikel mit Verweissystem stellen Herbst/Klotz (2003: 142) Einträge aus dem *Duden Oxford Großwörterbuch Englisch. Deutsch-English / Englisch-Deutsch* (1999) dar.

in der Muttersprache nicht parat hat, dann muss man unter dem Artikel zum Basiswort nach dem passenden Äquivalent suchen. Butina-Koller stellt bei einer statistischen Erhebung fest, dass relativ unabhängig von der Funktion des Wörterbuchs die Eintragung der Kollokationen etwas häufiger unter dem Kollokator zu finden ist als unter der Basis und circa 30% der Kollokationen doppelt verzeichnet sind.[44]

Bei der Rezeption einer Kollokation machen genau jene Kombinationen Probleme, bei denen der Kollokator in einer polysemen Lesart seiner "normalen" Bedeutung (d.h. außerhalb seines üblichen Kontextes) auftritt, und die äquivalente Bedeutung in der Muttersprache mit einem anderen Verb wiedergegeben wird. Das Verständnis wird eventuell erleichtert, wenn das Verb mit einer Reihe von Substantiven in der polysemen Lesart kollokiert. Mit der wörtlichen Übersetzung von *catch the train* mit **den Zug fangen* wäre nur wenig anzufangen, würde man die spezifische Bedeutung von *catch* nicht schon in der Kombination mit anderen Transportmitteln kennen. Gerade restriktive Kollokationen wie *catch fire* sind schwierig zu interpretieren, hat man in der Muttersprache mit *Feuer fangen* nicht zufällig das gleiche Bild (im Französischen *prendre feu* - 'Feuer *nehmen').
Da es sich bei solchen Beispielen immer um eine polyseme Lesart des Verbs handelt, muss die spezifische Bedeutung definiert werden. Hausmann betont, dass eine vollständige Definition des Kollokators ohne die Kollokation gar nicht möglich ist. Für ein Wort wie *Gefängnis* kann man ohne weiteren Ko-text mit dem Definieren der Bedeutung anfangen, für ein Wort wie *aufwerfen* muss man sich zuerst den Ko-text vorstellen (*Frage, Damm*): *Gefängnis* ist autonom, *aufwerfen* ko-textabhängig[45] (Hausmann 1997: 172). Die Autosemantika sind sehr wohl ohne Angabe des Ko-textes, nicht aber ohne Angabe des (außersprachlichen) Kontextes definierbar, so kann man Rezept anhand des Kontextes 'medizinisch/kulinarisch' unterscheiden (Hausmann 1997: 176). Benson hingegen hält gerade die Aufnahme von freien Wortkombinationen für unumgänglich "... to illustrate a sense of a polysemous entry in a general-purpose dictionary" (*BBI* 1986: ix). In Kapitel 3.4. wird deutlich, dass es meistens nicht die freien Wortkombinationen sind, die auch polyseme Substantive charakterisieren, sondern gerade die Kombinationen, in denen die Verben als Kollokate auftreten.
Es gibt auch Verben mit nur nur einer Lesart wie *hoffen.* Verben scheinen im Gegensatz zu Substantiven aber sehr viel häufiger mit einer polysemen Lesart vorzuliegen und diese systematischer anhand des sprachlichen Kontextes zu unterscheiden. Die Bedeutung von Kollokaten mit einem sehr weiten Kollokationsbereich wie *putzen* oder *hot* wird auch in synsemantischer Funktion primär durch ihre normale Bedeutung geprägt und ist nicht von einer spezifischen Basis abhängig (*Zähne putzen, hot food*). Die Basis bestimmt lediglich die arbiträre Wahl des Kollokators gegenüber alternativen Möglichkeiten wie **Zähne waschen* und **warm food*. In der Kollokation *Herz gehören* wird das *Herz* in einer übertragenen Bedeutung verwandt als *Liebe*, die Bedeutung des Kollokators bleibt erhalten, er entscheidet über die polyseme Lesart der Basis.

44 Die genauen Frequenztabellen und Diagramme für jedes der untersuchten Wörterbücher werden in Butina-Koller (2004) aufgeführt: 74-77, 97-100. Auch in den hier untersuchten deutsch/portugiesischen Wörterbüchern, die in beiden Sprachrichtungen und für keine spezifische Zielgruppe verfasst sind, bestätigt sich dieses Bild (vgl. Kapitel 6).

45 Hausmann unterscheidet zwischen "Kontext (außersprachlich) und Ko-text (sprachlich)" (1997: 172). Der Begriff 'Ko-text' wird hier nur übernommen wenn er von den zitierten Autoren verwendet wird. Ansonsten bezieht sich 'Kontext' auf die (abstrakte oder konkrete) sprachliche Umgebung, auf die Situation der Äußerung, wird mit "situativem Kontext" oder "Domäne" referiert.

Das Kollokationskonzept von Hausmann (und Benson) hat zu Wörterbüchern geführt, in denen man für die Substantive ausführliche, wohlgeordnete Einträge findet. Für das englische Nomen *honour* werden im *Oxford Collocations Dictionary* definitorisch vier polyseme Lesarten gegeben (1 *sth that makes you feel proud*, 2 *great respect*, 3 *good reputation*, 4 *award/official title*). Jede polyseme Lesart enthält Kollokate, die sich von den Kollokaten der anderen Lesarten unterscheiden, sie sind komplementär verteilt: "Definitions of head-words are given only insofar as they are necessary to distinguish different senses of the same word, when they have different collocations and need to be treated separately" (2002: x).
Für das Adjektiv *hot* werden fünf polyseme Lesarten gegeben (1 *of the weather*, 2 *of a person*, 3 *of a thing*, 4 *of food*: *not cold*, 5 *of food: spicy*). In der Definition wird der Kontext durch ein Archilexem benannt, als Kollokate werden Verben und Adverbien gezeigt[46]. Doch im Gegensatz zu den Kollokaten bei den polysemen Substantiven werden beim Adjektiv etliche Adverbien und Verben in den monosemen Einträgen fünf mal genannt (*a bit, quite, rather, really ... / be*). Die bedeutungsunterscheidenden Wörter des Kontextes, die substantivischen Kollokate der Verben oder Adjektive, werden nicht als Kombinationspartner aufgeführt. Sind die Substantiv-Kollokate beim Verb nicht verzeichnet, ist die Auflösung polysemer Verbformen über ihren Kontext wie im Eintrag von Cowie für *hammer into* (vgl. 3.1) nicht möglich. Man kann natürlich bei der Bestimmung der polysemen Lesart von *hammer into* - *hämmern/eintrichtern* auf generalisierende Lexeme wie *spitzer Gegenstand* oder semantische Merkmale wie [*Wissensgebiet*] ausweichen.
Würde aber unter den Einträgen der Adjektive oder Verben nicht eher der sprachliche Kontext von Substantiven interessieren als die Mehrfachnennung der meisten Kollokate? Diesbezüglich stellt sich auch die Frage, ob die Determination einer polysemen Verbform durch ganze Reihen von Kollokaten (*catch a cold/flew/aids/...*, *catch the bus/train/boat/...)* nicht sinnreicher beim Kollokator einzutragen wäre. Auch bei der freien Textproduktion ist es durchaus denkbar, dass man beim Verb nachschlägt. Üblicherweise erwartet man im Eintrag der Verben auch Substantive als Kontextangaben. Hätte man als Fremdpsrachenlerner das Verb *hammer into* parat, aber nicht das englische Substantiv für *Nagel* oder *Pfahl,* würde man es auch bei der Textproduktion mit einem einsprachigen Kollokationswörterbuch begrüßen, den substantivischen Kontext im Eintrag der Verben zu finden.
Möglicherweise sind in der lexikografischen Praxis andere Kriterien für den Benutzer von Vorteil als die strikte Umsetzung der Beschreibung der internen Beziehung der Kollokationspartner. Auch häufige Verben (wie *move*) werden im *Oxford Dictionary of Collocations* nicht als Lemma verzeichnet. Sucht man für die Textproduktion beispielsweise nach spezifischen Präpositionen des Verbs könnte man nur hoffen, dass es als Kollokator der Basis (mit einem Beispielsatz) zu finden ist. Ein weiteres Wörterbuch mit präziseren Verbeinträgen ist durch die Trennung der Kollokationen in Basis und Kollokator und der Umsetzung des Konzepts in der Lexikografie bei der Textproduktion immer von Nöten. Man könnte die Informationen auch beim Substantiv geben, da viele Kollokatoren jedoch einen weiten Kollokationsbereich haben, würde dies zu redundanten Einträgen führen.

46 Im *Oxford Dictionary of Collocations* werden folgende Kollokationstypen unterschieden: ADJ+**N**, QUANT+**N**, V+**N**, **N**+V, **N**+N, PREP+**N**, **N**+PREP, ADV+**V**, V+**V**, **V**+PREP, V+**ADJ**, ADV+**ADJ**, **ADJ**+PREP.

3.2.2. Kollokationen in allgemeinsprachlichen Wörterbüchern

Die Behandlung von Kollokationen in allgemeinsprachlichen Wörterbüchern bietet die Grundlage für eine Vielzahl von Untersuchungen einsprachiger deutscher und englischer Wörterbücher (Bahns (1997), Hausmann (1985, 2004), Heid (2004), Köster/Neubauer (2002), Lehr (1998)). Kritisiert wird die Darstellungsform von Kollokationen innerhalb eines Wörterbuchartikels. Die Kollokationsangaben erscheinen in der Mikrostruktur in den Definitionen und in den Beispielen (mitunter fett hervorgehoben) oder in Querverweisen.[47]

> **medo** (ê). [Do lat. metu] *S.m.* **1.** Sentimento de grande inquietação ante a noção de um perigo real ou imaginário, de uma ameaça; susto, pavor, temor, terror. **2.** V. *Receio* (1 e 2) ◆ **A medo**. Timidamente, hesitantemente: "É que, à noite, lírio branco, /Os astros guardam segredo /Dos beijos dados a medo ...." (Gonçalves Crespo, *Obras Completas*, p. 316) **Não ter medo de caretas**. Não se amedrontar com ameaças. **Pelar-se de medo**. Ter um medo que se péla. **Ter medo da própria sombra**. Assustar-se ou apavorar-se por qualquer motivo. **Ter um medo que se péla**. Ter medo excessivo; pelar-se de modo.
>
> (*Novo Dicionário Aurélio da Língua Portuguesa* 1986)

Wichtig wurden Kollokationen auch im Fremdsprachenunterricht, was die Defizite der Wörterbücher umso deutlicher macht. Die Ergebnisse lexikografischer Corpusanalysen unterstreichen die große Rolle von Kollokationen und Phraseologismen in der Sprache. Das breite Interesse an Kollokationen schlägt sich in neueren allgemeinsprachlichen Wörterbüchern auch in der Mikrostruktur nieder. Im *Longman Dictionary of Contemporary English* (2003) sind die Kollokationen in eigens hervorgehobenen Blöcken in den Einträgen der Basis vertreten sind. Das *Langenscheidts Großwörterbuch Deutsch als Fremdsprache* (2003), hebt die Kollokate in spitzen Klammern deutlich von den anderen Einträgen ab. Die Kollokate werden wie im *OCD* nach Wortarten geordnet und teilweise mit Beispielsätzen unterlegt:

> **Hoffnung** *die; -, -en;* **1 e-e H. (auf** etw. (Akk)) der starke Wunsch od. Glaube, daß etw. geschehen wird < e-e begründete, berechtigte, falsche, schwache H.; sich/j-m Hoffnung(en) machen; in j-m Hoffnung(en) (er)wecken; H. schöpfen; (keine, wenig) H. haben; j-m e-e/die H. nehmen; die H. aufgeben, verlieren > : *es gibt kaum noch H, daß er gesund wird; sie ging voller H. in die Prüfung ...*
>
> (*Langenscheidts Großwörterbuch Deutsch als Fremdsprache* 2003)

Die Präsentation der Kollokationen in der Mikrostruktur der von Butina-Koller untersuchten zweisprachigen Wörterbücher hingegen (vgl. Kapitel 3.2.1) ist in keiner Weise systematisch. Kollokationen erscheinen innerhalb der jeweiligen Einzelbedeutungen der Lemmata, in den Glossen oder in einem separaten Teil, der für Idiome oder phraseologischen Mehr-Wort-Einheiten bestimmt ist (Butina-Koller 2004: 77-92, 100-104). Eine konsequente Sortierung der Kollokationen wird weder alphabetisch noch kategoriell durchgeführt.

Die von Butina-Koller vorgeschlagenen Musterartikel orientieren sich an der Struktur des *OCD*. Die Kollokatoren erscheinen primär sortiert nach ihrer Kategorie, innerhalb einer Kollokationsart nach semantischen Beziehungen. Sind die Kollokatoren alle spezifisch, wird eine alphabetische Reihenfolge gewahrt. An die Stelle eines repräsentativen Beispielsatzes tritt die Übersetzung in der Zielsprache. Vor allem für die Situation der Textproduktion sollte die Form ersichtlich werden, in der eine Kollokation gewöhnlich benutzt wird, sowie

47 Einen kurzen Überblick zu dieser Thematik geben Herbst und Klotz in einem Kapitel über die "Erkennbarkeit von Kollokationsinformation" (Herbst/Klotz 2003: 85-88), Bahns (1996: 37-91) behandelt die Kollokationen in der Mikrostruktur des Wörterbuchartikels ausführlich in Theorie und Praxis.

die Eigenschaften des breiteren Kontextes (wie kopulative Verben und Valenz) und der Artikelgebrauch (Butina-Koller 2004: 138).
In den Musterartikeln sind einige (morpho)syntaktische Merkmale der Übersetzung zu entnehmen. Die Kollokationen erscheinen nach der allgemeinsprachlichen Übersetzung des Lemmas. Im Gegensatz zur Praxis in den meisten Wörterbüchern kommen die Idiome zusammen mit den restriktiven Kollokationen vor, dieser Bereich wird im Wörterbuchartikel durch das Zeichen ☒ markiert. Butina-Koller (2004: 44) begründet diese Vorgehen mit den Schwierigkeiten des Lexikografen bei der Unterscheidung von Kollokationen und Idiomen und der Unmöglichkeit für den Rezipienten der Fremdsprache im Falle des Nichtverstehens einer Kombination zu entscheiden, ob es sich um eine freie Wortverbindung, eine Kollokation oder ein Idiom handelt, um dann von einer nach diesen Kriterien unterteilten Mikrostruktur zu profitieren. Hier ein Musterartikel für ein Verb im aktiven Wörterbuch Russisch-Französisch[48]:

> **СОСТАВ||ИТЬ** ... **1** *(= ...)*: mettre, placer, ranger: ~... ranger la vaiselle dans le buffet; ~ ... : disposer en rangs **2** *(...)* déplacer, mettre: ... elle a déplacé les fleurs du bord de la fenêtre el les a mises par terre **3** *(= ...)* < ... > composer; former: <bouquet *m*, groupe *m*, mot *m* phrase *f* > || *(= ...):* < ... > composer: <solution *f*, mélange *m* > **4** < ... > dresser, établir: < un document: acte *m*, contrat *m*, rapport *m*, procès-verbal *m*; horaire *m*, liste *f*>; < ... > composer, rédiger: <recueil *m*, dictionnaire *m*, texte *m*, manuel *m*>; ~... cette lettre est mal rédigée *(*... mal tournée) ☒ ~ ... faire exception; ~ ...tenir compagnie; ~ ... se former une opinion; ~ ... se faire une idée de qch; ~ ...faire fortune; ~ ... cela fera une grosse somme; ~ ... faire le bonheur de qn; ~ ... cela ne sera guère difficile à faire; cela ne fait aucune difficulté; ~ ... poser une équation
>
> (Butina-Koller 2004: 144)

In den einsprachigen Kollokationswörterbüchern wird die Trennung der Einträge nach Basis und Kollokator streng durchgehalten. In einem zweisprachigen Wörterbuch können auch unter dem Verb ausführliche Angaben zu den Kollokaten stehen.

3.2.3. Informationstypen in (Kollokations)wörterbüchern

Auch für das Portugiesische gibt es ein Kontextwörterbuch, das auf dem Kollokationskonzept von Hausmann basiert, das *Dicionário Contextual Básico da Língua Portuguesa. Portugiesisches Kontextwörterbuch* (Pöll, 2000). Als Lemmata dienen ausschließlich Substantive, als Kollokate werden Adjektive, Substantive und Verben verzeichnet:

> **MEDO** n. m. "sentimento de grande inquietação ante a noção de um perigo (imaginário ou real) ou uma ameaça": ◆ ~ angustioso; ~ atroz, confrangedor ◆ ter ~ (de a.c., alg.): *Tive ~ que surgisse outra calamidade.* estar com, estar cheio de ~ ; sentir, experimentar ~: *Viver é encontrar situações em que experimentemos ~.* estar verde de ~; estar tolhido de ~; tremer de ~; fazer, meter ~ (a alg.): *As coisas que não compreendemos fazem-nos ~. A minha força mete-lhe ~.* provocar, causar ~: *O que é que provocou no João aquele ~ inexplicável?*
>
> (*Dicionário Contextual* 2000)

Während im *Oxford Collocations Dictionary* (2002) genauere Angaben zu grammatischen Konstruktionen oder morphosyntaktischen Präferenzen der Kollokationen nur aus den Beispielsätzen zu erschließen sind (vgl. Wörterbuchausschnitt in Kapitel 1), werden grammatische Angaben im *Dicionário Contextual* von Pöll explizit gemacht sowie in den Beispielsätzen und Kontextelementen verpackt (2000: VI). Auch eine Bedeutungsdefinition des Substantivs wird immer gegeben. Im *BBI Combinatory Dictionary of English* (1986)

48 Die russischen Angaben in kyrillisch sind hier nicht aufgeführt.

wird auf Corpusbelege in Form von ganzen Beispielsätzen verzichtet zugunsten illustrativer Phrasen und eine begrenzte Anzahl von syntaktischen Kontextangaben gemacht. Die im Vergleich zu den beiden anderen Kollokationswörterbüchern ausführlichen Angaben zu syntaktischen Eigenschaften im *BBI* sind dort Ausdruck eines anderen Kollokationsverständnisses, das Präpositionen und Subkategorisierungsinformationen als "grammatische Kollokationen" enthält (vgl. Kapitel 3.1.2). Alle drei Kollokationswörterbücher tragen die Kollokation unter der Basis ein und nehmen eine Partitionierung des Kollokationsbereichs nach syntaktischen und semantischen Kriterien vor.
Für den monolingualen Bereich gibt es detaillierte Beschreibungen der Angaben, die ein Eintrag für eine Kollokation in optimaler Weise enthalten soll. Hier spielen sehr viel mehr Komponenten eine Rolle als die bislang aufgezeigten Wörterbucheinträge vermuten ließen. Heid gibt in dem Artikel "On the presentation of collocations in monolingual dictionaries" (2004) folgende Übersicht für Substantiv-Verb Kollokationen:[49]

Level	Phenomenon	Examples
Lexical Combinatorics	two lexemes open/closed collocate list	*pay+attention* *{etwas/nichts...} halten von*
Morphosyntax	noun: singular/plural modification of N determination of N	*high hopes* *raise a (ADJ) question* *zur* (=definite) *Sprache bringen*
Syntax	verb valency noun valency	*raise* + OBJ (*question*) *cherish the hope that*
Semantics	synonymy	*Vorschlag machen/unterbreiten*
Pragmatics	diasystematic marks frequency in corpus	style, geographic use, ...

Abb. 7: Informationstypen für Substantiv-Verb Kollokationen (Heid 2004:731)

Bezieht man diese ganzen zu einer Kollokation gehörigen Informationen in die Beurteilung neuer Kollokationswörterbücher mit ein, erscheinen auch deren Einträge nur noch suboptimal. Das *Oxford Collocations Dictionary* (2002) gliedert seine Mikrostruktur in vorbildlicher Weise nach der Basis, der Wortart des Kollokators und innerhalb der syntaktischen Gruppe nach semantischen Beziehungen und Synonymen. Eine oder mehrere Kollokationen aus jeder Gruppe werden durch einen Beispielsatz aus dem Corpus begleitet. Eine detaillierte Angabe zu syntaktischen und morphosyntaktischen Eigenschaften einer Kollokation hingegen sucht man vergeblich.
Heid vergleicht in seinem Artikel auch noch ein Wörterbuch des Französischen mit dem Informationskatalog für Kollokationen und bescheinigt diesem eine gute Darstellung bezüglich syntaktischer Subkategorisierung und Valenzstrukturen für alle Kollokationen. Es handelt sich bei dem untersuchten Wörterbuch um das *Dictionnaire combinatoire du français - Expressions, locutions et constructions* (Zinglé/Brobeck-Zinglé 2003):

49 Die Forderung nach der Präsentation von Kollokationen im Kontext wird auch in Heid (2005) anhand von detaillierten Beispielen unterstrichen.

espoir
aimer sans espoir: phr. aimer sans espérer être aimé [...]
caresser un espoir, une espérance: phr. espérer qqch [...]
garder espoir: phr. continuer d' espérer
garder l' espoir de <inf>: phr. continuer d' espérer que <prop>
gonflé d' espoir: qual. qui espère beaucoup [...]
nourrir l' espoir de <inf>, que <prop>: phr. espérer <in>, que <prop>
nourrir l' espoir: phr. continuer d' espérer [...]

(*Dictionnaire combinatoire* 2003 (nach Heid 2004: 733)

Hier wird zu jeder einzelnen Kollokation eine Bedeutungsangabe gemacht, morphosyntaktische Informationen sind dem Eintrag der Kollokation zu entnehmen, und die syntaktische Präzisierung wird in detaillierten formalisierten Valenzmustern gegeben. Die Kollokationen sind alphabetisch geordnet, semantische Klassen von Kollokationen können nur durch einen Vergleich der Bedeutungen ermittelt werden. Was in beiden Wörterbüchern fehlt ist jegliche Angabe von Frequenzdaten und Heid (2004: 736) regt den Gebrauch eines einfachen Annotationsschemas nach dem Vorbild des *Collins COBUILD English Dictionary* (1995) auf einem visuell ansprechenden Frequenzband an.

Einen Vorschlag zur Konzeption von Kollokationseinträgen in zweisprachigen Wörterbüchern legt Zita Hollós 2005 vor. *Lernerlexikographie: syntagmatisch. Konzeption für ein deutsch-ungarisches Lernerwörterbuch* ist der Titel. Hollós verzichtet auf den Begriff 'Kollokationswörterbuch', denn der deutsche Wörterbuchgegenstand, von dem der ungarische abhängig ist, umfasst auf der lexikalischen Ebene außer den usuellen Kollokationen auch freie Kombinationen, diese sind in fremdsprachigen Produktionssituationen unumgänglich (2005: 80). Charakterisiert wird das Wörterbuch, für das die Musterartikel erscheinen als "deutsch-ungarisches, monoskopales, korpusorientiertes, polyfunktionales, dennoch produktionsbezogenes syntagmatisches Lernerwörterbuch mit einem primären Wörterverzeichnis der Basen und einem sekundären der Kollokatoren und zusätzlich mit einem ungarisch-deutschen Register zum Basis-Wörterverzeichnis" (Hollós 2005: 174).
Die Zielgruppe wird in der Charakteristik benannt. Das Kollokationskonzept orientiert sich an der Unterscheidung Hausmanns in Basis und Kollokator. Die Strukturtypen der lexikalischen Kollokationen werden mit Einschränkungen aus dem *BBI* (1986) übernommen, es verbleiben die fünf Kombinationsmöglichkeiten, die mit einem Basiswort beginnen (Hollós 2005: 42). Als Lemmata des Wörterbuchs werden Substantive, Verben, Adjektive und Adverbien verzeichnet. Der Eintrag erfolgt im Sinne eines aktiven Wörterbuchs wie bei Butina-Koller unter der Basis und dem Kollokator. Die Musterartikel für bestimmte Wörter sind nicht als Verbesserungsvorschlag für existierende allgemeine Wörterbücher gedacht, sondern stellen eine "wissenschaftlich fundierte Wörterbuchkonzeption in einer "gut konsumierbaren" Form" dar (Hollós 2005: 2).
Über den Formkommentar zu den verschiedenen Lemmazeichentypen werden recht ausführliche Angaben gemacht (Hollós 2005: 104-140). Ein "semantischer Kommentar" (durch Spitze Klammern gekennzeichnet) kann verschiedene pragmatische und morphologische Informationen enthalten. Die Aktanten des Verbs und deren Kasus werden, wenn sie als Objekte fungieren, im Syntagma vollständig aufgeführt. Konstruktionen, die einen Nebensatz verlangen, werden nicht verzeichnet (vgl. im folgenden Eintrag 'etw^{A} mit S. feststellen', statt 'mit Schrecken feststellen, dass').

Schre|cken der -s, -

❶ <nur Sg> *rémület, ijedség , félelem*
SUBS <selten> Furcht/Angst und S. *ijedség és rémület*
ADJ großer < hefiger *nagy, heves* * gehöriger *jókora* * ↑panischer *páni* || voller S. *f.mel tele*
☝ ***egy kis ijedséggel*** mit leichtem Schreck
VERB ergreift ≈ packt, befällt jn *elfogja* * lähmt jn *megbénítja* * ↑durchzuckt jn *belenyilall vkibe* * etw[A] mit S. feststellen, wahrnehmen *r.tel állapít meg, vesz észre vmit* * sich von dem S. erholen *kiheveri az i.et* * etw[N] erfüllt jn mit S. *r.tel tölti el vmi*
☝ ***megrémül/megijed*** Schrecken bekommen /<gespr> kriegen * ***rémületbe ejt vkit*** jn in (Angst und) Schrecken versetzen * *(széles körben)* ***rémületet kelt*** (<selten> Furcht/ Angst und) Schrecken verbreiten * ***megrémíti/megijeszti*** jm einen Schrecken ↑ einjagen/↑einflößen * ***megszabadítja a félelemtöl*** jm den Schrecken nehmen * ***az ifedségen kívül nem esik baja*** mit dem Schrecken davonkommen * ***csillapodik az ijedség*** (an) Schrecken verlieren
❷ <mst Pl> die Schrecken + Gen *vminek a borzalma(i), rémsége(i)*
des Krieges, des (Zweiten) Weltkrieges *a háború, a (második) világháború* * der Vergangenheit *a múlt* * des Todes *a halál*
SUBS Schönheit und S. *szépség és borzalom*
ADJ reale *valós*

(Hollós 2005: 177-178)

duf|ten duftete, hat geduftet

❶ <nur Sg> etwas[N] <mst Pl> duftet *vmi illatozik*
Blüten, Blumen *virágok* * Rosen, Flieder <Pl> *rózsa, orgona* * Mandeln <Pl> *mandula* * *s* Brot *kenyér*
❷ jemand/etwas[N] duftet nach etwas[D] <o Art> *vkinek/vminek vmilyen illata/szaga van*
Kaffee, Glühwein *kávé, forralt, bor* * Waffeln, Lebkuchen, Plätzchen <Pl> *gofri, mézeskalács, aprósütemény* * Gewürze <Pl>, Zimt, Lavendel *füszerek, fahéj, lavendula* * Parfüm *parfüm* * Bratwürste <Pl> *sült kolbász*
ADV <zu B und C> ↑verlockend *csábító(an)* * ↑köstlich *felséges(en)* * ↑herrlich *csodás(an)* * angenehm *kellemes(en)* * stark *erös(en)*
☝ *ínycsiklandozó(an)* ↑verführerisch

(Hollós 2005: 179)

Die Auswahl der einzutragenden Syntagmen erfolgt unter drei empiriebezogenen Gesichtspunkten. Die Ermittlung der usuellen Kotextpartner geschieht corpusbezogen, die Primärquelle ist das COSMAS-I Corpus[50] des Instituts für Deutsche Sprache. Die Stärke der lexikalischen Kohäsion wird repräsentiert durch den Gamma-Wert, für seine Berechnung wird unter anderem der Likelihood-Ratio-Test verwendet. Die Datenbasis spiegelt die Richtung des Wörterbuchs wieder, es ist deutsch-ungarisch angelegt und beschreibt die Kombinierbarkeit deutscher sprachlicher Einheiten (Hollós 2005: 24). Als Sekundärquelle dienen Einträge von deutschen Wortkombinationen in monolingualen deutschen Wörterbüchern und einem deutsch-ungarischen Wörterbuch. Anhand eines Vergleichs der beiden Quellen wird deutlich, dass die Sekundärquellen ohne die Corpusanalyse nicht ausreichen, da sie ungebräuchliches Sprachmaterial enthalten, während frequente Kombinationen fehlen (Hollós 2005: 27). Für das Ungarische stehen morphosyntaktisch analysierte Corpora von geeigneter Größe und Corpusrecherchetools, die eine statistische Kotextanalyse integrieren, nicht zur Verfügung (Hollós 2005: 66-68).
Darüber, ob eine Wortverbindung als Kollokation einzustufen ist, entscheidet der Grad der Kombinierbarkeit des zur Basis gehörenden Kotextpartners (Hollós 2005: 46). Ist dieser stark begrenzt kombinierbar, ist er ein Kollokator und damit die Wortkombination eine

50 http://www.ids-mannheim.de/kt/projekte/cosmas_I/

Kollokation. Mit dieser Vorgehensweise bestimmt Hollós die *intralingualen* Kollokationen. Als Grenze der Anzahl der Basen bei einer Standardkotextanalyse mit hoher Zuverlässigkeit gibt Hollós ca. 15 an, um das untersuchte Wort als Kollokator und damit die Kombination als Kollokation anzuerkennen. Der bei der COSMAS-I Corpusanalyse verstellbare Parameter "Zuverlässigkeit" (gering/mittel/hoch) bezieht sich auf den Status der Kombination als Kollokation und wird anhand der Gamma-Werte ermittelt. Liegt eine Vielzahl von Basen mit einem ähnlichen Gamma-Wert zu einem Kollokator vor, sollte bei der Aufnahme einer Kombination in das Wörterbuch "... der entscheidende Faktor die Wichtigkeit der Wortverbindung für Deutschlerner sein" (Hollós 2005: 50).
Die statistisch signifikanten deutschen Wortkombinationen werden unabhängig davon, ob sie sich durch die Größe des Kollokationsbereichs des Kollokators als freie Kombination oder Kollokation erweisen, wortwörtlich ins Ungarische übersetzt. Im Falle einer Nicht-Übereinstimmung der wortwörtlichen Übersetzung mit der ungarischen Norm muss eine Alternativlösung, d.h. ein Äquivalent, gefunden und festgelegt werden. Dies geschieht mit Hilfe mehrerer (deutsch-)ungarischer Wörterbücher.
Eine *interlinguale* Kollokation liegt vor, wenn die wortwörtliche Übersetzung in Ziel- und Ausgangssprache nicht übereinstimmt. Dabei kann eine interlinguale Kollokation in der Zielsprache durch eine nicht wörtlich äquivalente Übersetzung einer (intralingualen) Kollokation oder einer freien Wortkombination der Ausgangssprache entstehen (Hollós 2005: 72). Da die Bestimmung einer Kollokation von der Größe des Kollokationsbereichs des Kollokators abhängt, und die Bestimmung des Kollokationsstatus im Ungarischen anhand der verfügbaren Corpora aufgrund der Größe und der fehlenden Corpusrecherchetools nicht möglich ist, wird ausschließlich die Beschaffenheit der Wortkombination innerhalb der Ausgangssprache berücksichtigt. Die interlingualen Kollokationen sind im Gegensatz zu anderen Wortkombinationen "kontrastiv idiomatisch".
Das Kollokationskonzept der intra- und interlingualen Kollokationen äußert sich in der Mikrostruktur der vorgeschlagenen Wörterbuchartikel. Das Symbol '☝' der Schnellzugriffsstruktur verweist auf interlinguale Kollokationen. Der Pfeil '🡅' beim Kollokator verweist auf diesen als Lemma. Lemmatisiert werden nur die begrenzt kombinierbaren Kollokatoren, wodurch der Benutzer beim Erscheinen des Pfeils auf eine intralinguale Kollokation in der Ausgangssprache schließen kann und bei Bedarf weitere Kotextpartner des Kollokators unter dessen Artikel ermitteln kann. Das Ergebnis sind Wörterbuchartikel, in denen die häufigsten syntagmatischen Kombinationspartner strukturiert nach ihrer Kategorie und ihrem intra- und interlingualen Kollokationsstatus erscheinen.

Bei Butina-Koller (2004: 27-28) übernimmt die Wörterbuchfunktion hinsichtlich der Aufnahme von Kollokationen in ein zweisprachiges Wörterbuch eine tragende Rolle. Neben die allgemeingültigen Auswahlkriterien der Frequenz und der "Typikalität"[51] tritt bei einem zweisprachigen Wörterbuch der kontrastive Aspekt. Ein rezeptionsorientiertes Wörterbuch kann viele transparente Kollokationen aus seinem Lemmabestand ausklammern, da sich der Benutzer bei der Auswahl passender muttersprachlicher Übersetzungsäquivalente durch sein Sprachgefühl leiten lassen kann. Zusätzlich muss das passive Wörterbuch auch selten gebrauchte Kollokationen aufnehmen, wenn sie kontrastiv nicht transparent sind, denn für die Rezeption ist eine möglichst große Abdeckung der Sprache entscheidend. Stilistisch

51 Die "Typikalität" bezieht sich auf mehrere Aspekte im Gebrauch von Wortschatzeinheiten: stilistische Charakteristika, ggf. diatopische und diastratische Varietäten usw. (Butina-Koller 2005: 27)

bzw. konnotativ gefärbte und fachsprachliche Wortschatzeinheiten sollten hier zu finden sein. In einem produktionsorientierten Wörterbuch sind die transparenten Kollokationen zu verzeichnen, da man sich bei der Übersetzung nicht auf sein Sprachgefühl verlassen kann. Dafür kann sich ein aktives Wörterbuch auf die Bearbeitung von häufigen[52] und typischen Kollokationen konzentrieren, denn es ist für "neutrale" Situationen konzipiert.
Die Unterscheidung der Wörterbücher in passive und aktive entspricht für viele Sprachpaare eher einem linguistischen Ideal als der lexikografischen Realität, so auch für das portugiesisch/deutsche. Die meisten Wörterbücher sind bidirektional und jede Sprachrichtung sowohl für Textproduktion als auch für Textrezeption konzipiert. Herbst und Klotz formulieren das sprachkontrastive Kriterium daher auch allgemeiner (2003: 138). Als Kollokationen einzutragen sind nur die Kombinationen von Wörtern, bei denen die Gefahr einer falschen Übertragung besteht, wie im Falle von *Fahrrad fahren* und *ride a bike*. Dieser Standpunkt birgt gewisse Risiken für den aktiven Gebrauch. Ist eine Kollokation nicht im Wörterbuch verzeichnet, kann sich der Benutzer nie sicher sein, ob sie tatsächlich fehlt, weil ein wörtliches Übersetzungsäquivalent vorliegt, oder ob die Kollokation selten gebraucht wird und mit ihrer nicht äquivalenten Übersetzung aus Platzgründen keinen Eintrag erhält.
Im Portugiesischen ist die Übersetzung von *braun* normalerweise *castanho*. *Braunes Haar* wird einfach übersetzt zu *cabelo castanho*. Die *braune Haut* hingegen ist *pele moreno*. Im Deutschen existiert nur ein Wort, in dem diese semantische Unterscheidung nicht explizit gemacht wird. *Pele moreno* ist im Vergleich zu *cabelo castanho* im *PONS Standardwörterbuch P-D/D-P* (2002) und im *Langenscheidts Taschenwörterbuch Portugiesisch* (2001) als Kollokation vorhanden (jeweils als Übersetzung des deutschen Kollokators). Im *Langenscheidt* findet man unter dem Eintrag *moreno* die Übersetzung 'dunkel(haarig,-häutig)'. Man spricht zwar von einem *moreno* als einem 'dunkelhaarigen Typ', niemals würde sich *moreno* aber als Adjektiv direkt auf die *Haare* beziehen (*moreno* verhält sich im Bezug auf die Haare wie *brünett* im Deutschen). Da man im Wörterbuch *cabelo castanho* nicht findet, könnte man leicht dazu tendieren das Adjektiv, das die *Haut* näher modifiziert, auch auf die *Haare* anzuwenden.
Neben den transparenten Kollokationen fehlen in den Wörterbüchern besonders häufig diejenigen Wortkombinationen, bei deren Übertragung in die Zielsprache ein Übersetzungsäquivalent existiert, das einem anderen kollokationalen Strukturtyp angehört, und in dem die betreffenden Wörter frei kombinierbar erscheinen. *Fazer inveja* (**Neid machen*) existiert im Deutschen nur als *neidisch machen*. *Neidisch machen* gehört zu einem neuen Strukturtyp, den Hausmann 2004 einführt (vgl. Kapitel 3.1.3), in dem die Basis das Adjektiv ist und das Kopulaverb der unvorhersehbare Kollokator. Lexikalisch bleibt die Kollokation transparent, da es sich beim Adjektiv um ein Derivat des Substantivs handelt. Mit 323 extrahierten Okkurrenzen im Corpus *Cetempúblico* ist *fazer inveja* sehr frequent. Aufgrund der unterschiedlichen Realisierung der kollokationalen Bedeutung in beiden Sprachen, müsste *fazer inveja* als einzige Übersetzungsmöglichkeit von *neidisch machen* auf jeden Fall im Wörterbuch stehen. Doch ist *fazer inveja* oder das deutsche Äquivalent in keinem der untersuchten Wörterbücher verzeichnet.
In den portugiesischen Wörterbüchern sind selten Kollokationen zu finden, die eine wörtlich äquivalente Übersetzung besitzen. Doch das frequente *ter medo* (*Angst haben*) ist in mehreren Wörterbüchern vertreten. Die Übersetzung von *Angst haben*, die im portugie-

52 Die häufigen Kollokationen werden im Internet mit der Suchmaschine *Yahoo!* ermittelt. Genauere Angaben zu dem Verfahren gibt Butina-Koller (2004: 178).

sisch-deutschen Vergleich so einfach ist, gestaltet sich beispielsweise in der englischen Übersetzung viel komplizierter. Auch in diesem Fall, in dem der Wörterbuchbenutzer andere Sprachpaare vor Augen hat, ist für ihn der Eintrag der Kollokation informativ. Wörtlich äquivalente Übersetzungen aufzunehmen ist also durchaus angebracht, denn zieht man in Betracht, dass der Wörterbuchnutzer ja nicht weiß, wie einfach er es hat, ist er nach der Konsultation schlauer. Unter dem Aspekt der Platzeinsparung wäre das Nicht-Vorkommen einer Kollokation im zweisprachigen Wörterbuch auch eine Art der Informa-tion im Sinne einer möglichen Minimallösung. Wenn diese Vorgehensweise konsequent durchgehalten wird, ist jede nicht verzeichnete Kollokation wortwörtlich zu übersetzen.

In elektronischen Wörterbüchern ist die Darstellung transparenter Kollokationen und ausführlicher Kollokationsinformationen kein Problem. Wie Herbst und Klotz (2003: 256) schreiben, fördert "die lexikografische Forschung ... , insbesondere seitdem der Einsatz von Korpora bei der lexikografischen Arbeit selbstverständlich geworden ist, eine zunehmende Fülle von Details in Bezug auf Valenzpatterns, Kollokationen, phraseologischen Einheiten etc. ans Licht". Zu bewältigen ist die Flut detaillierter lexikalischer Informationen nur noch in Druckwörterbüchern mit mehreren Bänden, was für viele Wörterbuchtypen impraktikabel ist, oder im elektronischen Medium.
Etliche der großen einsprachigen Wörterbücher sind mittlerweile auch auf CD-Rom erschienen (vgl. Herbst/Klotz 2003: 251-266). Mitunter wird dem Wörterbuch ein Corpus an die Seite gestellt, in dem die Benutzer über das im Wörterbuchartikel vorgegebene Material hinaus weitere Beispiele finden können. In den elektronischen Versionen der Wörterbücher ist die Suche über ein Eingabefeld möglich, mittels Hyperlinking kann man zu einem anderen Wörterbuchartikel springen und über eine Volltextsuche beispielsweise zwei eingetragene Kollokationen bei *deadline* um sechs weitere ergänzen, die sich in den Wörterbuchartikeln der Kollokate befinden.[53] Wünschenswert wäre durch die Steigerung der Informationsmenge eine Flexibilisierung des Wörterbuchs über Filter, so dass die angezeigte Information den spezifischen Nachschlagebedürfnissen entspricht. Es könnte rezeptive oder produktive Bedürfnisse bedienen, bestimmte Varietäten ein- oder ausschließen, oder spezifischere Informationen (zu syntaktischen Konstruktionen, Kollokationen, Synonymen und Antonymen, ...) auf Wunsch nur verborgen anzeigen. Das Wörterbuch ist in diesem Sinne zu verstehen als umfassende lexikalische Datenbank.
Einen detaillierten Überblick über die Merkmale einer lexikalischen Datenbank haben Heid/Freibott 1990 gegeben. Sie betonen die Notwendigkeit Kollokationen als Lemmata zu behandeln: "sie haben, wie "Einwortlexeme", eigene Übersetzungen und meist auch eigene Varietätenmarkierungen; Kollokationen brauchen auch nicht mit Kollokationen übersetzt zu werden. ... Es ist auch möglich, Kollokationen als nicht gebräuchlich bzw. von der Verwendungsnorm nicht akzeptiert zu bezeichnen: Übersetzer und technische Autoren sind an dieser Art "negativer Information" vor allem dann interessiert, wenn es sich um "falsche Freunde" handelt" (Heid/Freibott 1990: 251). Ein Zugriff erfolgt von beiden Elementen der Kollokation, da die Datenbank sowohl zur Hin- als auch zur Herübersetzung dienen soll, wie auch über den Wortlaut der Kollokation. Ein an dieses Konzept angelehntes über-setzungsorientiertes Datenbankmodell Deutsch-Spanisch beschreibt Caro Cedillo (2004) in der Dissertation über *Fachsprachliche Kollokationen*.

53 Herbst/Klotz (2003: 260) bringen dieses Beispiel für die Volltextsuche im einzigen von ihnen beschriebenen elektronisch vorliegenden zweisprachigen Wörterbuch, dem *Duden Oxford Großwörterbuch Englisch. Deutsch-Englisch / Englisch-Deutsch* (1999).

Neben die beschriebene Möglichkeit, elektronische Wörterbücher einem menschlichen Benutzer über eine Suchmaske zugänglich zu machen, tritt ihre Funktion als lexikalische Datenbasis verschiedener maschineller sprachverarbeitender Systeme. Im Gegensatz zum Menschen kann ein maschinelles System mit einfachen Kollokationslisten gar nichts anfangen. Bei der Maschinellen Übersetzung oder Sprachgenerierung werden auf jeden Fall präzise grammatische und semantische Informationen zu Kollokationen benötigt. Wie umfangreich eine systematische Einteilung nach syntaktischen und morphosyntaktischen Kriterien wird, zeigen Braasch und Olsen in dem Artikel "Formalised Representation of Collocations in a Danish Computational Lexicon" (2002).
Das polyfunktionale maschinenlesbare dänische Lexikon soll mit detaillierten lexikalischen Informationen als Datenbasis für verschieden NLP-Aufgaben dienen. Verzeichnet werden die Substantiv-Verb Kollokationen (des Typs V+N) beim Verb. Begründet wird dieses Vorgehen durch die sequentielle Analyse eines Satzes von links nach rechts, und die Möglichkeit, potentielle Kollokationen beim Lexikonnachschlag des Verbs zu signalisieren. Durch den Fokus des Lexikons auf morphologische und syntaktische Merkmale, ist die Identifizierung der semantisch dominanten Basis im Erkennungsprozess keine triviale Aufgabe. Die Subtypen der Kollokation werden durch sie geprägt, da der Beitrag der Valenzstruktur des Substantivs die syntaktischen Möglichkeiten der Kollokation bestimmt. Anhand der abgedruckten Beispiele wird deutlich, dass eine ausführliche Darstellung der Kontextmöglichkeiten für jedes Valenzmuster der Kollokation (Passivierbarkeit, Numerus, Determination und Adjektiveinfügung) mit den Anforderungen an Print-Medien nicht mehr zu vereinen ist.
Der Ansatz von Braasch/Olsen erleichtert die Erkennung von Kollokationen beim Parsen. Bei der Sprachgenerierung geht die systematische Wahl eines der Kollokationspartner vom Substantiv aus - ein geeignetes Mittel zur Beschreibung der syntagmatischen semantischen Beziehungen zwischen zwei Wörtern stellen hier die lexikalischen Funktionen von Mel'čuk dar. Polguère (2000) beschreibt die formale Datenbasis *DiCo*, die als Grundlage von NLP-Systemen und von Print-Wörterbüchern konzipiert ist: "Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French". Die deskriptiven Prinzipien der *Explanatory Combinatorial Lexicology* von Mel'čuk (vgl. Kapitel 2.4.1) bilden die Grundlage für die Kodierung. Die lexikalische Datenbasis *DiCo* ist primär kombinatorisch konzipiert, sie soll automatisch in die Lexika anderer NLP-Systeme zu übersetzen sein und stellt eine vereinfachte und formalisierte Variante der ursprüng-lichen *Meaning-Text Theory* von Mel'čuk dar.
Die Verben werden als Werte der lexikalischen Funktionen beim Substantiv geordnet. Der Artikelgebrauch und Subkategorisierungseigenschaften werden bei den einzelnen Kollokationen verzeichnet. Die lexikalischen Funktionen werden durch "popularisierte" Ausdrücke wiedergegeben, das "Meta-Französisch" ist auch für den Sprachenlerner leicht zu verstehen. Hintergrund für die Umschreibung der lexikalischen Funktionen ist die Nutzung der Einträge für ein "general public dictionary". Das *Lexique actif du français (LAF)* soll eine Lücke schließen zwischen "theoretischer" und "kommerzieller" Lexikografie und wird direkt aus den Einträgen des *DiCo* kompiliert. Das Konzept der lexikalischen Funktionen, das in der maschinellen Sprachprozessierung und Computerlexikografie bereits Anwendung findet, kann als detaillierte Beschreibungsmethode für semantische Derivationen und Kollokationen auch in die allgemeine Lexikografie einfließen. Hier zwei Ausschnitte für

Substantiv-Verb Kollokationen, geordnet nach ihren syntagmatischen lexikalischen Funktionen, aus *LAF* und *DiCo*:

> **MEURTRE, nom, masc**
> ... **FAIRE UN M.** accomplir, commettre, perpétrer [ART ~]; tremper [dans ART ~] [*Il a refusé de tremper dans ce meurtre odieux*.] **CAUSER QUE X FASSE UN M.** pousser [N_x au ~] **RAISON D'UN M.** mobile [de ART ~] **S'OCCUPER D'UN M.** enquêter [sur ART ~]; élucider [ART ~], trouver l'auteur [de ART ~]; punir, châtier [ART ~]; venger [ART ~] **SERVICE DE POLICE QUE S'OCCUPE DES M.** brigade criminelle ...
>
> (*LAF* Polguère 2000: 523)

> MEURTRE
> ... /*Faire un M.*/
> {Oper1} accomplir, commettre, perpétrer [ART ~]; tremper [dans ART ~] ["Il a refusé de tremper dans ce meurtre odieux".]
> /*Causer que X fasse un M.*/
> {CausOper1} pousser [N=X au ~]
> /*Raison d'un M.*/
> {S1CausOper1} mobile [de ART ~]
> /*S'occuper d'un M.*/
> {Real-I} enquêter [sur ART ~]
> {Real-II} élucider [ART ~], trouver l'auteur [de ART ~]
> {Real-III} punir, châtier [ART ~]; venger [ART ~]
> /*Service de police que s'occupe des M.*/
> {S1Real-I/II} brigade criminelle ...
>
> (*DiCo* Polguère 2000: 521)

Die Beispielsätze stammen aus Corpora. Auf die Idiome, die das Substantiv enthalten, die aber lexikalische Einheiten mit einem eigenen Artikel bilden, wird mit Pointer verwiesen. Die Darstellung von Kollokationen in Wörterbüchern, die über das Internet zu konsultieren sind, wird ausführlich erst in Kapitel 6.1.2 diskutiert.

3.2.4. Aufnahmekriterien und Usus in Kollokationswörterbüchern

Das Platzproblem der Print-Medien betrifft nicht nur die Darstellung der Kollokationsinformation, oder im Falle von zweisprachigen Wörterüchern das häufige Ausschließen der transparenten Kollokationen, sondern primär die Frage welche Wortkombinationen als Kollokationen gelten, und damit die Abgrenzung von den freien Syntagmen. Verschiedene Autoren kritisieren die dichotome Unterteilung lexikalischer Wortkombinationen in Kollokationen und Ko-Kreationen oder in freie und kodierte Kombinatorik wie sie Hausmann durchführt. Zum einen besteht zwischen Kollokation und Ko-Kreation ein Gradient, zum anderen können Kollokationen in unterschiedliche Maße etabliert sein. Klotz (2000: 88-91) demonstriert dies mit dem englischen Verb *catch*. *Catch a fish* und *catch a ball* sind Ko-Kreationen, ihre Kombinierbarkeit ergibt sich aufgrund semantischer Mindestbedingungen. Als klare Kollokationen gelten *catch attention* und *catch fire*, denn die jeweilige Bedeutung des Verbs ist von der Basis abhängig. Unklar ist die Einordnung der Kombinationen *catch a train/bus/boat/...* und *catch a cold/AIDS/pneumonia/...* . Einerseits repräsentieren sie jeweils semantisch wohl definierte Kollokationsbereiche, was sie als Ko-Kreationen auszeichnet. Andererseits trifft die von Hausmann postulierte wenig begrenzte Kombinierbarkeit für Ko-Kreationen auf sie kaum zu. Klotz sieht in diesem Beispiel keinen Einzelfall. Er rät dazu die Unterscheidung von Kollokation und Ko-Kreation aufzugeben und mit Cowie (1978) alle diese Kombinationen als Kollokationen zu betrachten, die auf einer

Skala zwischen quasi völliger Offenheit (*open collocation*) und äußerster Restriktion auf nur einen möglichen Kollokationspartner (*restricted collocation*) angesiedelt sind.
Innerhalb eines Kollokationsbereichs kann der Kollokator in unterschiedlichem Maße etabliert sein. *Commit suicide* ist sehr viel frequenter als *commit fraud*, obwohl beide Nomina gleich häufig im Corpus vorkommen und die gleichen semantischen Mindestbedingungen [+action][+legally or morally wrong] erfüllen. *Catch the bus* und *take the bus* sind ungefähr gleich oft vertreten, während *catch the subway* im Vergleich zu *take the subway* viel seltener erscheint. Die höhere Frequenz im Corpus ergibt sich allein aufgrund der größeren Etabliertheit der Kollokation, ohne dass hier semantische Faktoren geltend gemacht werden können. Mit Cowie (1978) kann man unterscheiden zwischen etablierten Kollokationen und potentiellen Kollokationen, die eine Verträglichkeit auf der semantischen Ebene auszeichnet, aber keine signifikante Kookkurrenz.
Klotz kommt zu einem Konzept, bei dem sich die Signifikanz einer Kollokation einerseits aus dem Grad ihrer Restriktivität und andererseits aus dem Grad ihrer Etabliertheit ergibt (Klotz 2000: 99). Klotz sieht diese beiden Aspekte in Abhängigkeit voneinander: ist eine Kollokation restriktiv (d.h. mit kleinem Kollokationsbereich und daher spezifischer Bedeutung), so wird sie auch einen hohen Grad an Etabliertheit (Corpushäufigkeit) zeigen. Damit bliebe als einziges relevantes Maß, neben der syntaktischen Abhängigkeit, die Frequenz einer Kombination.
Hausmann wehrt sich gegen die Frequenz als Kriterium und erinnert an die Disponibilität des Wortschatzes: "Viele Kollokationen sind nicht frequent, aber dennoch verfügbar" (Hausmann 1985: 124). Dass mitunter auch niedrigfrequente Verbindungen unter besonderen semantischen Bedingungen zu den Kollokationen zählen, wenn sie ein idiomatisches Element enthalten, sich sehr restriktiv verhalten oder sprachkontrastiv relevant sind, und zusätzlich in eine Wörterbuch aufgenommen werden können, spricht nicht gegen das Frequenzkriterium als gruppierendes Maß. Die Auswertung der Kookkurrenzdaten in Kapitel 6 zeigt, dass niedrigfrequente Wortkombinationen nur selten für einen Wörterbucheintrag relevant sind. In speziellen Wörterbüchern, die bestimmte Sprachausschnitte genauer thematisieren, können natürlich auch diese Kombinationen interessieren.
Hausmann untermalt seine Kritik am Frequenzkriterium auch anhand einer Untersuchung zu den verbalen Kookkurrenzen von *Angst* und bemängelt, dass unter den Wortkombinationen, die nur einmal im Corpus vorkommen, viele banale Verbindungen zu finden sind.[54] Dass die höher frequenten Kookkurrenzen (*Angst haben, bekommen, machen, einjagen, einflößen, empfinden, ...*) tatsächlich alle Kollokationen bilden, wird stillschweigend hingenommen. So ist seine Kritik eher eine Affirmation des Frequenzkriteriums, denn bei einem einmaligen Vorkommen im Corpus, geht man nicht von einer usuellen Verbindung aus, von Etabliertheit wird hier nicht gesprochen.
Kathrin Steyer gibt als entscheidendes Kriterium für eine Kollokation die statistische Signifikanz an. Die Kollokation ist für sie kein phraseologisches Phänomen, sondern ein usuelles, da sie eine typische Verbindung repräsentiert (Steyer 2000: 110). Auch sie gibt zu bedenken, dass es neben den deutlichen Kollokationen wie *eingefleischter Junggeselle* und Ko-Kreationen wie *einen Baum anschauen* ein breites Feld an statistisch signifikanten Kollokationspartnern gibt, denen man nach Hausmanns Kriterien keinen kollokativen Status

54 Die Untersuchung zum Wortfeld 'Angst' wurde 1980 von Bergenholtz durchgeführt: *Das Wortfeld Angst. Eine lexikographische Untersuchung mit Vorschlägen für ein großes interdisziplinäres Wörterbuch der deutschen Sprache*. Stuttgart, Klett. Die Kritik Hausmanns findet man in Hausmann (1985): 124-126.

zusprechen würde und führt als Beispiele an: 'Baum + stehen, pflanzen, fällen, wachsen, klettern ..'. Dabei wird von Steyer übersehen, dass Hausmanns Phrasem-Begriff nicht auf bildliche Kollokatoren beschränkt bleibt: "Banale Verben können Kollokator sein: so sagt man im Französischen *einen Traum machen* (statt deutsch *haben*) ... " (Hausmann 2000: III). 'Baum + stehen' könnte also durchaus als Kollokation gelten, wäre sie sprachkontrastiv relevant.

Beim sprachkontrastiven Vorgehen wird die Bestimmung einer Kollokation immer nur sprachpaarabhängig vorgenommen. Vermutlich würden sich fast alle vermeintlich trivialen Kombinationen als unvorhersehbare Kollokationen entpuppen wenn man den sprachkontrastiven Vergleich nicht im bilingualen Bereich verwirklichte, sondern eine Vielzahl an Sprachen zur Komparation konsultieren würde. Dies betrifft zum einen die Kombinationen, die in ihrer grammatischen Ausprägung nur in einer Sprache als Kollokation gelten. Neben dem oben erwähnten Beispiel *fazer inveja* kann man auch *ter ciúmes* (*'Eifersucht haben') mit der Übersetzung 'eifersüchtig sein' anführen oder *descalçar* für *Schuhe ausziehen*. Außerdem können die Kombinationen, die sich in der einen Sprache entsprechend gewisser semantischer Mindestregeln bilden, zu Bezugspartnern von Kollokationen der anderen Sprache werden, sind dort in diesem Bereich die semantischen Kriterien der Kombinierbarkeit restriktiver anlegt, wie im Fall der "banalen" Kombination *braune Haut* für *pele moreno*. Natürlich kann man argumentieren, dass *pele moreno* wie im Deutschen *blondes Haar* mittels semantischer Selektion gebildet wird und insofern keinen idiosynkratischen Charakter einnimmt, doch werden die lexikalischen Solidaritäten auch von Hausmann ausdrücklich zu den Kollokationen gezählt (vgl. Abb. 5) und als Beispiel für Kollokationen angeführt (*schütteres Haar* 1985: 119).

In den früheren Artikeln von Hausmann wird der Charakter der Kollokation als "Halbfertigprodukt der Sprache im Sinne der Norm" betont (1985: 118). Kollokationen sind Kombinationen von "auffallender Üblichkeit", deren Affinität definiert wird als "Neigung zweier Wörter, kombiniert aufzutreten" (Hausmann 1984: 398-399). In den späteren Arbeiten Hausmanns tritt diese Auffassung von Kollokationen, die in ihrer Definition ein häufiges Auftreten in der Sprache impliziert, zunehmend in den Hintergrund zugunsten einer Situierung der Kollokationen innerhalb der kodierten Kombinatorik als phraseologische Kombinationen. Hier sind allein linguistische Kriterien ausschlaggebend für die Klassifi-kation.

Wie sieht das Verhältnis von Kollokationen zu Ko-Kreationen in der monolingualen lexikografischen Praxis der Kollokationswörterbücher aus? Das sprachkontrastive Kriterium ist hier nicht anzuwenden und damit geht auch die Forderung nach der intralingualen Arbitrarität verloren, die Hausmann an die Kollokationen stellt. Pöll nimmt in sein *Dicionário Contextual* die "geläufigsten Wortverbindungen" auf (2000: V). Im *BBI Combinatory Dictionary* wird zunächst ausführlich darauf hingewiesen, dass freie Kombinationen keine Aufnahme finden. Dazu zählen Kombinationen mit Verben, die in fast unbegrenzter Anzahl mit Substantiven kombinieren und die aufgrund der Bedeutung der beiden Kombinationspartner als vorhersehbar erscheinen: "*build bridges (houses, roads), cause damage (deafness, a death), cook meat (potatoes, vegetables), etc.*" (1986: XXV). Erstaunlicherweise sind dann doch fast alle diese Kombinationen unter den Einträgen der Substantive verzeichnet. Das *Oxford Collocations Dictionary* (2002: vii-viii) betrachtet Kollokationen zwischen den beiden Extremen der völlig freien Kombinationen (*see a man/car/book*) und der Idiome (*not see the wood for the trees*). Die gesamte intermediäre

Skala soll in die Einträge einbezogen werden: schwache Kollokationen (*see a film*), mittelstarke Kollokationen (*see a doctor*) und starke Kollokationen (*see danger/reason/the point*) sind vertreten.

Die monolinguale lexikografische Praxis ist demnach eher am Konzept von Grossmann/ Tutin, Cowie und Steyer orientiert, das die Etabliertheit der Kollokationen innerhalb einer großen Skala von Kombinationen als Kriterium zur Aufnahme in ein Wörterbuch nimmt. Damit wird der Tatsache Rechnung getragen, dass sich für den Wörterbuchbenutzer die Etabliertheit einer Kollokation durchaus als relevant erweist, es wird ja gerade auf typische Wortverbindungen verwiesen, und damit werden oft auch Verben mit einem sehr weiten Kollokationsradius als Kollokate der Substantive aufgeführt, wie *stay in* unter dem Nomen *hotel* und *put in* unter dem Eintrag *key* im *Oxford Collocations Dictionary* oder *cook potatoes* im *BBI Combinatory Dictionary* und *assar um frango* ('ein Huhn braten') im *Dicionário Contextual*.

Die Kollokationswörterbücher scheinen in der Praxis eine möglichst breite Abdeckung der Sprache anzustreben, sie verzeichnen die Kombinationen, die sich nach Corpusdaten als relevant erweisen. So ist die semantische Vereinbarkeit der Substantive der Gefühle im hier untersuchten Corpus *Cetempúblico* als Objekt von *ausdrücken* (*expressar/exprimir*) bei allen Kombinationen gegeben, *ausdrücken* kann man jegliche Art von Gefühlen. Geläufige Kombinationen im Deutschen sind beispielsweise *Bewunderung/Liebe ausdrücken*. Die portugiesischen Äquivalente *expressar/exprimir admiração* (20/17) und *expressar/exprimir amor* (6/10) sind auch im Corpus frequent. Weniger üblich, aber dennoch möglich, ist die Kombination *Hass ausdrücken*. *Hass* wird oft *empfunden, genährt, geschürt* aber in *Cetempúblico* ist *expressar/exprimir ódio* (0/3) nur selten zu finden (obwohl *ódio* häufiger ist als *admiração* oder *amor*). Im *Oxford Collocations Dictionary* sind entsprechend *express admiration* und *express love* verzeichnet, nicht aber *express hatred* (eine interlinguale Identität der Kollokationen wird in diesem Vergleich vorausgesetzt).

Die Frage, ob der Grad der Etabliertheit einer Wortkombination ausschlaggebend ist für die Aufnahme in ein Wörterbuch, hängt u.a. von der Zielgruppe des Nachschlagewerkes ab. Handelt es sich um allgemeine ein- oder zweisprachige Wörterbücher wird meist eine möglichst große Abdeckung der Sprache angestrebt, wodurch frequente Kombinationen sinnreich erscheinen. Die Behauptung, die Cowie (1978) formulierte und die von Klotz (2000) neu aufgegriffen wird, dass restriktive Kollokationen einen hohen Grad an Etabliertheit zeigen, konnte im Verlauf der Auswertung der Kookkurrenzdaten der Substantiv-Verb Kollokationen im Portugiesischen bestätigt werden. Ein Beispiel ist *acalentar esperança* ('Hoffnung hegen' / '*aufwärmen') mit 70 Okkurrenzen, das Verb *acalentar* tritt in dieser Lesart nur bei einem Gefühlssubstantiv auf. Dies bedeutet aber nicht, dass Kollokate mit einem weiteren Kollokationsbereich weniger frequent wären. *Alimentar* ('nähren') kommt mit etlichen weiteren Emotionsnomina vor und ist trotzdem als Kollokat von *esperança* 169 mal vertreten. Auffällig ist, dass *alimentar* in der übertragenen Bedeutung mit abstrakten Nomina im untersuchten Zeitungscorpus sehr viel häufiger vorkommt, als in den Wortkombinationen, in denen es seine primäre Bedeutung ('füttern', 'ernähren') beibehält.

Die Behauptung im *BBI* (1986: xxiv, vgl. Kapitel 3.1.2) hingegen, dass Kombinationen aus zwei Wörtern, die beide einen sehr großen Kollokationsradius haben, weniger frequent sind, gilt zwar für den Großteil der freien Kombinationen, doch sind auch hier bei einem Verb bestimmte Substantive üblicher und somit häufiger vertreten (vgl. oben). Kombinieren sehr

geläufige Verben mit ihren bevorzugten Substantiven, sind freie Kombinationen wie *Haus verkaufen* weit vorne in den Rankinglisten zu finden. Charakteristisch an der Kombination *Haus verkaufen* ist, dass man von der Verbsemantik her zwar nahezu alles verkaufen kann, zumindest in geschriebenen Corpora scheint aber nur der Verkauf bestimmter Dinge relevant. Zurückzuführen ist dies auf gesellschaftliche Normen, in denen der Verkauf eines Hauses mehr ins Gewicht fällt als der Verkauf eines Toasters, die Sprachrealität spiegelt immer auch gesellschaftliche Werte wider. Ob man freie Kombinationen in ein Wörterbuch aufnimmt als Kollokationen, ist eine Frage der zugrunde liegenden Kollokationstheorie.
In Übereinstimmung mit der lexikografischen Praxis in den Kollokationswörterbüchern werden die aufgeführten Kollokationen der Gefühlssubstantive in Kapitel 6 nicht unter dem Aspekt einer sprachkontrastiven Transparenz oder wegen ihres idiosynkratischen Charakters ausgewählt. Ausschlaggebend ist zunächst die Frequenz einer bestimmten Kookkurrenz und die Weiterverarbeitung der Daten mit statistischen Assoziationsmaßen. Das Vorkommen der einzelnen Substantive der Gefühle in Singular- oder Pluralform variiert in *Cetempúblico* zwischen dem einmaligen Vorkommen seltener Pluralformen und den hochfrequenten Formen mit über 20.000 Okkurrenzen. Aufgrund der stark variierenden Frequenz der einzelnen Substantive ist es nicht ganz einfach numerische Werte festzusetzen, nach denen die Kollokationskandidaten ausgewählt werden. Bei Substantiven, die nur ein paar hundert mal im Corpus erscheinen, sind auch schon wenige Kookkurrenzen mit demselben Verb aufschlussreich. Andererseits ist es nicht möglich, bei den sehr häufig vorkommenden Nomina alle Kombinationen aufzuführen, die über einer bestimmten Frequenz, einem Wert des t-score oder der Mutual Information liegen. In diesem Fall werden nur die Kollokationen genannt, die in der Rankingliste nach t-score Werten vorne liegen, hohe Mutual Information Werte haben oder unter semantischen Gesichtspunkten ins Auge fallen.
Nimmt man die Corpusanalyse und die dadurch ermittelten Frequenzen und Werte statistischer Assoziationsmaße von Kollokationen als einzige lexikografische Quelle, besteht immer die Gefahr, dass Kollokationen nicht als solche identifiziert werden, da sie das Frequenzkriterium nicht erfüllen, oder dass freie Kombinationen, die aufgrund ihrer Häufigkeit das Frequenzkriterium erfüllen, als Kollokationen klassifiziert werden. Ob eine freie Kombination wie *Haus verkaufen* auf Grund ihrer Häufigkeit als Beispiel einer usuellen, wenn auch vollkommen regelmäßigen Verbindung im Wörterbuch beim Nomen oder Verb erscheint, bestimmt der Lexikograf. Statistische Angaben können einen Lexikografen nur leiten und auf Besonderheiten hinweisen. Das definitive Kriterium für eine Aufnahme im Wörterbuch bleibt die menschliche Introspektion. Ob die Aufnahme der Kollokationen in die Wörterbücher von einem numerischen Wert geleitet wird, oder zu beurteilen ist hinsichtlich semantischer oder kontrastiver Bedingungen oder von den Bedürfnissen einer bestimmten Benutzergruppe abhängt, ist die Entscheidung der Lexikografen. Ein Mensch muss beurteilen unter welchem Aspekt die Kollokation in ein Wörterbuch aufzunehmen ist oder nicht.

3.3. Divergenzen bei der Übersetzung von Kollokationen

Bei der Übersetzung von Kollokationen traten bisher Unterschiede in verschiedenen linguistischen Bereichen zwischen den portugiesischen Kollokationen und ihren deutschen Entsprechungen auf. Am ausführlichsten wurde bisher die Wahl der Kollokate behandelt. Es besteht ein gewisses Kontinuum zwischen den beiden Polen der wörtlich äquivalenten Übersetzung der Kollokate und einer in ihrer Bedeutung stark abweichenden wörtlichen Übersetzung der Kollokate. Im Mittelfeld ist die Entscheidung über Äquivalenz bzw. Divergenz der Kollokate nicht immer eindeutig. Es handelt sich bei den nicht wörtlich äquivalenten Übersetzungen der Kollokate um kollokationale Divergenzen in Ausgangs- und Zielsprache:

Kollokate in ihrer wörtlichen Übersetzung äquivalent

ter esperança	- Hoffnung haben
expressar admiração	- Bewunderung ausdrücken
causar	- Kummer bereiten (verursachen)
alimentar esperança	- Hoffnung nähren (*ernähren/*füttern)
depositar esperança em	- Hoffnung setzen auf/in (*einzahlen/*hinterlegen)
apanhar susto	- Schreck bekommen (*ernten/*aufsammeln/*fassen)
incitar ódio	- Hass schüren (*anstacheln)
pregar susto	- Schreck einjagen (*einschlagen/*annageln)
acalentar esperança	- Hoffnung hegen (*aufwärmen)
roer-se de invejas	- vor Neid vergehen/platzen (*sich abnagen)

Kollokate in ihrer wörtlichen Übersetzung stark divergent

Abb. 8: Kollokationale Divergenzen

Mit der Übersetzung einer Kollokation können eine ganze Reihe weiterer Divergenzen zwischen zwei Sprachen auftreten, die im Bereich der Maschinellen Sprachverarbeitung klassifiziert und konkretisiert werden.[55]

Mitunter muss das Substantiv durch ein Adjektiv übersetzt werden, oder die Bedeutung des Substantivs wird in die Übersetzung der dichotomen Struktur mit nur einem Verb inkorporiert. Man spricht dann von kategorialen Divergenzen. Die Bedeutung des übersetzten Substantivs bleibt in den Derivaten transparent. Doch würden ein Vollverb und eine Adjektivkombination mit *sein* nicht mehr zu den Kollokationen zählen.

ter ciúmes	- eifersüchtig sein (*Eifersucht haben)
fazer inveja	- neidisch machen (*Neid machen)
ter admiração	- bewundern (*Bewunderung haben)

Syntaktische Divergenzen[56] können bei der Übersetzung der beiden Kollokationspartner innerhalb der Kollokation auftreten. Das 2. Argument des Verbs (hier die Basis), kann je nach der Rektion des Verbs als Akkusativ-, Dativ- oder Präpositionalobjekt realisiert werden. Im folgenden Beispiel wird ein Präpositionalobjekt im Portugiesischen gefordert,

55 Eine Klassifikation allgemeiner Übersetzungsprobleme bieten Dorr/Jordan/Benoit (1999: 4-12) im Bereich der Maschinellen Übersetzung und Heid (1997: 179-214) bei der Strukturierung von einsprachigen und kontrastiven elektronischen Wörterbüchern. Hier werden nur die Divergenzen besprochen, die sich bei der Übersetzung der untersuchten Substantiv-Verb Kollokationen als relevant erweisen.

56 Die Syntax des Portugiesischen wird in Kapitel 5.4.2 näher erläutert.

während das Verb in der Zielsprache ein Dativobjekt verlangt. Es handelt sich um interne syntaktische Divergenzen bei der Kollokationsübersetzung:

123532937: « Não confio nas declarações deste Governo ... intern: *em* PP -> Dat
(N) V *em* PP *de* PP extern: *de* PP -> Gen
Ich vertraue den Erklärungen dieser Regierungen nicht ...
N V Dat Gen

Kollokationen kann man nicht in allen Fällen als Nominalprädikate beschreiben, bei denen das Nomen als Prädikatskern selbst prädikative Funktion übernimmt. Wie Detges (1994) bemerkt (vgl. Kapitel 3.1.3) sind die Nomina der FVG keine Aktanten der Funktionsverben. Um eine Einheitlichkeit in der Nomenklatur zu wahren, und da die Frage der Abgrenzung von Funktionsverben gegenüber anderen verbalen Kollokaten nicht immer trivial ist, wird hier auch den Nomina der Funktionsverbgefügen Aktantenstatus zugesprochen. Häufig hängen weitere (externe) syntaktischen Divergenzen im Satz mit den Rektionseigenschaften des nominalen Kollokationspartners zusammen. Oben im Beispiel wird die syntaktische Realisierung des 3. Aktanten im Satz vom Substantiv regiert.
Kategoriale Divergenzen bei der Übersetzung der Substantiv-Verb Kollokationen führen immer zu einer syntaktischen Neustrukturierung des Satzes. Mit dem Wechsel vom Substantiv zum Adjektiv oder der Inkorporation des Substantivs in die Verbbedeutung ist das 2. Argument des portugiesischen Verbs als Objekt nicht mehr verfügbar. Im folgenden Fall ist die Realisierung des Adjektivs in einem Valenzmuster von *machen* angelegt:[57]

22080040: ... a celebridade das penitenciárias faz <inveja> ao próprio inferno . Akk-> Adj
N V **Akk** *a* Dat Dat -> Akk
... die Berüchtigtheit der Haftanstalten macht selbst die Hölle neidisch.
N V Akk **Adj**

Im nächsten Beispiel hängt das Präpositionalobjekt im Portugiesischen von der Valenz des Substantivs *admiração* ab. Im Deutschen wird das Gefühlssubstantiv in die Verbbedeutung integriert und das Objekt der Bewunderung vom verbalen Valenzrahmen subsumiert:

97563588: Dostom tem muita <admiração> por Massud . Akk -> Ø (V)
N V **Akk** *por* PP *por* PP -> Akk
Dostom bewundert Massud sehr.
N **V** Akk
≈ * Dostom hat große Bewunderung für Massud.
N V **Akk** *für* PP

Der Wechsel eines der beteiligten Satzglieder in den Valenzrahmen des anderen Wortes der Kollokation kann auch unabhängig von kategorialen Divergenzen erfolgen. Im folgenden Beispiel ist der fakultative 3. Aktant *aos outros árabes* im Portugiesischen vom Verb abhängig. Im Deutschen wechselt das Dativobjekt in der ersten Übersetzung in den Valenzrahmen des Substantivs und in den Genitiv. Wird der *Neid* im Deutschen mit Artikel gebraucht, ist die Realisierung des Aktanten des Substantivs auch nicht mehr fakultativ, sondern obligatorisch, sofern der *Neid* im Kontext noch nicht eingeführt ist. Zusammen hängt dies mit dem präsuppositionalen Gehalt des Artikels, der die Spezifikation des *Neids* verlangt. Steht *Neid* ohne Artikel, ist der Neider Präpositionalobjekt. In den portugiesischen Corpora kommt zwischen *causar* und *inveja* kein bestimmter Artikel vor:

57 Zur Valenz deutscher Verben, Adjektive und Substantive vgl. Helbig/Schenkel (1983) und Sommerfeldt/ Schreiber (1983).

113867064: ... a vida extravagante causava <inveja> aos outros árabes .
N V **Akk** *a* Dat
... das extravagante Leben erregte den Neid der anderen Araber. Dat -> Gen
N V *det* **Akk** Gen
... das extravagante Leben erregte Neid bei den anderen Araber. Dat -> *bei* PP
N V **Akk** *bei* PP

Syntaktische Differenzen bei der Realisierung einer Kollokation in der Ausgangssprache können zu Bedeutungsunterschieden führen, die sich in der Zielsprache lexikalisch niederschlagen. Im monolingualen Bereich liegt mit Klotz (2000) eine corpusbasierte Untersuchung vor über die Beziehung der Verbbedeutung zu den vom Verb eröffneten Valenzpattern und deren konkrete lexikalische Realisierung. In den folgenden Beispielen haben die unterschiedlichen Bedeutungen eines polysemen Verbs unterschiedliche Valenzrahmen, die auf den ersten Blick sehr ähnlich erscheinen. Bei genauerem Hinsehen fällt auch die abweichende Belegung der Valenzstellen durch die Verbargumente auf. Die abweichenden Bedeutungen eines polysemen Verbs determinieren bestimmte Subkategorisierungsrahmen, die hier zur Disambiguierung der Verbbedeutung verwendet werden können, ohne die lexikalische Füllung oder deren semantische Eigenschaften zu kennen. Die Realisierung eines bestimmten Subkategorisierungsrahmens kann als Indikator für die richtige Bedeutung des Verbs in der Ausgangssprache dienen, und daher auch über die Wahl des geeigneten Kollokats in der Zielsprache bestimmen. Das Vorkommen des Kollokats in einer spezifischen syntaktische Strukturen in der Ausgangssprache determiniert kollokationale Divergenzen in der Zielsprache:

erfüllen -> encher

Die Möglichkeit erfüllte die internationale Gemeinschaft mit Hoffnung
N V Akk *mit* **PP**
43213632: A ocasião encheu de <esperança> a comunidade internacional .
N V *de* **PP** Akk

erfüllen -> cumprir

... sie hatten es nicht geschafft, die großen Hoffnungen (Erwartungen) zu erfüllen, die ...
N **Akk** V
51073382: ... , não lograram cumprir as elevadas <esperanças> que ...
(N) V **Akk**

Unterschiedliche Subkategorisierungsrahmen können aber ebenso die Auswahl des Übersetzungsäquivalents eines polysemen Substantivs motivieren:

ter uma inclinação de + Nomen -> 'Neigung' im Sinne von 'Gefälle'

92280968: A pista terá uma <inclinação> de 30 graus , ...
Die Rennbahn wird eine Neigung von 30 Grad haben , ...

ter uma inclinação por|para + Nomen -> 'Neigung' im Sinne von 'Zuneigung', 'Hang zu', 'Talent'

19764720: « Eles tinham uma <inclinação> para o amor ordinário » .
"Sie hatten einen Hang zur gewöhnlichen Liebe".

46568203: ... , o novo presidente tem uma grande <inclinação> por Portugal e ...
.. . der neue Präsident hegt/empfindet/hat eine große Zuneigung zu Portugal und ...

Wird *inclinação* mit 'Zuneigung' übersetzt ist das passende Kollokat von *ter* im Deutschen 'hegen' 'empfinden' oder 'haben', während im Zusammenhang mit 'Neigung' das portugiesische Kollokat nur mit 'haben' übersetzt werden kann .

Thematische Divergenzen entstehen durch die unterschiedliche Abbildung der Verbargumente auf die subkategorisierten Komplemente in den beiden Sprachen. Die Vertauschung der Subjekt- mit der Objektposition bei zweistelligen Verben und die Realisierung des thematischen Arguments als Akkusativobjekt in der einen Sprache und als Präpositionalobjekt in der anderen Sprache sind Beispiele für diese Art der Divergenz:

63871899: -- Eu acho que ninguém gosta da <pena> de morte ... N -> Dat
N V *de* **PP** + PP -> N

-- Ich denke, dass niemandem die Todesstrafe gefällt ....
Dat **N** V

≈ -- Ich denke, dass niemand die Todesstrafe mag ... (syntaktische Divergenz) PP -> Akk
N **Akk** V

50580067: - « Peço ao Meritíssimo <pena> suspensa , ... Akk -> PP
(N) V Dat **Akk** + Dat -> Akk

- " Ich bitte Hochwürden um Bewährungsstrafe, ...
N V Akk *um* **PP**

3.4. Äquivalenzbeziehungen zwischen deutschen und portugiesischen Substantiven

Die Divergenzen bei der Übersetzung der Kollokationen in Ausgangs- und Zielsprache betreffen die Übersetzung des Verbs, die syntaktische Realisierung weiterer Satzteile sowie die mögliche Übersetzung des Substantivs mit einem verbalen oder adjektivischen Derivat. Die folgenden Überlegungen beziehen sich auf mögliche Ambiguitäten, die sich bei der Übersetzung der Basis ergeben. Wenn man vom Konzept Hausmanns ausgeht, würde man beim Nachschlagen einer Kollokation im Wörterbuch das Substantiv wählen. Hat man dieses parat kann man in ausführlichen monolingualen (Kollokations)wörterbüchern unter dem Eintrag des Nomens nach Angaben zu den Kollokaten suchen. Die richtige Wahl des Kollokats erfolgt über semantische Beziehungen zu verwandten Kollokaten oder die Beispielsätze. Auch bei Mel'čuk (vgl. Kapitel 2.4) sind die syntagmatischen lexikalischen Beziehungen beim Substantiv verzeichnet. Mel'čuk gruppiert die Kollokate systematisch nach lexikalischen Funktionen. Die Wahl des Kollokats oder der mögliche kategoriale Wechsel der Basis in einer anderen Sprache sind abhängig vom spezifischen Nomen und der Funktion die das Verb in Bezug auf das Nomen übernimmt.
Da es im bilingualen Bereich noch keine Kollokationswörterbücher gibt, und in den kleinen zweisprachigen Wörterbüchern, wie sie fürs Portugiesische vorliegen, nur wenige Kollokationen verzeichnet sind, bräuchte man, um bei der Übertragung einer Kollokation in die Zielsprache ein monolinguales Kollokationswörterbuch konsultieren zu können, zuerst die Übersetzung der Basis. Die Übersetzung der Basis ist nicht durch den obligatorischen syntagmatischen Bezug auf ein weiteres Wort gekennzeichnet. Es handelt sich zunächst um die Übersetzung einfacher Lexeme. In den einsprachigen Wörterbüchern erfolgt die Bedeutungsangabe der Substantive über die semantischen Beziehungen des Substantivs zu anderen Substantiven: Synonyme, Antonyme, Hyperonyme oder Meronyme werden beispielsweise genannt. Bei Mel'čuk finden diese Beziehungen Ausdruck in den paradigma-

tischen lexikalischen Funktionen. Unter dem Gesichtspunkt der Bedeutungsdifferenzierung des Nomens sind auf der syntagmatischen Ebene gerade auch Kollokationen verzeichnet.
In den kleineren zweisprachigen Nachschlagewerken findet man häufig unter dem gesuchten Substantiv mögliche Übersetzungsäquivalente ohne genauere syntagmatische oder paradigmatische Angaben. Mit einer Vielzahl an Übersetzungen ist **Wut** ("raiva, fúria, furor; sanha; mania"*)* im *Langenscheidt* (2001) präsent. Die Polysemie des Substantivs zeigt sich in der Mikrostruktur der Wörterbucheinträge, die die Bedeutungsunterschiede zwischen den Übersetzungen strukturiert: "Das Semikolon trennt eine gegebene Bedeutung von einer neuen, verschiedenen: zwischen verwandten Begriffen steht ein Komma. Wesentliche Bedeutungsunterschiede bzw. verschiedene Wortarten werden durch Zahlen oder auch durch Buchstaben gekennzeichnet" (*Langenscheidt* 2001: 13). Bei der Aufzählung der Übersetzungsäquivalente für *Wut* kann sich der Benutzer aber nicht sicher sein, wie sich die "verwandten Begriffe" gegenseitig und von den "verschiedenen Bedeutungen" unterscheiden. Sind *raiva, fúria, furor* Synonyme, oder differieren sie hinsichtlich semantischer, stilistischer, register- oder kontextbezogener Merkmale? Nicht zu erklären ist auch die semantische Differenz zu den "verschiedenen Bedeutungen" *sanha* und *mania*.
Bei **Aufregung** werden die folgenden Übersetzungsmöglichkeiten präzisiert: "agitação, excitação, aflição; *freudige:* alvoroço" findet man im *Langenscheidt* (2001). In den Benutzerhinweisen steht, dass die Bedeutungsunterschiede gekennzeichnet sind durch: "a) kursiv erscheinende Synonyme in runden Klammern; b) durch portugiesische bzw. deutsche Ergänzungen oder Erklärungen in *Kursivschrift*". Diese bezeichnen vorangestellt bei Substantiven den Anwendungsbereich, bei Verben die Objekte. Nachgestellt nennen sie bei Substantiven sprachliche Kombinationsmöglichkeiten, bei Verben die Subjekte. Im *Pons* (2000) gibt es keine Benutzerhinweise, und der "Anwendungsbereich" *freudig* aus dem *Langenscheidt* erscheint hier mit einem anderen Nomen. Verwandte Übersetzungsäquivalente im *Langenscheidt* werden durch wesentliche Bedeutungsunterschiede im *Pons* markiert: "1. (*Verwirrung)* agitação, alvoroço 2. (*freudig*) excitação". Über ein deutsches Substantiv wird die Semantik der anderen Untereinheit des Wörterbuchartikels konkretisiert.
Die Einträge in den bilingualen Wörterbüchern spiegeln syntagmatische und paradigmatische Relationen wider, die sich gegenseitig bedingen. Mitunter trägt das Auftreten eines spezifischen Kollokats zur Bedeutungsdifferenzierung bei. Die Wahl des richtigen Übersetzungsäquivalents des Substantivs ist abhängig von einem Kollokat. Für ***furor*** findet man im *Langenscheidt* (2001): "a) Raserei; Wut; Tobsucht b) Begeisterung; ***fazer*** ~ P Furore machen". Auch die Unterteilung von ***pena*** im *Pons* (2000) wird mit Kollokationen belegt: "1. (DIR) Strafe; - **capital/de morte** Todesstrafe; ~**suspensa** Bewährungsstrafe; **cumprir uma** ~ eine Strafe verbüßen 2. (*pesar*) Leid, Kummer; (*piedade*) Mitleid; **eu tenho/sinto** ~ **dele** er tut mir Leid; **tenho muita** ~! es tut mir sehr Leid!; ... 3. (*de ave*) Feder". Für die Wörter innerhalb einer Untereinheit mit ähnlichen semantischen Merkmalen differenziert der Eintrag im *Pons* die Bedeutungen über die Angabe eines weiteren äquivalenten Substantivs der Ausgangssprache. Sie wird aber auch wie im *Langenscheidt* durch Kollokationen präzisiert: "... b) Kummer; Leid; Erbarmen; Qual; **meter~** traurig stimmen; Mitleid hervorrufen; **é uma** ~ es ist schade; **faz-me** ~, **tenho** ~ es tut mir Leid; **valer a** ~ sich lohnen, die Mühe wert sein; ...".
Hinsichtlich der Lemmatisierung besteht in monolingualen Wörterbüchern die Praxis zwei lexikalische Einheiten dann als zwei Lexeme zu betrachten und in unterschiedlichen

Einträgen zu behandeln, wenn sie über unterschiedliche etymologische Wurzeln verfügen. Dies ist der Fall bei *pena* deren dritten Bedeutung als '(Vogel-)(Schreib)Feder' im *Aurélio* (1986) separat in einem eigenen Eintrag erläutert wird. Dieses Verfahren findet in den bilingualen Wörterbüchern keine Anwendung, dort werden die Bedeutungen der homonymen Wortform innerhalb eines Eintrags auf der gleichen Ebene mit den Entsprechungen des Polysems verzeichnet.

Die unterschiedliche Granularität der lexikalischen Bedeutung eines Lexems in Ausgangs- und Zielsprache ist fast immer gegeben. Ist das Lexem der Ausgangssprache semantisch differenzierter als das der Zielsprache, kommt es zu einem Informationsverlust (*ser/estar* -> *sein* = zielsprachliches Hyperonym). Der Informationsverlust kann durch eine Paraphrase ausgeglichen werden. Stehen für ein Lexem mehrere Übersetzungsäquivalente zur Wahl, die sich durch semantische Merkmale unterscheiden, die in der Ausgangssprache keinen lexikalischen Niederschlag finden, wird ein Informationszuwachs erzwungen (*flor* -> *Blume/Blüte* = zielsprachliches Hyponym).[58] Die lexikalische Selektion wird in den Wörterbüchern über die Mikrostruktur, Synonyme und Kontextangaben erleichtert.

Die Polysemie der Lexeme und ihre im bilingualen Vergleich differente Verteilung führt zu komplexen Äquivalenzrelationen zwischen den Sprachen, bei denen ein Lexem gleichzeitig in Divergenz- und Konvergenzbeziehungen stehen kann. Abbildung 9 zeigt die Zuordnung der 40 untersuchten portugiesischen Gefühlssubstantive zu ihren deutschen Äquivalenten. In der linken Spalte befinden sich die von Mel'čuk und Wanner (1994) gewählten deutschen Substantive. Die Linien verdeutlichen die Übersetzungsmöglichkeit der deutschen und portugiesischen Substantive. Im Falle einer Rückübersetzung der portugiesischen Substantive wäre die Liste der 40 deutschen Substantive beträchtlich zu erweitern. Nur wenige deutsche Substantive sind im Wörterbuch mit genau einer portugiesischen Übersetzung vertreten. Dies betrifft beispielsweise *Hoffnung* (*esperança*), *Hass* (*ódio*) oder *Liebe* (*amor*). In der Rückübersetzung ist *esperança* (+ 'Erwartung') und *amor* (+ 'Liebling, Leidenschaft') aber polysem. Auch diese Übersetzungen bilden daher keine 1:1 Äquivalenz ab, sondern Teilbeziehungen. Die Übersetzung erfolgt nach den Einträgen im *PONS Standardwörterbuch* (2000) und im *Langenscheidts Taschenwörterbuch* (2001).

58 vgl. Heid/Freibott (1990: 247-248). Herbst/Klotz (2003: 117) bezeichnen die Äquivalenzbeziehungen als 'Divergenz' und 'Konvergenz'.

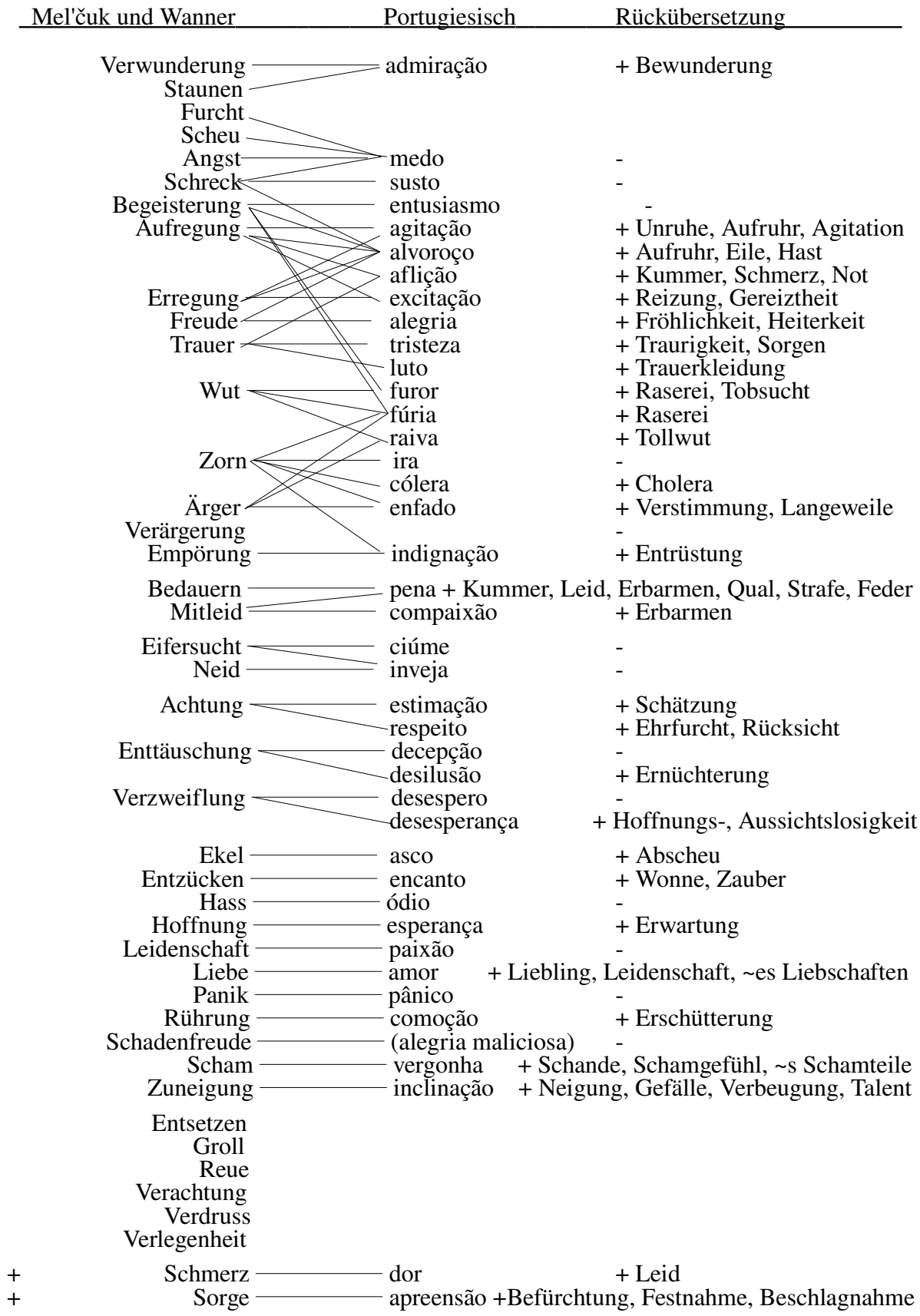


Abb. 9: Übersetzungsäquivalenz der 40 untersuchten Gefühlssubstantive im portugiesischen Wortfeld und bei Mel'čuk/Wanner (1994)

Die Teilnehmer am Wortfeld der untersuchten portugiesischen Gefühlssubstantive entsprechen nur in etwa den 40 deutschen Gefühlssubstantiven bei Mel'čuk und Wanner (1994). Einige der bei Mel'čuk und Wanner behandelten Substantive werden gar nicht berücksichtigt, weitere mit nur einer Übersetzung, obwohl mehrere Äquivalente in den Wörterbüchern gegeben sind, andere mit allen Übersetzungsmöglichkeiten. Die Auswahl der portugiesischen Substantive erfolgte nach verschiedenen Kriterien: zum einen sollen die geläufigen Gefühlssubstantive enthalten sein, daneben sollen aber auch repräsentative Beispiele (vermeintlicher) Synonyme und einige Polyseme vertreten sein, um Unterschiede und Ähnlichkeiten in deren Kookkurrenzverhalten zu demonstrieren (vgl. Kapitel 6.3). *Dor* und *apreensão* werden zusätzlich aufgenommen, da ihre spanischen Äquivalente in der automatischen semantischen Klassifikation von Wanner (2005) (vgl. Kapitel 2.4.2) vorkommen, und die portugiesischen Kookkurrenzdaten zum Vergleich interessant erscheinen.
Im Rahmen der vorliegenden Untersuchung erweist sich die Auswahl bestimmter Substantive mit mehreren Übersetzungsäquivalenten und das Fehlen anderer nicht als problematisch, da die Einträge der Gefühlssubstantive bei Mel'čuk und Wanner zwar als Vergleichsbasis für die Einträge der portugiesischen Substantive in Kapitel 6 dienen, eine umfassende Komparation der auf semantischen Kriterien beruhenden Klassifikation der Substantive und ihrer Kookkurrenzdaten mit dem Portugiesischen aber nicht angestrebt wird. In Kapitel 7.3 wird die von Mel'čuk und Wanner postulierte Vereinbarkeit von Verben und Substantiven mit bestimmten semantischen Merkmalen im Deutschen mit den portugiesischen Daten anhand einiger Beispiele verglichen. Eine Verifikation der bei Mel'čuk und Wanner durch menschliche Introspektion gewonnenen Kollokationsmöglichkeiten der Substantive mit den Verben anhand von großen Corpora ist nur in der gleichen Sprache durchführbar.

Die Abgrenzung der Bedeutungen polysemer Lexeme wird in den Wörterbüchern mit Hilfe stilistischer und registerbezogener Merkmale sowie über paradigmatische und syntagmatische Relationen zu anderen Lexemen strukturiert und erläutert. Im Rahmen dieser Arbeit interessiert besonders die Bedeutungsdifferenzierung durch spezifische Kollokate. Die Möglichkeiten der Auflösung der Ambiguitäten polysemer Wortformen lassen sich an den Kookkurrenzdaten messen. Unterscheiden sich die Übersetzungsäquivalente nur geringfügig in ihrer Bedeutung, werden sie ein ähnliches Kollokationsverhalten zeigen. Differieren sie in ihrer Semantik deutlich, sind die Kollokate mitunter komplementär verteilt. In der Maschinellen Sprachverarbeitung werden die Kookkurrenzdaten des polysemen Nomens zur automatischen Disambiguierung seiner Bedeutung verwandt. In den Einträgen der Gefühlssubstantive erweist es sich als praktikabel dann für ein Lexem verschiedene Subeinträge anzusetzen, wenn die Kollokationen für die verschiedenen Bedeutungen stark abweichen.
Die Wahl des äquivalenten Substantivs in der Fremdsprache wird in den Einträgen allein über syntagmatische Relationen vollzogen. Betrachtet man die aufgeführten Ambiguitäten, die bei der Übersetzung der Basis entstehen, ist die von Hausmann vertretene These, dass die Verwendung der Basis für die Sprachrezeption und Sprachproduktion immer banal und transparent ist, nicht mehr zu verstehen. Die Bedeutung der Basis bleibt in der Kollokation zwar (meist) transparent, nur um welche Bedeutung es sich bei polysemen oder homonymen Substantiven handelt, wird von sprachlichen Kontext und in vielen Fällen direkt vom Kollokat festgelegt. Das Substantiv determiniert funktional die Übersetzung der Kollokate, darüber hinaus trägt das Kollokat bei polysemen Substantiven genauso zu deren Bedeutungsdifferenzierung bei wie das Substantiv bei den verbalen polysemen Kollokaten.

4. Lusitanistik und Kollokationen

4.1. Forschungsüberblick

Die Behandlung von Kollokationen in der Lusitanistik beschränkt sich auf Untersuchungen weniger Linguisten, deren Ansätze und Ergebnisse im Folgenden kurz vorgestellt werden. Der Beitrag von Bernhard Pöll in Form eines Kontextwörterbuchs *(Dicionário Contextual Básico da Língua Portuguesa*, 2000) ist das einzige Kollokationswörterbuch für das Portugiesische (Aufbau und Struktur wurden schon in Kapitel 3.2.3 kurz verdeutlicht). In ihm werden Adjektive, Substantive und Verben als Kollokate der 1000 gebräuchlichsten Substantive des portugiesischen Grundwortschatzes verzeichnet. Dem lexikografischen Nachschlagewerk ging die Dissertation *Portugiesische Kollokationen im Wörterbuch: ein Beitrag zur Lexikographie und Metalexikographie* (1996) voraus. Hier wird vor allem der zugrunde liegende Kollokationsbegriff, der sich an Hausmann orientiert, die Corpuszusammensetzung und der Aufbau der Wörterbuchartikel näher erklärt.

Das Corpus auf dem das Kontextwörterbuch basiert, besteht zur Gänze aus Texten aus Portugal. Pöll sieht darin keine grundlegende Einschränkung für Benutzer, die der brasilianischen Norm folgen, denn abgesehen von oberflächlichen Unterschieden in der Orthographie weichen die beiden großen Varietäten des Portugiesischen in der Standardsprache nur unwesentlich voneinander ab (*Dicionário Contextual* 2000: VIII). Andere Autoren hingegen betonen, dass sich die Abweichungen der beiden Varietäten auch in der Lexik und Morphosyntax der Standardsprache niederschlagen.[59] Vereinfacht gesagt, handelt es sich um ein ähnliches Verhältnis wie zwischen dem amerikanischen und dem europäischen Englisch, je nach Standpunkt wird auf Gemeinsamkeiten oder Unterschiede hingewiesen (doch sind die Differenzen im Portugiesischen gravierender). In den hier untersuchten Wörterbüchern werden Wörter und Ausdrücke, die auf Brasilien beschränkt sind, durch (*bras*) markiert. Sind sie hingegen nur in Portugal bekannt, erfolgt im Falle vom *Pons Standardwörterbuch* (2002) und dem *Dicionário de Alemão-Português* (1989) keine Kennzeichnung, *Aurélio* (1986) führt sie unter *(lus)*, das *Langenscheidts Taschenwörterbuch Portugiesisch* (2001) unter *(port)*. Die Extraktion von Substantiv-Verb Kollokationen im Rahmen dieser Arbeit aus je einem Corpus der schriftlichen Standardsprache Portugals und Brasiliens zeigt, dass auch viele Kollokationen varietätenspezifisch sind (vgl. Kapitel 6.4).

In Kapitel 2 wurde die lexikalische Akquisition von Kollokationen aus Corpora dargestellt. Als weitere Quelle für ein Kollokationswörterbuch bieten sich ausführliche Definitionswörterbücher der betreffenden Sprache an, die dann entsprechend umzuarbeiten sind. Für gut beschriebene Sprachen wie das Französische führt dies allein zu quantitativ befriedigenden Ergebnissen (Hausmann 1989: 1012). Für das Portugiesische bietet sich diese Möglichkeit der Kollokationsakquisition nicht an, da Wörterbücher mit entsprechenden Informationen fehlen (Pöll 1996: 147). Dies bedeutet zum einen, dass es für das Portugiesische keine Wörterbücher wie das *Langenscheidts Großwörterbuch Deutsch als Fremdsprache* gibt (vgl. Kapitel 3.2.2), in denen die Kollokate explizit angegeben werden, und zum anderen, dass auch keine ausführlichen Definitionswörterbücher vorliegen, denn aus diesen geschieht die Akquisition aus allen Teilen der Mikrostruktur.

59 Einen Überblick auf Deutsch zu dieser Thematik gibt Noll (1999): *Das brasilianische Portugiesisch*. Portugiesisch wird außerdem in folgenden Ländern gesprochen: Mosambik, Angola, Kapverden, Guinea-Bissau, São Tomé und Príncipe. Hier ist mitunter eine kreolisierte Form des Portugiesischen anzutreffen, oder das Portugiesische hat den Status einer Minderheitensprache. Größere elektronisch verwertbare Corpora aus diesen Ländern sind nicht verfügbar.

Die portugiesische Wörterbuchlandschaft fällt im Vergleich zu vielen anderen europäischen Sprachen eher spärlich aus (Portugal ist mit ca. 10 Millionen Einwohnern auch ein kleines Land, während Brasilien mit ca. 200 Millionen Einwohnern immer noch zu den Schwellenländern zählt), und in den Wörterbüchern, die in Kapitel 6.2 näher untersucht werden, sind nur vereinzelte Informationen zu Kollokationen enthalten. Einen Überblick über die Lexikografie des Portugiesischen bietet Woll (1990) von den Anfängen bis heute, Schmidt-Radefeldt (2000) und Silva (1994) behandeln die zweisprachige Lexikografie Deutsch-Portugiesisch/Portugiesisch-Deutsch, und Oksefjell/Santos (1998) elektronische Versionen portugiesischer Wörterbücher. Die Auswahl der Wörterbücher, die dem Vergleich mit den corpusbasierten Kollokationsdaten dienen, wird in Kapitel 6 motiviert.

Die Darstellung von Kollokationen in deutsch-portugiesischen Wörterbüchern und in einsprachigen portugiesischen Verblexika wird in einigen Abschnitten des Artikels von Welker (2002) über die *Behandlung von Phraseologismen* kritisiert. Zur Phraseologie zählt Welker Idiome, Kollokationen und FVG. Situiert werden die Phraseme in den Wörterbuchartikeln der Verben, denn Ziel der Arbeit ist es, Musterartikel für ein eigenes deutsch-portugiesisches Verblexikon zu schreiben. In existierenden zweisprachigen Wörterbüchern fehlen Angaben zu eventuellen Restriktionen bzw. zur Festgeprägtheit der Phraseme, meistens fehlen auch Angaben zur "externen Valenz", der Einsatz von stilistischen Markierungen oder der Kennzeichnung *figuriert* ist inkonsistent, und Phraseme werden nur selten von den anderen Bedeutungen getrennt (Welker 2002: 408-409). Dieses Bild bestätigt sich in den beiden zweisprachigen Wörterbüchern, die in Kapitel 6 Verwendung finden. Auch die Repräsentation der Phraseologismen in den existierenden Verblexika des Portugie-sischen leidet unter den gleichen Einschränkungen.[60]

In den angebotenen Musterartikeln für das angestrebte Verblexikon erscheint die Systematisierung der Kollokate der Verben interessant. Zu jeder Verb-Lesart werden die erforderlichen oder möglichen Ergänzungen angegeben, sie erscheinen in drei unterschiedlichen Klammer-Typen: ( ) für Hyperonyme, [ ] für Beispiele für die kein Hyperonym zu finden ist, und > < für Ergänzungen, die nur durch ein Lexem oder wenige bestimmte Lexeme realisiert sind. Weniger überzeugt die Eintragung der Kollokate unter den normalen Lesarten der Verben. Bei Welker bleibt in Kollokationen die ursprüngliche Bedeutung der Komponenten immer erhalten (2002: 401), tatsächlich aber bestimmen häufig gerade die Substantive eine besondere Lesart der Verben. Die FVG stehen zusammen mit den (teil)idiomatischen Wendungen abgesondert von den normalen Lesarten der Verben in einer Rubrik, in der das Verb frei konjugierbar ist. In einer zweiten Rubrik werden die Verben mit Paradigmadefekten aufgeführt (*das schlägt dem Fass den Boden aus*), eine dritte umfasst verbale Idiome, die ohne den Kontext eines Substantivs funktionieren (*einen zwitschern*).

In den *Untersuchungen zur portugiesischen Phraseologie* von Christine Hundt werden FVG nur in den Bereich der Peripherie der Phraseologie eingeordnet, da sie nicht als idiomatisch im eigentlichen Sinne gelten (1994a: 46). Phraseologismen werden varietäten-spezifisch behandelt, und zahlreiche morphosyntaktische, lexikalische und semantische Varianten der Phraseolexeme Brasiliens und Mosambiks im Vergleich zum europäischen Portugiesisch zitiert. Der Terminus 'Kollokation' kommt in dieser Arbeit nicht vor, doch werden auch hier über die Untersuchung verteilt Kollokationen dargestellt, diejenigen mit einem

60 (Welker 2002a: 22, 24). In dem auf Deutsch erschienen Artikel Welker (2002) wird nur ein portugiesisches Verblexikon neben einem französischen und einem italienischen analysiert. Im Internet findet man eine leicht modifizierte Version des Artikel auf Portugiesisch, mit einem weiteren portugiesischen Verblexikon.

idiomatischen Element. Ist ein Verb im Phraseolexem enthalten, bildet es die Basiskomponente unter der die phraseologische Einheit erscheint (Hundt 1994a: 78).
In einer weiteren Arbeit[61] präzisiert Hundt ihre Einteilung anhand von Substantiv-Verb Konstruktionen, die sich durch ihre Reproduzierbarkeit und wiederholte Kookkurrenz auszeichnen, und kommt zu einer Dreiteilung dieser Verbindungen: 1. zur Substitution eines Vollverbs (*fazer uma afirmação = afirmar* - 'eine Bestätigung geben/*machen'), 2. zum Ausdruck von Aktionsart (*estar a disposição - zur Verfügung stehen/*sein* +durativo), 3. Einheiten mit einem figurativen Nominalelement und einem FV (*dar a dianteira a alg.* - 'jdm die Führung (über)geben'), oder einem Verb, das in der Kombination mit einem bestimmten Substantiv die Charakteristika eines FV einnimmt (*despertar a atenção* - 'Aufmerksamkeit (er)wecken'). Durch die Übersetzung wird deutlich, dass es sich in allen drei Fällen um Kollokationen handelt. Von Hundt wird nur die dritte Gruppe aufgrund ihrer partiellen Idiomatizität im Kernbereich der Phraseologie situiert (Hundt 2004b: 268-270).
Im Wörterbuch der *Idiomatik Deutsch-Portugiesisch* (2002) von Hans Schemann hingegen, sind etliche Informationen zu transparenten Kollokationen und FVG enthalten. Diese erscheinen jedoch nur sporadisch bei wenigen Lemmata (vgl. Kapitel 6.2). Vom gleichen Autor stammt eine umfangreiche Monografie zum Verb *dar* ('geben')[62], in der die Auffassung vertreten wird, dass "... Übertragung als Wesensmerkmal einer natürlichen Sprache Idiomatizität konstituiert" (Schemann 1981: 15). Das nicht-idiomatische *dar* gibt die Bedeutung "jem. etwas Konkret-Zählbares geben" wieder. In allen weiteren Kontexten ist die Bedeutung von *dar* idiomatisch. Aufgrund der Kontextpartner des Verbs in verschiedenen Bedeutungen wird die Übersetzung von *dar* ins Deutsche systematisiert. Die Gefühlssubstantive bilden eine eigene Gruppe, mit ihnen bedeutet *dar* "eine (den Körper duchziehende) seelische Empfindung hervorrufen" (Schemann 1981: 80). *Dar* wird innerhalb dieser Gruppe ins Deutsche nicht mit *geben* übersetzt, sondern mit *machen* (*Freude, Spaß, Sorge*(*n*)), *bringen* (*Ärger, Sorgen*), *erregen* (*Ekel*) oder *verursachen* (*Leid, Schmerz*). Abgesehen von unzureichenden Beschreibungskriterien des Merkmalbereichs mit [seelisch], [vital] lässt sich kein Archilexem oder Merkmal finden, mit dessen Hilfe im einzelnen entscheidbar wäre, ob *machen* oder *bringen* in Verbindung mir einem bestimmten Lexem der Norm entsprechen oder nicht (Schemann 1981: 84). Auch für die duale Unterscheidung [positiv] - [negativ] finden sich zahlreiche Gegenbeispiele.
Die häufige Erwähnung deutscher Lusitanisten im vorliegenden Abschnitt steht nicht nur in Zusammenhang mit der Verstehbarkeit der Literatur für den nicht-portugiesischsprachigen Leser, sondern leitet sich zum Großteil aus der Tatsache her, dass auf Portugiesisch zu diesen Themen keine linguistischen Studien existieren. Auch gezielte Untersuchungen zu Kollokationen im Portugiesischen beschränken sich auf die Arbeiten einiger Autoren, die alle aus Brasilien stammen. Die europäischen Lusitanisten beschäftigen sich bevorzugt mit einer Teilklasse der Kollokationen, den FVG, über die es auf Deutsch und Portugiesisch zahlreiche Arbeiten gibt.[63]

61 Hundt, Christine (1994b): "Construções de verbo + substantivo: estrutura, semântica e posição dentro da fraseologia".

62 Schemann, Hans (1981): *Das idiomatische Sprachzeichen. Untersuchung der Idiomatizitätsfaktoren anhand der Analyse portugiesischer Idioms und ihrer deutschen Entsprechungen.* Tübingen, Max Niemeyer.

63 In Form eines kurzen bibliographischen Verweises werden hier nur die ausführlicheren Artikel erwähnt: einen zweisprachigen Vergleich stellen Athayde (2001) (portugiesisch-deutsch) und Côco (2001) (portugiesisch-spanisch) an, Döll/Hundt (2002) grenzen FVG von anderen komplexen Verbalausdrücken ab, und Athayde (2002) beschreibt den Artikelgebrauch in FVG.

Döll/Hundt (2002) betrachten FVG nicht als integrierte Klasse im Paradigma der Kollokationen, sondern grenzen diese voneinander ab: "Hinsichtlich der Analyse ergibt sich bei Substantiv-Verb Kollokationen, daß der nominale Bestandteil Aktant zum Verb ist, die Nominalgruppe ist folglich im Unterschied zu FVG pronominalisierbar und erfragbar" (2002: 164). Bei Côco (2001) hingegen "sind FVG zweifelsohne der übergeordneten Bezeichnung 'Kollokationen' zuzurechnen" (2001: 39), mit einem Verweis erinnert sie an Burger, der die FVG die "am stärksten reguläre Untergruppe" (2003: 52) im Bereich der Kollokationen nennt. Da auch in den regulären Verbindungen nicht durchweg die gleichen Verben in Frage kommen, ist ein Phraseologisierungsaspekt vorhanden.

Louro (2001) behandelt Kollokationen in ihrer Dissertation *Enxergando as Colocações: Para ajudar a vencer o medo de um texto autêntico*[64] unter einem sprachdidaktischen Aspekt. Brasilianischen Englischlernern wird die Angst vor authentischen Texten genom-men, indem sie einige Zeit mit dem Kollokationskonzept vertraut gemacht werden. Dies befähigt die Schüler weiterhin, natürlich klingende Zusammenfassungen zu schreiben, weil sie Kollokationen aus den Ursprungstexten übernehmen. In einer Auswertung beobachtet Louro beachtliche Erfolge, was die Lese- und Schreibkompetenz ihrer Schüler betrifft, nachdem sie diese über didaktisches Material gezielt auf die Wichtigkeit von Wortkombina-tionen hingewiesen hat. Die Auswertung bezieht sich auf 'benennende Kollokationen' (*colocações denominadoras*) worunter Verbindungen verstanden werden, die die typische Funktion des Substantivs übernehmen. Die Realisierung des zweiten Lexems neben dem Substantiv ist im Englischen durch Adjektive (*high school*), Substantive (*crystal ball*), Verben [Partizip I] (*cleaning woman*), Verben [Partizip II] (*inverted commas*), Adverbien (*well-being*), Präpositionen (*by-product*), Numerale (*first aid*) oder Buchstaben (*T-shir*t) möglich. Dargestellt wird auch die genaue Kollokationsauffassung der Autorin, die sich an Hausmann orientiert, und die Einträge der Nominalkomposita in englischen und brasilianischen Wörterbüchern.

Ein didaktisches Anliegen hat auch Orenha (2004a), die in dem Artikel "Aplicações léxico-terminográficas da lingüística de corpus: relato da elaboração de um glossário bilíngüe de colocações na área de negócios" die Vorgehensweise bei ihrer Dissertation[65] beschreibt. Die Kompilierung eines bilingualen Kollokationsglossars aus dem Wirtschaftsbereich anhand vergleichbarer Corpora ist ihr Ziel. Integriert ist das Projekt in ein Seminar für Übersetzer und Dolmetscher, die gleichzeitig den richtigen Sprachgebrauch lernen und bei der Akquisition der Übersetzungsäquivalente helfen. Aus zwei Wirtschaftscorpora mit Zeitungs-texten extrahieren die Studenten mögliche Übersetzungen für vorgegebene Kollokationen in der jeweils anderen Sprache. Durch den Umgang mit verschiedenen Extraktionstools qualifizieren sie sich zusätzlich für ihr späteres Berufsbild.

Akquisition von Kollokationen und "semantische Prosodie" behandelt Sardinha (1999) in dem Artikel "Padrões lexicais e colocações do português". Er bemerkt, dass es vor seiner Arbeit keine corpusbasierten Studien im Portugiesischen zu Kollokationen gab. In der portugiesischen Sprachwissenschaft bezeichnet der Ausdruck *padrão* normalerweise die Standardsprache (der einzelnen portugiesischsprachigen Länder), Sardinha verwendet ihn ausschließlich für "linguistische Regelmäßigkeiten" (lexikalische, (morpho)syntaktische,

64 Die englische Übersetzung des Titels wird im Internet gegeben: *Learning collocations: to help read a text*, wo die Arbeit auch verfügbar ist.

65 Orenha, Adriane (2004b): *A compilação de um glossário bilíngüe de colocações, na área de negocios, baseado em corpus comparável*. São Paulo, Universidade de São Paulo, FFLCH, Dissertação de Mestrado. Die Dissertation ist im Internet oder in deutschen Bibliotheken nicht verfügbar.

semantische, stilistische), die in einem Corpus zu beobachten sind. Formal ist der *padrão* über das wiederholte Auftreten kookkurrenter Einheiten im Corpus zu identifizieren (Sardinha 1999: 3). Die Zuordnung einer Wortkombination zum lexikalischen Standard (*padrão lexical*) wird über die Signifikanz der Kookkurrenz mit dem t-Test und der Mutual Information geregelt. Ist der t-score größer als 2, und liegt die MI über 3, handelt es sich um Kollokationen, die zum *padrão* gehören, und nicht um bloße Kookkurrenzen. Die semantische Prosodie ist die Konnotation, die ein bestimmtes Lexem beinhaltet. So hat *cause* eine negative semantische Prosodie, mit *cause* werden unangenehme Wörter wie *problem(s), damge, death, disease* assoziert, während die Kookkurrenzen von *provide* (*assistance, care, jobs, opportunities*) ein semantisch neutrales bis positives Profil ergeben. Zu ähnlichen Ergebnissen kommt Sardinha im Portugiesischen. Die häufigsten Kookkurrenzen von *causar* sind die Wörter *problemas, danos, morte(s), prejuízos*, die Werte von t-Test und MI bestätigen, dass sie Kollokationen sind.

Von Sardinha stammt auch eine ausführliche Monographie zur Corpuslinguistik im Portugiesischen (2004) und ein Sammelband zur *Língua portuguesa no computador* (2005) ('Portugiesische Sprache im Computer').[66] Wie schon die Dissertation von Orenha, ist auch diese Literatur über deutsche Bibliotheken nicht erhältlich. Ebenso kann auf zwei Werke der brasilianischen Linguistin Stella Tagnin nur ohne deren Kenntnis verwiesen werden, das eine handelt von "idiomatischen und üblichen Ausdrücken" (1989), das andere ist ein Wörterbuch verbaler Kollokationen englisch-portugiesisch/portugiesisch-englisch (1999a).[67] In zwei Artikeln über "verbal collocations", die in europäischen Kongressakten enthalten sind, erläutert Tagnin (1999b, 2002) ihr Kollokationskonzept, die Struktur des bilingualen Wörterbuchs, und zeigt einige exemplarische Einträge. Die portugiesischen Kollokationen werden aus elektronisch verfügbaren Corpora transkribierter gesprochener Sprache und aus Zeitungstexten extrahiert, die beide nicht linguistisch annotiert sind. Die Bestimmung der Kollokationen in den Exzerptionsdateien, die sich je nach Wortform unterscheiden, geschieht manuell. Die englischen Kollokationen stammen aus dem COBUILD Projekt und dem Brown Corpus.

Als Kollokation qualifiziert sich bei Tagnin eine "recurrent, non-idiomatic, cohesive, arbitrary lexical combination, whose constituents are contextually restricted" (1999b: 400). Verbale Kollokationen bestehen aus einer Basis, normalerweise einem Nomen, das in der Objekt- oder Subjektposition steht, und einem Kollokat, dem Verb. Im Gegensatz zur Auffassung von Hausmann betrachtet Tagnin auch in den Verbindungen eines Adjektivs oder Adverbs mit einem Verb, das Verb immer als Kollokat. Dementsprechend erfolgt die Gliederung in den Wörterbüchern. Als Lemmata dienen die Basen, respektive Substantive, Adjektive und Adverbien. Die spezifischen Bedeutungen polysemer Wörter werden als einzelne Lemmata behandelt. Unter jedem Lemma erfolgt die Aufzählung der Kollokate in alphabetischer Reihenfolge, doch werden Synonyme zusammengefasst. Jeder Kollokation folgt ein Beispielsatz aus dem Corpus. Der fakultative oder obligatorische Gebrauch des bestimmten und/oder unbestimmten Artikels wird explizit verzeichnet. Für jede Kollokation wird die äquivalente Kollokation in der anderen Sprache gegeben, ebenfalls gefolgt von einem Beispielsatz. Sind mehrere Übersetzungsäquivalente in der Zielsprache möglich,

66 Sardinha, Tony Berber (2004): *Lingüística de Corpus*. São Paulo, Editora Manole.
Sardinha, Tony Berber (2005): *A Língua Portuguesa no Computador*. Campinas, Mercado de Letras.

67 Tagnin, Stella E. O. (1989): *Expressões Idiomáticas e Convencionais*. São Paulo, Ática.
Tagnin, Stella E. O. (1999a): *Convencionalidade e Produção de Texto: um Dicionário de Colocações Verbais Inglês/Português; Português/Inglês.* Tese de Livre-Docência, Universidade de São Paulo.

werden diese alle separat aufgeführt, sie können sich bezüglich des Nomens oder des Verbs unterscheiden (z.B. *ajuda, cortar a* und *auxílio*, *cortar o* für *aid, cut off).* Ist in der Zielsprache keine Kollokation als Übersetzungsäquivalent vorhanden, werden die Kombinationen der Zielsprache durch das Symbol ➔ eingeführt: *admittance, gain ~ to* ➔ *conseguir entrar* oder *aground, run* ➔ *encalhar* (2002: 739).
In den Einträgen von *account* wird deutlich, dass auch eine Lesart des englischen Nomens durch unterschiedliche Substantive im Portugiesischen wiedergegeben werden kann. Als Lesarten differenzierende Einheiten des im Englischen monosemen Lemmas dienen nicht nur die Kollokate, sondern ebenso typische Präpositionen:

account $_{n.}$ **[report]**
account, give/render an -*I believe you gave a very good account of what happened. (Cob.)*
relato, fazer um -*O delegado vai fazer um relato rápido dos fatos e enviar ao promotor. (FSP)*
account, take into -*Alicia communicated that she respected her daughter's feeling and took them into account. (Lerner)*
conta, levar/ter em -*Passados 28 anos, todos teriam guardado o segredo, mesmo levando-se em conta que ele hoje valeria milhões de dólares. (Veja)*
account, take ~ of -*... but now it is time to take account of the psychological side ... (Backlund)*
atenção, dar ~ a -*Não gostaram do filme, não dão muita atenção à figura do diretor ... (OESP)*

account $_{n.}$ **[explanation]**
account, call sb. to -*I was called to account for my conduct by the headmistress. (Cob.)*
contas, chamar alg. às -*Ele foi chamado às contas para explicar a razão da briga.*

account $_{n.}$ **[bank]** ...

(Tagnin 1999b: 406)

4.2. Portugiesische linguistische Ressourcen im Internet

Aus den demografischen Daten der portugiesischsprachigen Länder resultiert eine geringe Partizipation an romanistischen oder europäischen linguistischen Projekten. Unter den acht Sprachen des EuroWordNet ist Portugiesisch beispielsweise nicht vertreten. In einem aktuellen Sammelband zur romanistischen Corpuslinguistik[68], der aus fast hundert Beiträgen besteht, geht es in nur fünf Artikeln um die portugiesische Sprache. Vier davon beschreiben die Corpusannotation historischer portugiesischer Texte und die diachrone Entwicklung der sprachlichen Umgebung des Pronomen *se*. Der verbleibende Artikel[69] von Alencar berichtet über das Tool *Constructor*, das Suchanfragen für Recherchen in den WWW-Versionen portugiesischer Corpora konstruiert. Kodiert sind diese Corpora im Format der IMS Corpus Workbench (vgl. Kapitel 2.3.3). Zur Verfügung steht dort ein eigenes Anfragemodul, der **C**orpus **Q**uery **P**rocessor, der mit einer Anfragesprache arbeitet, die auf der Syntax regulärer Ausdrücke basiert.
Um in CQP eine Suchanfrage nach einem Adverb auf *-mente* zu stellen, das zwischen einem Substantiv und einem Adjektiv steht, ist folgender Suchausdruck möglich: "set Context 2s; [pos="N.*"&pos!="NUM.*"] [] {0,0} [word=".mente"&pos="ADV.*"] [] {0,0} [pos= "ADJ.*"] within 1s;" (Alencar 2002: 149). Alencar möchte mit dem *Constructor* besonders den Benutzern ohne Programmierkenntnisse die Suche in den Textcorpora erleichtern. Er entwickelt einen Katalog sprachlich ausformulierter Möglichkeiten, aus denen der Anwen-

68 Pusch, Claus D. / Raible, Wolfgang (eds.) (2002): *Romanistische Korpuslinguistik: Korpora und gesprochene Sprache.* Pusch, Claus D. / Kabatek, Johannes / Raible, Wolfgang (eds.) (2002): *Romanistische Korpuslinguistik: Korpora und diachrone Sprachwissenschaft* II. Tübingen, Gunter Narr.
69 Alencar, Leonel F. de (2002): "Der *Constructor* - ein interaktives Werkzeug für Recherchen in portugiesischen Korpora auf dem WWW". Band I: 147-154.

der auf einer graphischen Benutzeroberfläche über Eingabefelder und Schaltflächen wählt. Die Anfragen werden dann mit *Javascript* in CQP übersetzt.
Integriert war die Erstellung des *Constructors* in das Projekt *Processamento Computacional do Português*, das 1998 die Koordination der Aktivitäten auf dem Gebiet der automatischen Sprachverarbeitung des Portugiesischen übernahm. Auf verschiedene Forschungszentren verteilt wird das Projekt heute von der *Linguateca*[70] verwaltet. Ein Ziel der *Linguateca* ist es, allen Interessierten Informationen und Zugang zu existierenden elektronischen Medien und sprachverarbeitenden Tools des Portugiesischen zu bieten. Ein ausführlicher Katalog verweist auf Internetadressen, unter denen beispielsweise Tagger, Parser, Redaktionshilfen, Sprachsynthesesysteme, Maschinelle Übersetzer und Verbkonjugatoren (meist kostenfrei) zu konsultieren oder herunterzuladen sind. Auch die Forschungseinrichtungen, die sich mit der elektronischen Verarbeitung der portugiesischen Sprache befassen, sind hier auf einen Blick zu finden. Unter dem Katalog der Ressourcen stehen zahlreiche Links zu Corpora aller Sparten, Wörterbüchern und Lexika, didaktischem Material und weiteren Informationen, die die portugiesische Sprache betreffen.
Die *Linguateca* unterhält eigene Projekte und kollaboriert mit verschiedenen anderen Gruppen und bietet einen direkten Zugang zu den Ressourcen. Neben der Abfrage alignierter Parallelcorpora englisch/portugiesisch ist die morphosyntaktische Auszeichnung beliebiger Texte möglich, Baumbanken und eine Kollektion der kompletten Webseiten auf Portugiesisch aus dem Jahre 2003 sind verfügbar. Ein Schwerpunkt liegt auf der Vereinheitlichung, Bereitstellung und Abfrage großer Textcorpora über das WorldWideWeb. Im Projekt AC/DC (*Acesso a corpora / Disponibilização de corpora*) liegen alle Corpora mit ausführlichen Informationen und in zwei Versionen vor. Die eine enthält den Text nach einer linguistischen Vorverarbeitung, in der zusammenstehende Zeichen oder Zeichenketten tokenisiert und Sätze getrennt werden. In manchen Corpora werden zusätzlich Struktur des Textes und Metadaten mit SGML-Tags markiert. In Kapitel 5.1 werden Ausschnitte aus Corpora nach diesen Schritten der linguistischen Vorverarbeitung gezeigt.
Die linguistische Annotation der Corpora erfolgt mit dem syntaktischen Analysetool PALAVRAS[71]. Für jede Einheit des Corpus wird das Lemma, die grammatische Kategorie, morphologische Charakteristika und die syntaktische Funktion in Form von Attribut-Wert Paaren notiert. Der Anfang des ersten Satzes aus *Cetenfolha* hat in der annotierten Version folgendes Format, er stammt aus einem Teilcorpus, der von der *Linguateca* zu beziehen ist:

```
<s>
Brasília=Pesquisa=Datafolha      [Brasília=Pesquisa=Datafolha] PROP F S @SUBJ>
publicada   [publicar] V PCP F S @N<
hoje      [hoje] ADV @ADVL>
revela    [revelar] <fmc> V PR 3S IND VFIN @FMV
um        [um] <arti> DET M S @>N
dado      [dar] V PCP M S @>N
supreendente   [supreendente] <DERS> N M S @<SUBJ
$:
```

Attribute und Werte werden auf der Webseite des Projekts *Visual Interactive Syntax Learning* (VISL), zu dem PALAVRAS gehört, ausführlich erläutert. Beide Versionen der Corpora sind in der *Linguateca* im Format der Corpus Workbench gespeichert, als Anfragesprache für die annotierte Version wird CQP benutzt. Auf eine Integration des *Constructors* wurde verzichtet, dem unerfahrenen Benutzer werden auf einer zusätzlichen Seite Beispiele für Suchanfragen gezeigt und Links verweisen auf die Anfragemöglichkeiten

70 http://www.linguateca.pt
71 http://visl.sdu.dk/visl/pt

von CQP innerhalb der IMS Corpus Workbench. Die größten Corpora im AC/DC-Projekt sind *CETEMPúblico* und *CETENFolha*, aus ihnen erfolgt die Akquisition der Substantiv-Verb Kollokationen mit dem Programmpaket PECCI im nächsten Kapitel.[72]
Ebenfalls verfügbar über die *Linguateca* sind die partiell manuell korrigierten Baumbanken der ersten 1 Millionen Wörtern aus den beiden Corpora in dem Projekt *Floresta Sintá(c)tica*. Analysiert und annotiert wird ebenfalls mit PALAVRAS. Aus dem Quellcode im projektinternen VISL-Format wird die Visualisierung der Baumstruktur in einer grafischen Oberfläche generiert, er liegt auch konvertiert in die Ausgabeformate von Penn Treebank und Tiger-XML vor. Mehrere Abfragetools stehen zur Verfügung. Dependenzbäume werden direkt aus einem Format der Constraint Grammar generiert, das mit Tokens für die Kopf-ID's angereichert wird (#2->3), die auf den übergeordneten Knoten verweisen, und dadurch die Abhängigkeiten der einzelnen Wörter im Syntaxbaum verdeutlichen. Kopf-ID's sind in dem von *Floresta Sintá(c)tica* präsentierten CG-Format nicht vorhanden, wodurch der erste Satz von *Cetenfolha* in der automatisch generierten Baumbanken-Version identisch ist mit der annotierten Version der Corpora. Der erste Satz der manuell korrigierten Version unterscheidet sich in den fett markierten Werten von der automatisch generierten Analyse:

```
<s>
BRASÍLIA     [Brasília] PROP F S @NPHR
Pesquisa=Datafolha   [Pesquisa=Datafolha] NPROP F S @SUBJ>
publicada   [publicar] V PCP F S @IMV @#ICL-N<
hoje     [hoje] ADV @<ADVL
revela   [revelar] <fmc> V PR 3S IND VFIN @FMV
um       [um] <arti> DET M S @>N
dado     [dado] N M S @<ACC
supreendente   [supreendente] ADJ M S @N<
$:
```

Das Beispiel verdeutlicht die Fehleranfälligkeit der maschinellen Prozessierung und die Notwendigkeit der manuellen Korrektur. In den letzten beiden Zeilen wurde durch die falsche Analyse der Wortartenzugehörigkeit eine falsche grammatische Funktion und Abhängigkeit dargestellt. In den oberen Zeilen wird die syntaktische Kategorie präzisiert.
Das Angebot der *Linguateca* zeigt ein breites Spektrum an Institutionen und Projekten, die sich mit der elektronischen Verarbeitung der portugiesischen Sprache beschäftigen, und eine Vielzahl an Ressourcen und Tools, die disponibel sind. Die Aktivitäten auf dem Gebiet der elektronischen Sprachverarbeitung in den portugiesischsprachigen Ländern sind sehr viel ausgedehnter, als verfügbare Print-Medien vermuten ließen. Die Evaluierung der computationellen Prozessierung des Portugiesischen ist ein weiterer Schwerpunkt der *Linguateca*. Eine "Morfolympiade" des Portugiesischen wurde ausgetragen, und eine Bibliografie verweist auf Arbeiten zur Evaluierung auf verschiedenen Gebiete der elektronischen Sprachprozessierung. Der Katalog mit der allgemeinen Literatur zur maschinellen Sprachverarbeitung des Portugiesischen enthält fünf Artikel, die dem Titel nach Kollokationen behandeln (Dias/Nunes 2001, Fiker/Foley 2004, Leffa 1997, Sardinha 1999b, Tagnin 2000), von denen einer über das Internet oder hiesige Bibliotheken zu beziehen ist. Dias/Nunes (2001) zeigen einen genetischen Algorithmus für die Extraktion von Mehrworteinheiten (N-Grammen) aus nicht-annotierten multilingualen Textcorpora. Die schlechten Precision-Werte für das Französische und Portugiesische im Vergleich zum Englischen führen sie zurück auf die ausgeprägte Flexion sowie die Bildung komplexer Terme mit Präpositionen in den beiden romanischen Sprachen und die Favorisierung von 2-Grammen durch den Algorithmus.

72 CETEMPublico: **C**orpus de **E**xtractos de **T**extos **E**lectrónicos **M**CT/Público (Ministério da Ciência e da Tecnologia), CETENFolha: **C**orpus de **E**xtractos de **T**extos **E**lectrónicos **N**ILC/**F**olha (Núcleo Interinstitucional de Lingüística Computacional).

5. Extraktion der Substantiv-Verb Kollokationen - das Programmpaket PECCI

Das vorgestellte Anwendungsprogramm PECCI (Program for the Extraction of Collocations and Cluster Information) eignet sich zur Extraktion von Substantiv-Verb Kollokationen aus portugiesischen Corpora, die zwar eine linguistische Vorverarbeitung durchlaufen haben, aber keine linguistische Annotationen wie Lemmaangaben oder POS-Tags enthalten (vgl. Kapitel 2.2). Zwei weitere Perl-Programme (Cetemp und Cetenf) erledigen zunächst die Aufgabe der Corpusaufbereitung, um verschiedene Texte in einem einheitlichen Format zu speichern.

5.1. Corpusbeschreibung und Corpusaufbereitung

Grundlage für die Untersuchung bilden zwei Textcorpora, die über die *Linguateca* erhältlich sind. Das erste Corpus stammt aus Portugal, es umfasst 2.600 Ausgaben der Tageszeitung *Público* zwischen 1991 und 1998, dies entspricht ca. 174 Millionen Wörtern mit einer Größe von 1,2 Gigabyte. Das zweite Corpus kommt aus Brasilien. Die 365 Ausgaben der *Folha de São Paulo* von 1994 bilden den Text, mit ca. 24 Millionen Wörtern und 178 Megabyte. Genaue Informationen zu den Corpora und ihrer Aufbereitung sind unter http://www.linguateca.pt/CETEMPublico bzw. http://www.linguateca.pt/CETENFolha zu finden.
Beide Corpora sind innerhalb der *Linguateca* im Projekt AC/DC eingebettet (vgl. Kapitel 4.2), d.h. Wortkonkordanzen und -frequenzen sowie verschiedene Distributionen der POS-Tags können mit einer einzeiligen Abfrage, in CQP kodiert, ermittelt werden. Der Nachteil dieser Art der Anfrage liegt in der Wiederholung der Anfrage für jedes einzelne Wort und der Verarbeitung der Ergebnisse, da diese nicht automatisch weiterverwertet werden können, zudem bricht die Ausgabe nach 15.000 Fundstellen ab.
Beide Corpora liegen also in einer annotierten Form vor, zum Herunterladen der annotierten Versionen sind jedoch nur Teilcorpora verfügbar, die zur Kollokationsextraktion keine geeignet Größe haben (*CETEMPúblico* ca. 11 Millionen annotierte Wörter, *CETENFolha* ca. 7 Millionen annotierte Wörter).[73] Am IMS selbst liegt in der Corpus Workbench, die mit CQP als Anfragesprache arbeitet, ebenfalls nur die nicht annotierte Form des Textes vor. "The CQP Query Language Tutorial" (Evert 2005b) zeigt die sehr umfangreichen Möglichkeiten von CQP auf. Diese gehen weit über die verfügbaren Funktionen im AC/DC-Projekt hinaus. Vor allem im Hinblick auf Ausgabeformat und Speichermöglichkeiten der Suchergebnisse bieten sich verschiedene Lösungen, Makros können eingebunden und CQP als Child-Prozess in einem Perl-Programm gestartet werden. Durch das Einbinden von CQP wären die Module 1, 2 und 3 des im folgenden vorgestellten Perl-Programm redundant, man könnte direkt auf Frequenzdaten und Konkordanzen der Okkurrenzen und Kookkurrenzen der Lemmata zugreifen.
Das Programmpaket PECCI arbeitet mit den Textdateien, die in nicht annotierter Form vollständig von der *Linguateca* zur Verfügung gestellt werden. Die fehlenden POS-Informationen und Lemmaangaben müssen durch eine komplizierte Mustersuche mit den Deklinations- bzw. Konjugationsformen der gesuchten Substantive und Verben nachvoll-

73 Seit dem 25.10.2005 ist, als Ergebnis einer längeren Korrespondenz mit der *Linguateca*, nun auch die annotierte CQP-Version als Download verfügbar. Sie kann hier nicht mehr berücksichtigt werden, da der programmierpraktische Teil am Anfang der Diplomarbeit stand.

zogen werden. Die Gewinnung von Frequenzdaten und Konkordanzen aus großen Textcorpora ist sehr zeitintensiv. Anfragen auf der Corpus Workbench brauchen durch datenspezifische Kompressions-Algorithmen, die Binärkodierung und Indizierung der Corpora nur einen Bruchteil der Zeit. Mit dem Perl-Programm PECCI wird eine Möglichkeit der lexikalischen Akquisition von Kollokationen aus linguistisch nicht-annotierten Corpora vorgestellt, auch macht die Postprozessierung (Modul 4, 5, 6) einen großen Teil des Programms aus.[74] Die beiden von der *Linguateca* zur Verfügung gestellten Corpora haben eine linguistische Vorverarbeitung durchlaufen (vgl. Kapitel 2.2 und 2.3), d.h. die Satzgrenzen sind erkannt, die Wörter (hier nur teilweise) tokenisiert und die Textstruktur und weitere Metadaten mit SGML-Tags markiert. Da sich die Darstellung des portugiesischen und des brasilianischen Corpus leicht unterscheidet, werden die beiden Originalcorpora zunächst in ein einheitliches Format gebracht und die Wörter nummeriert.

CetemPublico1.7 (Originalcorpus):

```
<ext n=1 sec=clt sem=92b>
<t>
Um
revivalismo
refrescante
</t>
<p>
<s>
O
7
e
Meio
é
um
ex-libris
da
noite
algarvia
.
</s>
```

wird zu *Cetempúblico*:

```
<ext n=1 sec=clt sem=92b>
<t> 3: Um revivalismo refrescante </t>
<p>
<s> 13: O 7 e Meio é um ex-libris da noite algarvia . </s>
```

CETENFolha-1.0 (Originalcorpus):

```
<ext id=1 cad="Opinião" sec="opi" sem="94a">
<s> <t> PT no governo </t> </s>
<s> <a> Gilberto Dimenstein </a> </s>
<p>
<s> BRASÍLIA Pesquisa Datafolha publicada hoje revela um dado supreendente: recusando uma postura radical, a esmagadora maioria (77%) dos eleitores quer o PT participando do Governo Fernando Henrique Cardoso . </s>
<s> Tem sentido -- aliás, muitíssimo sentido . </s>
</p>
```

wird zu *Cetenfolha*:

```
<ext id=1 cad="Opinião" sec="opi" sem="94a">
<t> 3: PT no governo </t>
<a> 5: Gilberto Dimenstein </a>
<p>
<s> 33: BRASÍLIA Pesquisa Datafolha publicada hoje revela um dado supreendente : recusando uma postura radical , a esmagadora maioria ( 77 % ) dos eleitores quer o PT participando do Governo Fernando Henrique Cardoso . </s>
<s> 38: Tem sentido -- aliás , muitíssimo sentido . </s>
</p>
```

74 "CQP is a useful tool for interactive work, but many tasks become tedious when they have to be carried out by hand; macros can be used as templates, providing some relief; however, full *scripting* is still desirable (and in some cases essential). - similarly, the output of CQP requires post-processing at times: better formatting of KWIC lines (especially for HTML output), different sort options for frequency tables, frequency counts on normalised word forms (or other transformations of the values)" (Evert 2005b: 37).

Die Zahl zu Beginn eines Satzes gibt die Nummer des letzten Wortes im Satz wieder. Gezählt werden alle Wörter (enthält mindestens eines der Zeichen 'A-Za-zÀ-ü0-9') in Sätzen, Überschriften, Autorennamen und Listenelementen. Im Falle von CetemPublico1.7 wird der Text wieder in Zeilen geschrieben. Bei CETENFolha1.0 werden alle Satzzeichen ',?!...«»' (außerhalb der Metadaten) sowie die Klammern '()' tokenisiert, was im Originalcorpus nur mit dem Punkt '.' geschehen ist. Die SGML-Tags <s> und </s> werden analog zu CetemPublico1.7 nur noch um Sätze gesetzt. In den Exzerptionsdateien der Substantive, die von PECCI generiert werden, findet der Anwender eine um alle SGML-Tags reduzierte Ausgabe:

> 15285: « Mas hoje [ontem] , quando perceberam que não era para adoptar , começaram também a aparecer pessoas interessadas em receber também as mães » , explicou uma fonte do Fórum Estudante , que está a organizar a Missão Crescer em <Esperança> . *Cetempúblico*

> 17899: Vamos confrontar os titulares que tudo perderam com esses meninos que são a derradeira <esperança> de um título neste ano : enquanto os titulares de Telê privilegiam o meio-campo , com três, quatro , às vezes cinco volantes congestionando o setor , o que , na prática , ... *Cetenfolha*

Die Anzeige der Suchergebnisse im AC/DC-Projekt für *CETEMPúblico* und *CETENFolha* divergiert in der Darstellung der Metadaten (Anfrage: "[Ele]sperança"].

CETEMPúblico:

> Ext **139** (soc, 92b): Mas hoje [ontem], quando perceberam que não era para adoptar, começaram também a aparecer pessoas interessadas em receber também as mães», explicou uma fonte do Fórum Estudante, que está a organizar a Missão Crescer em **Esperança** .

CETENFolha:

> par Esporte-94a-des-3: Vamos confrontar os titulares que tudo perderam com esses meninos que são a derradeira **esperança** de um título neste ano: enquanto os titulares de Telê privilegiam o meio-campo, com três, quatro, às vezes cinco volantes congestionando o setor, o que, na prática, ...

Da im Zusammenhang mit der Kollokationsextraktion in dieser Arbeit die Metadaten eher unwichtig erscheinen und diese bei Bedarf vollständig über die aufbereiteten Corpora nachvollziehbar bleiben, wird die kurze numerische Angabe der genauen Fundstelle bevorzugt.
Für die Corpusaufbereitung sind die Programme Cetemp und Cetenf (Quellcode im Anhang) zuständig. Auf der beiliegenden CD2 sind die Corpora in ihrer komprimierten Form enthalten, das heißt sie müssen entpackt und durch das Ausführen von Cetemp und Cetenf in das geeignete Format gebracht werden, bevor man PECCI startet.

5.2. PECCIs Programmarchitektur

Die Extraktion von Substantiv-Verb Kollokationen aus einem linguistisch nicht annotierten Corpus bringt erhebliche Umstände bei der Implementierung eines entsprechenden Programms mit sich. Ein großer Teil des Mehraufwands entsteht durch die komplizierte Mustersuche der Konjugationsformen der einzelnen Verben.[75] Damit wird die meist von einem Tagger vorgenommene Lemmatisierung simuliert. Die gesamte Konjugation eines Verbs wird als regulärer Ausdruck formuliert, der als Suchmuster dient. Ausgelassen werden die Formen, die mit anderen Wortarten homograph sind, beispielsweise kann *morro* 'ich sterbe' oder 'Hügel' bedeuten. Das Lemma ist die Infinitivform des Verbs, auch die konjugierten Vorkommen werden unter diesem Eintrag gespeichert. Prinzipiell sollten alle Verben, die sich im Corpus befinden, lemmatisiert vorliegen. Im Programm beschränkt sich die Bearbeitung auf 226 Verben. Deren Frequenzen werden von Modul 1 gezählt und im Verzeichnis des zugrunde liegenden Corpus gespeichert.

75 Diese kann man den *Verbformen Portugiesisch zum Nachschlagen* von Freire (1985) entnehmen.

Das Auffinden der untersuchten Nomina gestaltet sich leichter. Wegen der im Portugiesischen fehlenden morphosyntaktischen Kennzeichnung der Deklination am Nomen, dienen die Singular- und Pluralform des Nomens als Lemmata. Singular- und Pluralform eines Nomens werden separat extrahiert, da sie sich mitunter im Kollokationsverhalten unterscheiden. Im Gegensatz zu den Verben, von denen zunächst nur die Frequenz im Corpus interessierte, werden von jedem Nomen die gesamten Konkordanzen extrahiert und als Subcorpus gespeichert.
Es werden exemplarisch 40 portugiesische Substantive der Gefühle behandelt, die in einem Sample zusammengefasst werden, das hier einem Wortfeld entspricht. Die Unterteilung in verschiedene Samples entspringt der Idee, die von Mel'čuk und Wanner (1994) vorgenommene Klassifizierung deutscher Gefühlssubstantive anhand ihrer semantischen Merkmale und der Beschreibung ihres Kollokationsverhaltens durch lexikalische Funktionen mit Ergebnissen für das Portugiesische zu vergleichen. Dabei entsprechen die 40 untersuchten portugiesischen Emotionsnomina nur in etwa den deutschen von Mel'čuk und Wanner aufgeführten 40 Gefühlssubstantiven (vgl. Abb. 9). Auch das Clustering der Kookkurrenzdaten mit Modul 5 und 6 geschieht auf der Grundlage von einem Sample. Es können beliebig viele weitere Samples initiiert und bestehende Samples um zusätzliche Nomina ergänzt werden. Die Ergebnisse werden für jedes Corpus separat gespeichert damit mögliche Unterschiede zwischen den Corpora nachvollziehbar bleiben. Die Exzerption der Nominakonkordanzen geschieht mit Modul 2.
In Modul 3 werden die Konkordanzen der Kookkurrenzen der Nomina mit den Verben extrahiert, archiviert und die Frequenzdaten verwaltet. Durch das Fehlen einer weiteren Tagging-Information, die angibt, ob bestimmte Verben als Voll- oder Hilfsverb agieren, ist es notwendig, diejenigen Verben, die in beiden Funktionen auftreten können, aus einem Subcorpus zu extrahieren, das um alle Okkurrenzen mit einem lemmatisierten Vollverb im Suchraum reduziert ist. Aus dem Subcorpus werden zum Schluss diejenigen Verben gesucht, die häufig als grammatischer Modifikator dienen. Damit soll die doppelte Extraktion von zusammengesetzten Zeiten, Passivkonstruktionen und Verbalperiphrasen vermieden werden. Die Konkordanz, die folgenden Nebensatz enthält: " *, que* ***tinha alimentado*** *a* ***<esperança>*** *de ver a chegada dos deputados portugueses ,* " (Cetempúblico, 745714) ('... hatte genährt ... Hoffnung ...'), wird nur als Kookkurrenz der beiden Wörter *alimentar esperança* gezählt und gespeichert. In den Konkordanzen der Verben sollten nur die Kookkurrenzen, in denen sie als Vollverb fungieren, enthalten sein, wie "*318503:* ***Temos <esperança>*** *de mais História e menos mito ,* " (Cetempúblico) ('Wir haben Hoffnung ...'). Da aber nur 226 Verben lemmatisiert sind, funktioniert dieses Verfahren nur bedingt.
In Modul 3 wird auch die Größe des Fensters um das Substantiv festgelegt, in dem nach den Verben zu suchen ist. Trotz der mangelnden linguistischen Annotation der Corpora kann durch gezielte Suchanfragen zu bestimmten Wortarten eine vorher festgelegte Kollokationsart extrahiert werden. Die syntaktische Struktur lässt sich für Nomen-Verb Kollokationen im Portugiesischen bedingt über die Wortstellung simulieren. Meist lässt sich das nächste rechts vom Verb stehende Nomen als dessen Objekt identifizieren, während es sich bei einem links vom Verb stehenden Nomen sowohl um dessen Subjekt als auch dessen Objekt handeln kann (das Vorkommen als Subjekt ist häufiger). Der Suchraum wird in der Default-Einstellung auf drei Wörter vor dem Substantiv und eines dahinter festgelegt, da sonst zu viele falsche Kookkurrenzen extrahiert werden. Er kann bei Bedarf global oder einzeln an jedem Verb verändert werden.

Außerdem werden in Modul 3 Informationen zu den Umgebungsdaten einer Substantiv-Verb Kookkurrenz gegeben. Darunter wird die Ermittlung und Zählung der unmittelbar zwischen und hinter den beiden Wörtern vorkommenden frequenten Wörter verstanden, worunter Artikel, Präpositionen und Konjunktionen fallen. Genauere Angaben z.B zu. präferiert vorkommenden Adjektiven zwischen einer bestimmten Substantiv-Verb Kookkurrenz, sind wegen der fehlenden POS- und Lemmainformation in der Implemen-tierung zu aufwendig und daher nicht enthalten. Auch der Subkategorisierungsrahmen kann aufgrund der mangelnden Annotation der Corpora nicht systematisch erfasst werden. Er kann nur vom Menschen aus den Kookkurrenzkonkordanzen ausgelesen werden. Modul 3 liefert die ersten für den Lexikografen interessanten Daten, die Kookkurrenzkonkordanzen für die Nomina mit den Verben. Die Kookkurrenzfrequenztabellen bilden die Eingabe für die Berechnungsmodelle der folgenden Module.
Modul 4 berechnet das Kollokationspotenzial mit einem vom Benutzer wählbaren Assoziationsmaß (vgl. Kapitel 2.1) für die Kookkurrenzen und stellt das Ergebnis für jedes Nomen eines Samples einzeln und für das gesamte Sample in Form einer Rankingliste zur Verfügung (Ausschnitte aus diesen Ausgabedateien werden im Anhang gezeigt). In diesem Zusammenhang erscheint es angebracht für den lexikografischen Gebrauch eines der oben vorgestellten statistischen Assoziationsmaße zum Ranking zu benutzen und die Werte eines zweiten Maßes zum Vergleich daneben darzustellen (vgl. Kapitel 6). In der Ausgabe erscheint auch die Suchraumeinstellung, die Frequenz der Substantive und der Verben und deren Kookkurrenz. Eine ähnliche Ausgabe kann man auch sortiert nach Verben erhalten.
In Modul 5 und 6 ist das Clusterverfahren K-Means implementiert, das die Kookkurrenzfrequenzen zum Aufbau einer Matrix verwendet, die den Vektorraum der Nomina (Modul 5) oder Verben (Modul 6) repräsentiert. Die Cluster berechnen sich über die Minimierung der euklidischen Distanzen der Vektoren. Die Ergebnisse und Berechnungsmodelle von Modul 5 und 6 werden ausführlich in Kapitel 7 behandelt.
Die Programmarchitektur (Abb. 10) gibt in grafischer Form noch einmal eine Übersicht über die verschiedenen Module:

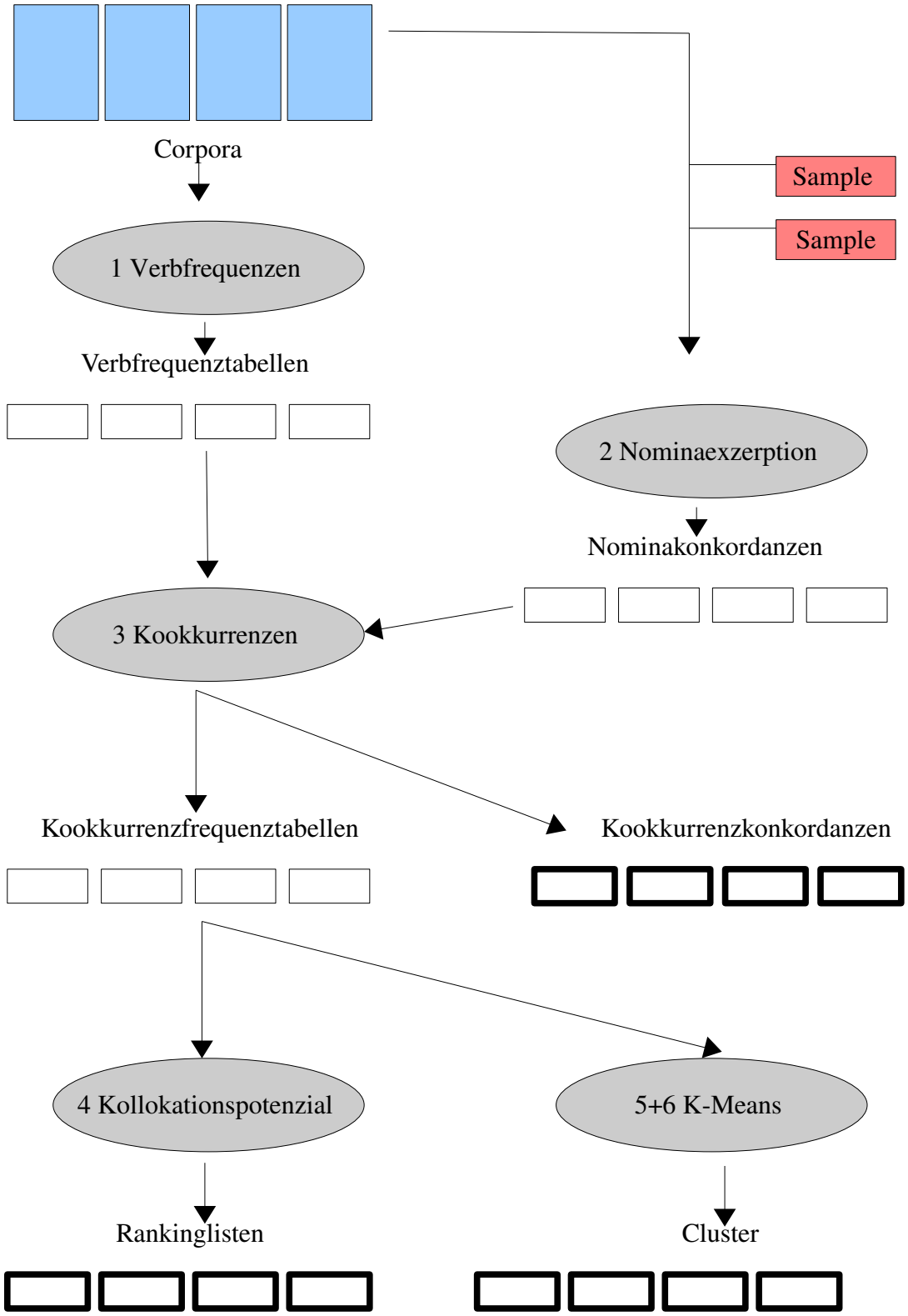


Abb. 10: PECCIs Programmarchitektur

Mit PECCI können die einzelnen Extraktions- und Exzerptionsschritte separat gesteuert werden. Die Corpora und Samples, sowie die Module und deren Parameter werden über einen Benutzerdialog ausgewählt. Zunächst trifft man die Corpus- und Samplewahl, die beschriebenen Module entsprechen den nummerierten Optionen im dritten Schritt.

```
heike@linux:~/D> ./Pecci
Möchten Sie ein bereits existierendes Sample bearbeiten?
1 = Substantive der Gefühle = Sample 'Sentimento'
X = Neues Sample
1
Mit welchem Corpus möchten Sie arbeiten?
1 = Cetempublico
2 = Cetenfolha
3 = Testcorpus 2 Millionen
1
Folgende Möglichkeiten stehen zur Auswahl
0 = Fundstellen neuer Verben
1 = Anzahl der Fundstellen der Verben
2 = Extraktion (zusätzlicher) Substantive
3 = Extraktion der Substantiv-Verb Kollokationen
4 = Berechnung des Kollokationspotenzials
5 = Clusterverfahren K-Means Nomina
6 = Clusterverfahren K-Means Verben
Es kann nur ein Modul oder mehrere Module ausgewählt werden.
(Eingabe getrennt durch Leerzeichen)
1 2 3 4 5 6
Die Extraktion der Verben ist abgeschlossen.
Bisher wurden 4 Substantive extrahiert (Singular- und Pluralform separat):
admiração admirações aflição aflições susto sustos vergonha vergonhas
Geben Sie nun die (zusätzlich) zu extrahierenden Substantive in Singular-
und Pluralform durch Leerzeichen getrennt ein.
esperança esperanças
Wie möchten Sie das Kollokationspotenzial berechnen?
Es können eins oder zwei der Assoziationsmaße gleichzeitig ausgewählt
werden. Default-Einstellung sind t-score und Mutual Information.
1 = t-score
2 = log-likelihood
3 = chi-square
4 = Mutual Information
1 2
Wo soll die Signifikanzgrenze liegen?
1
Wie sollen die Clusterzentren von K-Means für Nomina bestimmt werden?
1 = Zufallsgenerierte Clusterzentren
2 = Selektion des ersten Clusterzentrums und der Schrittgröße
3 = Einzelne Nomina bestimmen die Ausgangsclusterzentren
4 = Zufallsgenerierte Clusterzentren, K-Means wird 100 mal gestartet und
die Ergebnisse gesammelt
1
Wie viele Cluster sollen gebildet werden?
10
Wie sollen die Clusterzentren von K-Means für Verben bestimmt werden?
1 = Zufallsgenerierte Clusterzentren
2 = Selektion des ersten Clusterzentrums und der Schrittgröße
3 = Einzelne Verben bestimmen die Ausgangsclusterzentren
3
Die nummerierten Verben finden Sie in SentimentoCetemp/ClusterVerb/
zinfoverb.
Eingabe der einzelnen Clusterzentren gefolgt von einem Leerzeichen.
1 4 33 54 67 101 118 204
heike@linux:~/D>
```

Die Programmdokumentation von PECCI im Anhang bietet nähere Angaben zu den einzelnen Dateien und Modulen und zur Verzeichnisstruktur. Im Anhang sind auch die Installationshinweise zu den beiliegenden CDs zu finden, die den Quellcode, die Corpora in komprimierter Form, und die Extraktionsergebnisse für das Sample der Substantive der Gefühle enthalten. Ebenfalls im Anhang werden die relevanten Ausgabedateien partiell in gedruckter Form wiedergegeben.
Durch die Modularisierung des Programms kann man die Berechnung des Kollokationspotenzials und das Clusterverfahren K-Means beliebig oft mit verschiedenen Parametern ausführen, ohne die zeitintensiveren Module 1, 2 und 3 neu zu kompilieren. Bei einem Corpus in der Größe von *Cetempúblico* und der hier untersuchten Anzahl von Substantiven und Verben benötigt Modul 1 mit einem 1,66 GHz Prozessor ca. 48 Stunden, Modul 2 ca. 8 Stunden und Modul 3 ca. 20 Minuten, was sowohl auf die Größe des Corpus (1,2 Gb) als auch auf die zeitintensive Mustersuche (in Perl) zurückzuführen ist. Modul 4, 5 und 6 berechnen mathematische Modelle anhand der Frequenzdaten und laufen in Sekundenschnelle. Durch die Verwaltung der Konkordanzen und Frequenzdaten in Textdateien entsteht ein zusätzlicher Speicherbedarf. Wünschenswert wäre ein Zugriff auf das Originalcorpus über Pointer. Diese könnten zusammen mit anderen Ergebnissen der Corpusprozessierung in einer Datenbank abgelegt werden, wie es beispielhaft in der IMS Corpus Workbench geschieht (vgl. Kapitel 2.3.3).

5.3. Evaluierung der Extraktionsergebnisse

Im Folgenden soll es nicht um eine Bewertung der Gültigkeit der extrahierten Substantiv-Verb Kookkurrenzen als Kollokationen gehen. Der Wert der Extraktionsergebnisse in Form von Rankinglisten für den Lexikografen wird erst das Thema von Kapitel 6 sein. Vielmehr werden die Möglichkeiten und Grenzen eines Programms, das mit einem linguistisch nicht annotierten Corpus arbeitet, gezeigt. Der Aufwand bei der Programmierung und Validierung verdeutlicht, wie essentiell ein Text für die Corpusabfrage ist, der Informationen zu POS-Tags und Lemmata enthält. Die Evaluierung der Kookkurrenzdaten bestätigt weiterhin, dass ein fensterbasiertes Verfahren immer nur einen Teil der Kollokationen extrahiert (bei einem kleinen Fenster) oder (bei einem großen Fenster) erheblich verfälschte Ergebnisse liefert. Anhand einer Auswertung der Kookkurrenzdaten von PECCI lassen sich Recall und Precision für unterschiedliche Suchraumeinstellungen berechnen, positionelle Eigenheiten verschiedener Verben beschreiben und die syntaktischen Relationen der Substantiv-Verb Kollokationen im Portugiesischen exemplifizieren. Zunächst zeigt ein Vergleich der Extraktionsergebnisse von PECCI mit den Konkordanzen der *Linguateca*, die auf linguistisch annotierten Corpora beruhen, Stärken und Schwächen in beiden Versionen.[76]

5.3.1. Anfragen und Ergebnisse bei PECCI und der Linguateca

Die aufwendige Lemmatisierung der Verben in PECCI, über die auch Homographen ausgefiltert werden, birgt die erste Fehlerquelle im Programm. Zwar ist das Vorkommen der Konjugationsform des Verbs im Vergleich zum gleich geschriebenen Wort in einer anderen

76 Die maschinelle Akquisition lexikalischer Daten ist bis heute nicht ganz fehlerfrei. Wie viele inkorrekt klassifizierte Extraktionsdaten sich in den Ergebnissen befinden und wie viel Sprachmaterial aufgrund einer falschen Zuordnung fehlt, ist stark abhängig von der untersuchten Sprache, der Fragestellung, dem Grad der linguistischen Corpusaufbereitung und den dazu verwendeten Tools.

Wortart meistens weniger frequent, doch fehlen homographe Konjugationsformen in der Zählung und den Konkordanzen. So wurde die 1. Person Singular Präsens von *morrer* im Suchausdruck getilgt. Dadurch fallen 67 Vorkommen für *morro* in der Bedeutung 'ich sterbe' weg, aber es fehlen auch die 310 Okkurrenzen von *morro* für 'Hügel'. Könnte man in diesem Fall noch argumentieren, dass im Suchraum -3 bis +1 Wörtern um ein Gefühlssubstantiv die verbale Bedeutung sehr viel wahrscheinlicher ist, wäre bei anderen Verben die Lesart als Substantiv frequenter, wie für *manifesto* ('Manifest' oder 'ich bekunde') in: "132493608 ... <manifesto> de <amor> ..." (Cetempúblico) ('Manifest der Liebe'). Die Vorkommen polysemer und homonymer Verbformen werden daher immer ignoriert, um falsche Extraktionen zu vermeiden, was zu einer Minderung des Recalls führt.[77]
Die Ambiguität der Wortformen stellt auch für den Tagger der *Linguateca* ein Problem dar(zu den Tools der Corpusaufbereitung der *Linguateca* vgl. Kapitel 4.2). In den 310 Okkurrenzen, in denen *morro* mit dem Suchausdruck [word="morro" & pos="N.*"] extrahiert wird, befinden sich zahlreiche Beispiele, in denen *morro* in verbaler Funktion vorkommt: "*Ext 511727 (pol, 91b): «Por ela eu choro de alegria e* ***morro de tristeza*** *..."* (Cetempúblico) ('sterbe ich vor Traurigkeit'). Betrachtet man die Suchergebnisse für *manifesto* als Verbform zeigt sich, dass 10 der 20 angezeigten Okkurrenzen in diesem Kontext falsch getaggt sind: "*Ext 500123 (pol, 96a): ... do que de um* ***manifesto*** *politicamente articulado* ." (Cetempúblico) ('ein politisch artikuliertes Manifest'). Als Substantiv kommt *manifesto* weitaus häufiger vor, in den 1968 Okkurrenzen scheint sich *manifesto* auf den ersten Blick aber nicht in seiner verbalen Funktion zu befinden.
Während der Programmierung von PECCI traten Fehler auf, die sich erst bei einer Durchsicht der extrahierten Daten zeigten. Sie weisen weitere Parallelen zu falschen POS-Tags und einer fragwürdigen Corpusverarbeitung in der *Linguateca* auf. Wieder betreffen sie homographe Wortformen. Anfänglich wurde in PECCI die Großschreibung des Anfangsbuchstabens erlaubt, um die Verben und Substantive auch am Satzanfang zu extrahieren. Dadurch wurden jedoch auch Kombinationen wie *Restaurar a Esperança* gewonnen, in denen das Verb substantiviert im Namen einer militärischen Operation vorliegt: "*7998229: Mas a aprovação da Operação* ***Restaurar a <Esperança>*** *, ...* " (Cetempúblico). Die eigentlichen 6 Vorkommen in der beabsichtigten Substantiv-Verb Konstruktion im Sinne von 'Hoffnung wiedererlangen' wurden um 132 falsche Positive angereichert. Die Extraktion wurde später dahingehend präzisiert, die Großschreibung des ersten Buchstabens auf das erste Wort im Satz zu beschränken, um dadurch die Extraktion von Eigennamen zu vermeiden (Substantive schreiben sich im Portugiesischen sonst klein). In der *Linguateca* werden die einzelnen Wörter des Ausdrucks *Restaurar a Esperança* nur manchmal als Bestandteile eines Eigennamens erkannt und als "PROP" getaggt. In vielen Fällen erhält *Restaurar* auch im substantivierten Ausdruck den verbalen Tag "V" mit dem Attribut "INF" und ist in den Konkordanzen auf die Anfrage [word="Restaurar" & pos="V.*"] zu finden.

77 Neben den Substantiven sind es die Adjektive, die häufig homograph zu einer Konjugationsform des Verbs sind, meistens zur Form des Partizip II. Ein deutsches Beispiel ist: 'er hat erfahren - er ist erfahren'. In der portugiesischen Übersetzung verhalten sich die Formen äquivalent (*tem experimentado - está experimentado*) und werden in solchen Fällen aus den Suchmustern getilgt. Die Unterscheidung zwischen Adjektiv und Partizip II ist nicht immer eindeutig und wird kontrovers diskutiert (vgl. Stadler (1996)). Beispiele von Homonymie, bei denen die Bedeutungen der Wortform etymologisch nicht verwandt sind (wie im Fall von *morro*), sind seltener. Ein Homograph ist ein Wort das polyseme und/oder homonyme Ambiguitäten in seiner Schreibweise vereint.

Ein ähnliches Phänomen ist zu beobachten, wenn es um die Gleichlautung von Eigennamen, Akronymen und Gefühlssubstantiven geht. Wird der erste Buchstabe groß geschrieben, kann *Ira*, der 'Zorn', auch ein weiblicher Vorname sein. In den Akronymen sind alle Buchstaben groß - *IRA* steht für 'Irish Republican Army', *IRAS* ist die Abkürzung für 'Infra-Red Astronomic Satellite'. Dieses Problem wurde in PECCI für die Extraktion der Substantive zunächst durch das Suchmuster abgefangen, das Case-Insensitivity nur für den ersten Buchstaben des Suchausdrucks generiert, und durch die Beschränkung der Großschreibung auf das erste Wort im Satz. Der Zorn (*ira* 839*)* kommt im Vergleich zum Akronym seltener vor (*IRA* 3470), der Eigenname *Ira* ist 430 mal präsent. Im annotierten Corpus der *Linguateca* ist festzustellen, dass die Schreibung von *IRA* im Originalcorpus abgeändert wurde zu *Ira*. Dies führt zu einem Ergebnis, das bei einer Suche nach [lema="Ira"] 3383 Fundstellen mit einem "PROP"-Tag liefert, sie enthalten das Akronym und den Vornamen, die 15 Fundstellen mit dem "N_prop"-Tag bestehen aus dem Akronym *IRAS*. Bei einer Suche nach [lema="IRA"] verbleiben 20 Fundstellen, in denen das Akronym *IRA* vollständig groß geschrieben bleibt.[78] Warum die Corpusveränderung erfolgte ist nicht ganz klar, sie ist falsch und macht die Sätze zunächst schwer lesbar, da Großbuchstaben auch im Portugiesischen das übliche Kennzeichen eines Akronyms sind.

Um die komplexen Suchausdrücke der Verblemmata nicht noch weiter zu verkomplizieren, wurden die Verben in PECCI zunächst mit dem Asterisk (*) im Suchausdruck als Platzhalter für alle möglichen Pronomina extrahiert. Im Portugiesischen werden an das Verb mit einem Bindestrich Reflexiv- und unbetonte Personalpronomina angehängt. Im Kasus differente Pronomina können miteinander kombinieren und verändern dabei ihre Form, nach bestimmten Endungen des Verbs werden den Pronomina weitere Buchstaben vorangestellt (der letzte Buchstabe der Verbform entfällt dann) und in bestimmten Zeiten können die Reflexiv- und Personalpronomina sogar zwischen dem Stamm des Verbs und seiner Endung stehen.[79]

Durch den Asterisk als Platzhalter für das Pronomenparadigma waren aber nicht beabsichtigte Kombinationen in den Verbfundstellen enthalten. Am auffälligsten verhielt sich *cessar-fogo* der 'Waffenstillstand', der den eigentlichen 2308 Vorkommen des Lemmas *cessar* in verbaler Funktion 6191 falsche Fundstellen hinzufügte.[80] Der Fehler wurde behoben durch das Speichern des Pronomenparadigmas in einer Variablen, die den Asterisk im Suchausdruck der Verben ersetzt. Ausformuliert hat der Suchausdruck in PECCI für das Verb *cessar* nun die folgende Form (Pronomina sind fett gedruckt), er wäre noch kürzer zu formulieren und damit zu optimieren, was aber auf Kosten der Übersichtlichkeit und der einfachen Wartbarkeit ginge:

cess(o|ar|as|a|amo(s?)|ais|am|ada|adas|ado|ados|ava|avas|ávamo(s?)|áveis|avam|ei|aste|ou|ámo(s?)|astes|aram|ara|aras|áramo(s?)|áreis|arei|arás|ará|aremo(s?)|areis|arão|e|es|emo(s?)|eis|em|asse|asses|ássemo(s?)|ásseis|assem|ares|armo(s?)|ardes|arem|aria|arias|aríamo(s?)|aríeis|ariam|ando|á)(-(**me|te|se|nos|vos|o|lo|no|os|los|nos|a|la|na|as|las|nas|mo|ma|mos|mas|to|ta|tos|tas|lho|lha|lhos|lhas|lhe|lhes|no-(lo|la|los|las)|vo-(lo|la|los|las)**)))?(-(ei|ás|á|emos|eis|ão|ia|ias|íamos|íeis|iam))?

In der *Linguateca* ist der Suchausdruck sehr viel kürzer [lema="cessar" & pos="V.*"]], soll das Verb mit einem klitischen Pronomen erscheinen wird der Suchausdruck erweitert [lema="cessar\+.*" & pos="V\+PERS.*"].

78 Mit PECCI werden 72 Okkurrenzen von *IRA* zusätzlich zur Ausgabe in der *Linguateca* extrahiert, was sich aus dem häufigen Auftreten des Akronyms in Überschriften erklärt, die man in den Konkordanzen der *Linguateca* nicht findet.

79 Vgl. Hundertmark-Santos Martins (1982: 118-124, 308-309).

Durch die Beschränkung der Großschreibung auf den ersten Buchstaben des ersten Wortes im Satz werden Verben und Nomina in ihrer Funktion als Eigennamen von PECCI nicht mehr extrahiert (es sei denn sie stehen am Anfang des Satzes). Auch die Fehler, die sich durch einen unpräzisen Suchausdruck ergaben, sind durch eine Programmmodifikation behoben. Das Problem der fehlenden Fundstellen für homographe Verbformen ist auf diesem Wege jedoch nicht lösbar.
Eine erhebliche Diskrepanz ergibt sich bei der Zählung der Wörter und Sätze in den Corpora durch PECCI und den diesbezüglichen Angaben auf der Webseite der *Linguateca*. Die Angaben zu *Cetempúblico* belaufen sich auf 191.687.833 Wörter in 7.082.094 Sätzen. Als Wörter werden alle Tokens gezählt, die einen Buchstaben oder eine Zahl enthalten. Die Zählung der Wörter findet bei PECCI in allen durch SGML-Tags markierten Strukturen statt, die Corpustext enthalten. Dies können Titel, Autorennamen, Listenelemente und Sätze sein. Dennoch werden nur 174.184.724 Wörter gefunden. Als Sätze werden in *Cetempúblico* offenbar nur die Strukturen gezählt, die von den Satz-Tags <s> </s> umgeben sind. Würde man Titel, Autorennamen und Listenelemente mitberechnen, käme man auf über 8 Millionen Sätze. Die von PECCI gefunden Anzahl an tatsächlichen Sätzen (7.049.916) ist etwas geringer als die in der *Linguateca* genannte. Die leichte Divergenz mag darauf beruhen, dass die aktuelle Distribution des Corpus im Internet die neueste Version des Textes enthält, während sich die Angaben der Webseite der *Linguateca* auf eine frühe Version des Corpus beziehen. Diese wurde im Verlauf der Jahre mehrmals von festgestellten Fehlern bereinigt, worunter vor allem die Eliminierung von Dubletten und Buchstabenfolgen mit ungültigen Zeichen fällt. Deren Anzahl entspricht in etwa der Verminderung der Sätze von Version 1.0 zu Version 1.7, die als Download zur Verfügung steht. Der Unterschied in der Wortanzahl kann an diesem Sachverhalt aber nicht liegen, die Abnahme der Sätze beträgt ca. 0.5%, die Differenz in der angegebenen Wortanzahl zu PECCI aber über 9%. Möglicherweise wurden von der *Linguateca* die Elemente in den Metadaten als Wörter gezählt (<*ext n=6686 sec=pol sem=92a*>).
Auch für *Cetenfolha* wurden statt der von der *Linguateca* genannten 25.475.272 Wörter mit PECCI nur 23.461.475 gefunden, wiederum in allen durch SGML-Tags gekennzeichneten Textbestandteilen, nicht aber in den Angaben der Metadaten. Die Anzahl der tatsächlichen Sätze in *Cetenfolha* beträgt 1.295.838. Bei der *Linguateca* werden hier aber, im Gegensatz zum Verfahren bei *Cetempúblico,* die Zeilen mit Autorennamen, Listenelementen und Titeln mitgezählt, was dann genau die angegebenen 1.597.807 Zeilen mit Corpustext ergibt.
Die Satzanzahl ist für die weitere Verwertung der Frequenzdaten z.B. mit den statistischen Assoziationsmaßen nicht relevant, sie dient ausschließlich Informationszwecken. Die Anzahl der Wörter im Gesamtcorpus hingegen fließt bei der Berechnung dieser Werte immer mit ein. Es zeigt sich jedoch, dass beide Corpora so groß sind, dass die divergente Wortanzahl bei *Cetempúblico* keine, bei *Cetenfolha* nur minimale Abweichungen zur Folge hat.

80 In den deutsch/portugiesischen Wörterbüchern findet man *cessar-fogo* nicht als Übersetzung für 'Waffenstillstand' (auch nicht für 'Waffenruhe'). Im *Langenscheidts Taschenwörterbuch Portugiesisch* (2001) und dem *Dicionário de Alemão-Português* (1989) wird stattdessen *armistício* angegeben. In *Cetempúblico* ist dieses Wort nur 267 mal vertreten. Allein im *PONS Standardwörterbuch* (2002) ist das viel frequentere *cessar-fogo* als Übersetzung aufgeführt. Anhand einer kurzen Durchsicht der Fundstellen ist schnell klar, dass *armistício* eher einen 'Waffenstillstandsvertrag' bezeichnet, wofür auch das häufige Auftreten des Kollokats *assinar* ('unterzeichnen') spricht, während *cessar-fogo* die allgemeinere und geläufigere Übersetzung von 'Waffenstillstand' ist.

5.3.2. Verbindividuelle Suchraumeinstellung

Dieser Teil der Evaluierung beschäftigt sich mit den Auswirkungen der Veränderung der Fenstergröße auf die Recall- und Precision-Ergebnisse bei der Extraktion der Kookkurrenzdaten. Damit ist eine manuelle Durchsicht eines Teils der Ergebnisse unter dem Aspekt der syntaktischen Gültigkeit der extrahierten Konkordanzen gemeint. Die individuelle Einstellung des Suchraums wird in PECCI am Verb vorgenommen und nicht am Substantiv. Darin spiegelt sich die Ansicht wieder, dass bestimmte Verben in einheitlicher Weise mit bestimmten Gruppen von Substantiven kollokieren. Bestimmte Verben zeigen mit einigen Teilnehmern eines semantisch motivierten Wortfeldes (wie den Substantiven der Gefühle) ein relativ einheitliches, starkes Kollokationsverhalten. In diesem Fall kann man das Fenster vergrößern, in dem nach den Substantiv-Verb Kollokationen gesucht wird. Auf diese Weise lassen sich die Recall-Ergebnisse für bestimmte Kollokationen auf fast 100% steigern, ohne dass die Precision-Ergebnisse darunter leiden.
Die Substantiv-Verb Kollokationen *alimentar ódio* und *nutrir ódio* werden aus *Cetempúblico* bei der Default-Suchraumeinstellung von drei Wörtern vor dem Substantiv und einem dahinter 15 mal bzw. 13 mal extrahiert (*alimentar* und *nutrir* bedeuten '(er)nähren', in der Kombination mit Gefühlssubstantiven haben beide Ausdrücke die Bedeutung 'hegen' oder 'nähren') . Der t-Test platziert *alimentar* auf Platz 5 und *nutrir* auf Platz 8 der Rankingliste mit den Werten 3.83 bzw. 3.60. Die Mutual Information ordnet der Kollokation *nutrir ódio* den Wert 7.46 zu, der im Vergleich zu *alimentar ódio* (MI = 4.61) sehr viel höher liegt, als die Differenz ihrer Frequenz vermuten lässt, dies liegt daran, dass *nutrir* im Gesamtcorpus nur 555 mal vorkommt, während *alimentar* 11.053 mal erscheint. Für diese beiden Substantiv-Verb Kombinationen, die als Kollokationen auf jeden Fall in einem Wörterbuch zu führen sind, könnte man die Anzahl der Fundstellen bei gleich-bleibenden Precision-Werten noch einmal erheblich steigern, würde man die Suchraumein-stellung vergrößern. Betrachtet man die 20 Okkurrenzen von *ódio*, in denen *nutrir* satzintern, aber außerhalb der Default-Einstellung des Suchraums von -3/+1 vorkommt, stellt man fest, dass sich in 18 der 20 Fälle *nutrir* ebenfalls auf *ódio* bezieht, d.h., dass 18 Kookkurrenzen als falsche Negative zählen: "... -- *um* ***<ódio>*** *como o que Jacques Chirac e François Mitterrand* ***nutriam*** *um pelo outro , ...*" (Cetempúblico, 67127825) (' -- ein Hass, wie der den Jacques Chirac und François Mitterrand gegeneinander hegten, ...'). Ihr Vorkommen wird, obwohl die Wörter in der gewünschten syntaktischen Relation stehen, nicht extrahiert. Der Recall wird nach folgender Formel berechnet:

$$Recall = \frac{true\ positives}{true\ positives + false\ negatives} \qquad Recall\ 'nutrir\ ódio' = \frac{13}{13 + 18} = 0.42$$

Demnach liegt der Recall von *nutrir ódio* bei einer Suchraumeinstellung von -3/+1 nur bei 42%. Die Formel für die Precision ist:

$$Precision = \frac{true\ positives}{true\ positives + false\ positives} \qquad Precision\ 'nutrir\ ódio' = \frac{13}{13 + 0} = 1.0$$

Die Precision liegt bei der kleinen Fenstergröße bei 100%, da keine falschen Positive (Kookkurrenzen, in denen die beiden Wörter nicht in der gewünschten syntaktischen Beziehung stehen) in den Extraktionsergebnissen zu finden sind. Würde man das Fenster auf 12 Wörter vor und hinter dem Substantiv erhöhen, erhielte man 30 richtige Positive und die

F-measure ((Recall + Precision) : 2) würde auf 97% steigen. Eine Kookkurrenz wird auch in diesem großen Suchraum nicht extrahiert, da sich die beiden Wörter noch weiter entfernt voneinander befinden. Die Konkordanz enthält jetzt ein falsches Positiv, in dem sich *nutrir* nicht auf *ódio* bezieht.
Ähnlich verhält es sich mit der Kollokation *alimentar ódio*. Man könnte den Recall bei einer Ausdehnung des Suchraums erheblich verbessern. Hier kann man aber anhand des deutlich geringeren Wertes der Mutual Information schon ahnen, dass das Kollokationsverhalten nicht ganz so gefestigt ist. Den 15 Kookkurrenzen innerhalb des kleinen Fensters von -3/+1 stehen 23 Sätze gegenüber, die die Kollokation satzintern aber außerhalb des engen Suchraums enthalten. In den 23 Sätzen kommen 8 Sätze vor, in denen sich *alimentar* und *ódio* nicht aufeinander beziehen. Würde man die Fenstergröße wie bei *nutrir* wieder auf ±12 ändern, hätte man in den nun 32 extrahierten Kookkurrenzen 5 falsche Sätze. Dies ließe sich vermeiden, würde man die Fenstergröße nur auf ±6 ändern, in den 27 Kookkurrenzen innerhalb dieses Suchraums befindet sich dann nur ein falscher Satz. Der Recall beträgt bei einer Einstellung von ±6 Wörtern 89%, die Precision liegt bei 96%. Bei einer Fenstergröße von ±12 ist der Wert des Recalls zwar 96%, die Precision sinkt aber auf 84 %. Der F-Score bleibt annähernd gleich.
Auswirkungen hätte die Vergrößerung des Suchraums auch auf das Rankingverhalten. Mit der optimalen Einstellung von ±12 für *nutrir* und ±6 für *alimentar*, stünde *nutrir* in der Rankingliste von *ódio* nun auf Platz 2 (5.38), *alimentar* auf Platz 3 (5.16) - behält man für die anderen Verben die enge Suchraumeinstellung bei. Beim Substantiv *admiração* ('Bewunderung', 'Verwunderung') würde *nutrir* von Platz 6 auf Platz 3 wandern. *Alimentar* hingegen spielt als Kollokat bei *admiração* nur eine untergeordnete Rolle, im Vergleich zu *nutrir admiração* (43 Kookkurrenzen) kommt *alimentar admiração* nur 5 mal vor. Umgekehrt verhält sich die Verteilung der beiden Kollokate bei *esperança,* mit der Default-Suchraumeinstellung -3/1 ist *alimentar* mit einem Vorkommen von 169/181 (Sg./Pl.) sehr frequent, während *nutrir* nur 8 mal (mit der Pluralform) erscheint. Wie dieses kleine Experiment zeigt, sind die vermeintlich synonymen Verben auch innerhalb eines Wortfeldes nur bedingt austauschbar, in Kapitel 6 wird deutlich, dass sie sich (mitunter) auch in ihrem Subkategorisierungsverhalten unterscheiden.
Als äußerst problematisch (weil zeitaufwendig) erweist es sich, eine für jedes Verb optimale Suchraumeinstellung individuell nach der manuellen Durchsicht der Ergebnisse festzulegen. Für bestimmte Kookkurrenzen würden sich die Anzahl der richtig extrahierten Fundstellen erheblich verbessern: Bei einer Änderung der Suchraumeinstellung von -3/+1 auf die optimalen Werte bei *nutrir* und bei *alimentar*, hätte man für *nutrir admiração* 43 statt 28 richtig extrahierte Fundstellen, bei *alimentar admiração* würde die Zahl dieser Kookkurrenzen von 3 auf 5 ansteigen. Für den Großteil der Verben würden sich die Precision-Ergebnisse bei einer allgemeinen Anhebung der Fenstergröße aber erheblich verschlechtern. Besonders davon betroffen sind die semantisch entleerten Verben der Funktionsverbgefüge.
Die häufigsten Funktionsverben im Portugiesischen sind: *estar, ter, pôr, fazer, dar, tomar*. Ihre Verwendung in Funktionsverbgefügen in Verbindung mit einem Substantiv, an Stelle eines einfachen Prädikats, legt ein enges Zusammenstehen mit dem Substantiv nahe (vgl. Kapitel 3.1.3). Funktionsverben kombinieren mit einer größeren Anzahl von Substantiven als die semantisch reicheren Verben, die als Kollokationspartner eine eingeschränktere Anzahl von Nomina zur Verfügung haben. Bei einer Vergrößerung des Fensters würden sich

die Funktionsverben daher oft auf andere Substantive beziehen, die sich in der Nähe des gesuchten befinden.[81]
Besonders für die Verben *estar* und *ter* ergibt sich zusätzlich das Problem, ihr Auftreten als Funktionsverb oder Vollverb von ihrer Verwendung als Hilfsverb zur Bildung der zusammengesetzten Zeiten, von Passivkonstruktionen oder Verbalperiphrasen zu unterscheiden. Unter einem Lemma werden in PECCI alle möglichen Funktionen vereint.[82] Die Extraktion von Fundstellen eines Substantivs mit Hilfsverben in die Kookkurrenzkonkordanzen wird in PECCI vermieden durch die permanente Reduktion der ursprünglichen Substantiv-Konkordanz um die Fundstellen mit den lemmatisierten Vollverben während eines Programmdurchlaufs (Modul 3). Da nur relativ wenige Verben lemmatisiert sind, bleiben viele Fundstellen von Auxiliaren mit nicht lemmatisierten Vollverben erhalten. Das Fehlen der linguistischen Annotation im Corpus und die damit nicht gegebene Möglichkeit, Fundstellen, in denen das Verb nur als Auxiliar fungiert, über die Anfrage zu ignorieren, ist nicht befriedigend zu kompensieren. Durch funktionsspezifische Anfragen mit Hilfe verschiedener POS-Tags ließe sich die falsche Extraktion vermeiden.

Konsequenzen aus dem verbalen Verhalten

Zu dem Aufwand, der sich mit einer verbindividuellen Fenstergröße verbindet, kommt hinzu, dass sich eine Einstellung am Verb auch nur für einige Teilnehmer des lexikalischen Feldes als optimal erweist. Für *nutri*r kann man hier folgende Substantive in der Reihenfolge ihrer Frequenz in einem Suchraum von ±12 anführen: *admiração, ódio, respeito, peixão, amor, esperança(s), entusiasmo, raiva.* Das Vorkommen mit den anderen Substantiven der Gefühle liegt bei 1 oder 2 Kookkurrenzen und es zeigt sich, dass in den meisten dieser Fälle eine falsche Kookkurrenz extrahiert wird. Bei Mel'čuk und Wanner wird 'hegen' als $Oper_1$ von den folgenden sechs Nomina genannt, die durch keinerlei semantische Generalisierung unter einer bestimmten Kategorie zusammenzufassen sind: "Achtung, Groll, Hass, Hoffnung, Leidenschaft, Zuneigung" (1994:138). Diese Kategorie könnte man durch die portugiesischen extrahierten Nomina verifizieren und den deutschen Substantiven noch zwei weitere hinzufügen: 'Bewunderung' für *admiração* und 'Begeisterung' für *entusiasmo*.
Die präferierten Kollokationspartner von *nutrir* hätte man auch schon durch einen Blick in die Rankingliste der Verben erfassen können (sowohl für den weiten wie für den kleinen Suchraum). Die obigen Beispiele haben aber gezeigt, dass ein großes Fenster mit guten Precision-Ergebnissen sicher auf eine Kollokation hinweist.[83] Dies bedeutet, dass gerade Substantive und Verben, die ein starkes Kollokationsverhalten zeigen, häufig weit auseinander stehen. Dies macht deutlich, dass sich für ein Verfahren, das mit angemessenen Recall- **und** Precision-Ergebnissen aufwarten will, eine syntaktische Annotation der Corpora als unerlässlich erweist. Da Corpora in der Form von Baumbanken, die nicht nur syntaktische Kategorien bestimmen, sondern auch deren grammatische Funktionen, noch keine geeignete Größe zur Kollokationsextraktion haben, wären gerade im Zusammenhang mit der Kollokationsspanne die Annotation der Corpora mit Ergebnissen aus Chunking-

81 *Ter admiração* kommt in *Cetempúblico* 120 mal in der kleinen Suchraumeinstellung -3/+1 vor. Eine deutsche äquivalente Übersetzung als Funktionsverbgefüge 'Bewunderung haben'* existiert nicht, man würde 'bewundern' oder 'Bewunderung hegen' sagen. Bei der Betrachtung der 271 satzinternen Vorkommen von *ter* außerhalb des Suchraums von *admiração* stellt man schnell fest, dass eine Vergrößerung des Suchraums vor allem zu falschen Positiven führen würde. Nur eine Vergrößerung des Fensters auf vier Wörter vor dem Substantiv, brächte noch richtige Kookkurrenzen: "*172156583: (...)* ***Tenho*** *cada vez mais* ***<admiração>*** *pelos agentes da polícia , ...* " (Cetempúblico) .

verfahren interessant. Die Effizienz einer flachen syntaktischen Analyse mit Hilfe von rekursivem Chunking, das Phrasenstrukturen zusammenfasst, wird in Heid (2005) vorgestellt. Die Substantive werden jetzt als Kopf einer NP extrahiert, die Einschränkung des Suchraums auf eine bestimmte Wortanzahl ist in diesem Verfahren nicht mehr relevant.
Hätte man ein syntaktisch annotiertes Corpus zur Verfügung, könnte man Kollokationskandidaten eventuell auch danach bestimmen, wie viele Wörter sich im Durchschnitt zwischen ihnen befinden. Bestimmte Verben, die eine starke kollokationale Bindung mit dem Substantiv aufweisen, tendieren dazu, weiter entfernt vom Substantiv zu stehen. Eine Konstruktion wie folgendes Beispiel, bei dem 12 Wörter zwischen Basis und Kollokat stehen, ist eher selten: "... , *o* ***<respeito>*** *que se lhe* ***tinha*** *era o mesmo que hoje em dia se pode* ***nutrir*** *, por exemplo , por um Sylvester Stallone : ...* " (Cetempúblico, 160304622). Doch kann man anhand der deutschen Übersetzung sehr schön zeigen, wie die von der Basis ausgelöste Assoziation den Leser über den ganzen Bereich hinweg bis zum Kollokat begleitet: ' ... die Achtung, die er vor ihm hatte, war die gleiche wie die, die man heute beispielsweise für Sylvester Stallone empfinden kann'.
Wie es scheint, steht nie ein Substantiv zwischen den Kollokationsbestandteilen, das ein ebenso starkes Kollokationsverhalten mit dem Verb aufweist, aber einer anderern grammatischen Relation angehört. Nur innerhalb derselben syntaktischen Konstruktion können mit einem Verb mehrere Substantive stehen, die mit diesem ein starkes Kollokationsverhalten zeigen, wenn es sich im Sinne des Britischen Kontextualismus um ein lexikalisches Set handelt: "150235739 ... , *pessoa por quem* ***nutro*** *há muito tempo e ininterruptamente a maior* ***estima*** *e um grande* ***<respeito>*** *humano e profissional . »* " (Cetempúblico) (' ... hegen ... Wertschätzung ... Achtung ...'). (Vgl. auch das Beispiel in Kapitel 2.3, bei dem 11 bzw. 16 Wörter zwischen Basis und Kollokaten stehen.)
Da sich eine Optimierung des Suchraums weder für die einzelnen Substantiv-Verb Kombinationen noch für das Verhalten einzelner Verben mit dem gesamten Wortfeld automatisieren lässt, muss man einen Default-Wert bestimmen. Dieser wurde hier sehr klein und für alle Verben gleich gewählt, um keine falschen Kandidaten aufzunehmen. Bei einem Vorkommen des Verbs bis zu drei Wörter vor dem Substantiv oder einem Wort dahinter kann man sich relativ sicher sein, dass es sich auch auf dieses bezieht. Auf Kosten dieser Sicherheit gehen natürlich viele Okkurrenzen außerhalb des Suchraums verloren. Bei einer kleinen Einstellung des Suchraums können Ergebnisse erzielt werden, bei denen die Precision 100% beträgt. Diese könnten dann zum Beispiel in Parsing-Verfahren als Disambiguierungshilfe eingesetzt werden. Der kleine Suchraum wurde auch unter dem Aspekt gewählt, dass dem Lexikografen eine Rankingliste mit schlechten Precision-Werten nicht unbedingt von Nutzen wäre. Er müsste sehr viel Zeit mit der Durchsicht falscher Ergebnisse verbringen. Bei einem kleinen Suchraum kann er auf die Richtigkeit der Ergebnisse vertrauen und bei Bedarf in die Dateien mit dem satzinternen nicht extrahierten Vorkommen schauen und daraus weitere Informationen über das Kollokationsverhalten erhalten. Auch im Bezug auf die Clusterverfahren und den von ihnen benötigten

82 Die *Linguateca* zählt 758.365 Vorkommen von *ter* als Hilfsverb und 491.441 Vorkommen als Vollverb (und Funktionsverb). In PECCI gibt es für *ter* nur eine Zahl (1.177.009), die beide Funktionen vereint. Aus diesem Grund sind die von PECCI extrahierten Frequenzdaten der Verben, die auch Hilfsverbfunktionen ausüben nicht ganz korrekt, denn die Vorkommen als Hilfsverb sollten in der Zählung eigentlich nicht enthalten sein.

83 Den gegenteiligen Schluss, dass es sich im Falle einer Verschlechterung der Precision-Ergebnisse um keine Kollokation handelt, kann man dagegen nicht ziehen. Vor allem die Verben, die Bestandteile von Funktionsverbgefügen sind, stehen immer nahe beim Nomen.

Kookkurrenzdaten scheint dieses Vorgehen die bessere Wahl zu sein. Die Ergebnisse - die Cluster mit Substantiven, die mit den gleichen Verben kollokieren - erscheinen durch eine Minderung in den Frequenzdaten weniger verfälscht als durch inkorrekte, zusätzlich extrahierte Verben.

5.3.3. Syntaktische Relationen der Kollokationen

Die mit PECCI gewonnenen Extraktionsdaten sind homogen in Bezug auf die Wortart der Kookkurrenzpartner, doch beinhalten sie verschiedene Möglichkeiten der syntaktischen Kombinierbarkeit der beiden. Die Substantiv-Verb Kookkurrenzen bestehen weitgehend aus Paaren, in denen das Gefühlssubstantiv im Akkusativ steht, als direktes Objekt eines transitiven Verbs. Die Substantiv-Verb Kookkurrenzen, bei denen zwischen Basis und Kollokat Präpositionen stehen, sind dazu im Vergleich eher selten (eine Vorstellung von der Verteilung erhält man in Kapitel 6.2, wo die Kollokationen der einzelnen Gefühlssubstantive verzeichnet sind). Auch die Kombinationen mit Präpositionalobjekt können eine Kollokation ("*32897587: Os ingleses devem estar* ***roídos de <inveja>*** *com a libra tão fraquinha ...*" (Cetempúblico) ('Die Engländer sind vor Neid sicher schon geplatzt, bei dem schwachen Pfund ...')) oder eine freie Wortverbindung sein ("*140004683: .... , dado que o Governo* ***gosta de <agitação>*** *:*" (Cetempúblico) ('... angenommen, dass der Regierung Aufregung gefällt:')). Im Portugiesischen kontrahieren einige Präpositionen mit dem nachfolgenden Artikel, im nächsten Beispiel wird *em+uma* zu *numa*: "*91407325: Disse o analista que os britânicos* ***mergulharam numa <tristeza>*** *« desproporcionada » , ...*" (Cetempúblico) ('Der Analytiker sagte, dass die Briten in einer unverhältnismäßigen Trauer versunken waren').

Die Substantive der Präpositionalgefüge können im Dativ oder Akkusativ stehen. Abhängig ist dies im Fall der Präpositionalobjekte von der Rektion der Verben. Die Verben legen auch die Präpositionen fest, die einleitende Bestandteile der Präpositionalobjekte sind. Bestimmte Verben und deren zugehörige Präpositionen verlangen den Akkusativ (beispielsweise 'warten auf' im Deutschen sowie das portugiesische Äquivalent *esperar por* (*'warten für')), häufiger steht das Präpositionalobjekt aber im Dativ, wie bei *gostar de* oder *roer-se de* in den Sätzen oben. Nur im Falle von Adverbialbestimmungen ist der Kasus von der Präposition abhängig.

Das indirekte Objekt der ditransitiven Verben wird im Portugiesischen immer von der Präpositionen *a* eingeleitet. Die Präposition übernimmt hier die Funktion der Kasusmarkierung, denn der Kasus ist im Portugiesischen am Substantiv oder einem Determinanten über die Morphologie nicht zu bestimmen: '*123650375: ... , que só a sua visão* ***infundia <respeito> ao inimigo*** *.* " (Cetempúblico) ('..., dass allein sein Anblick dem Feind Respekt einflößte').[84] Diese Verwendung der Präposition *a* ist vom Vorkommen als einleitender Bestandteil eines Präpositionalobjekts mit bestimmten Verben zu trennen, im zweiten Fall erscheint die Wahl der Präposition idiosynkratisch, und ist daher als Bestandteil des Präpositionalobjekts zu verzeichnen (wie in *responder a uma pergunta* 'auf eine Frage antworten').

84 Auch der Akkusativ wird in besonderen Fällen im Portugiesischen mit der Präposition *a* gebildet: wenn der Name Gottes oder das Pronomen *quem* als Akkusativobjekt steht, bei einigen Personalpronomina zur Hervorhebung des Akkusativobjekts oder wenn der Sinn des Satzes ohne Präposition nicht eindeutig wäre - im Falle einer Abweichung von der normalen Satzstellung (*Come o gato ao rato* ('Frisst die Katze die Maus')). Auch in diesen Fällen handelt es sich nicht um Präpositionalobjekte. Einen Überblick über die Verwendung der Präposition *a* gibt Hundertmark-Santos Martins (1982: 423-436).

Das Dativobjekt ditransitiver Verben ist als Basis in Substantiv-Verb Kollokationen nicht zu finden. Der 3. Aktant ist üblicherweise Rezipient und typischerweise ein belebtes Argument, seine genaue Identität ist für die Wahl der Kollokate offensichtlich nicht ausschlaggebend. Die Spezifizität der Kollokate ergibt sich in Abhängigkeit von der lexikalischen Realisierung des 1. oder 2. Aktanten. Das Auftreten des 3. Aktanten ist insofern relevant, als dass die Bedeutung des Kollokats mitunter abhängig ist von der Realisierung verschiedener Valenzrahmen der Verben und Substantive (vgl. Kapitel 3.3).
Für das Portugiesische ist die normale Satzstellung 'Subjekt + Verb + Objekt1 + Objekt2', doch können diese Positionen variieren. Der Suchraum wurde in PECCI so gewählt (das Verb kann sich bis zu drei Wörter links vom Substantiv befinden oder ein Wort rechts davon), dass er auch die normale Wortstellung der Subjektrealisierung der Gefühlssubstantive umfasst. In den Kollokationswörterbüchern und der Kollokationstheorie werden die Substantiv-Verb Kollokationen, in denen die Basis die Funktion des 1. Aktanten einnimmt (S+V), von der Objektrealisierung der Basis unterschieden (V+S). Kombinationen, in denen die Substantive der Gefühle Subjektfunktion im Satz übernehmen sind im Vergleich zur Realisierung als Objekt eher selten. Ein Beispiel zeigt folgender Satz: "*851287: Sua* ***<ira> desperta*** *as Eríneas , ...* " (Cetenfolha) ('Ihr Zorn weckt die Erinnyen').
Die Verben, die direkt hinter dem Nomen stehen, leiten oft auch einen verkürzten Nebensatz im Passiv ein. Das Verb steht dann in der Form des Partizips der Vergangenheit. In der entsprechenden Aktivkonstruktion übernimmt das Substantiv die Funktion des direkten Objekts: "*116042654: , explica Clara entre o* ***<alvoroço> provocado*** *por algumas cadelas com o cio que se haviam soltado , ...*" (Cetempúblico) ('... zwischen der Aufregung, die von einigen Hündinnen verursacht wurde ...'). Der Suchraum wurde auf ein Wort rechts vom Nomen eingeschränkt, da schon bei einem Vorkommen des Verbs zwei Wörter rechts vom Nomen das Verb meistens zu einer anderen syntaktischen Struktur gehört: "*... , suficiente para* ***gerar*** *um enorme* ***<alvoroço>*** *e* ***provocar*** *a* ***mobilização*** *de um apreciável aparato policial . "* (Cetempúblico, 117169303). Hier bezieht sich *provocar* auf *mobilização*, mit *alvoroço* ('Aufruhr') kollokiert *gerar* ('erzeugen').
Die Ergebnisse sind hinsichtlich der grammatischen Relationen der Kollokation manuell zu sichten. So stehen der Kollokation *ira desperta* mit dem Emotionsnomen als Subjekt und einer Fundstelle im Corpus, 10 Kookkurrenzen von *despertar ira* gegenüber. Eine generelle Trennung der Ergebnisse aufgrund der Wortstellung erweist sich nicht als adäquat, da die Umstellung der Normalform (Verb + Objekt, Subjekt + Verb) relativ oft erfolgt. Eine syntaktische Corpusaufbereitung wäre für präzisere Extraktionsergebnisse entscheidend.
Häufig fällt die syntaktische Struktur einer Substantiv-Verb Kombination auch einheitlich aus. In den 334 Vorkommen von *depositar esperanças em* ('Hoffnungen setzen auf') ist *esperanças* immer Akkusativobjekt, in den 11 Kookkurrenzen mit *diminuir* ('nachlassen')immer Subjekt. Für jede Substantiv-Verb Kombiantion existiert eine Exzerptionsdatei, die nur einen kleinen Ausschnitt um den Suchraum aus den Konkordanzen zeigt, um die Ergebnisse in übersichtlicher Form zu präsentieren. Zu bestimmten Umgebungsdaten werden die genauen Frequenzen aufgeführt, z.B. ob das Nomen direkt auf das Verb folgt oder ein Artikel dazwischen steht, ob eine Präposition dem Nomen folgt oder ihm vorangeht und ob bestimmte Konjunktionen typisch sind.

5.3.4. Weitere Evaluierungsmöglichkeiten

Zum Abschluss soll noch eine Evaluierung erfolgen, die die Anzahl der Sätze betrifft, die nicht extrahiert werden, da das Verb in der Lemmaliste fehlt. Dies geschieht anhand des kleinen Corpus *Cetenfolha*. Dort findet man für *admiração* 176 Konkordanzen. Davon werden 82 als extrahierte Kookkurrenzen mit einem der lemmatisierten Verben in dem kleinen Suchraum von -3/+1 aufgeführt. Weitere 23 syntaktisch korrekte Kookkurrenzen wären mit einer größeren Suchraumeinstellung zu extrahieren. 34 Sätze fehlen hingegen, weil *admiração* mit einem Verb vorkommt, das nicht in der Lemmaliste steht. In dem Rest der 37 Sätze kommt *admiração* überwiegend in Substantiv-Komposita wie *objeto de admiração* ('Objekt der Bewunderung') vor. Ähnliche Relationen erhält man auch bei einer manuellen Durchsicht der Ergebnisse anderer Substantive. Dabei fällt auf, dass die Sätze, die nicht extrahiert werden, weil die kookkurrierenden Verben in der Lemmaliste fehlen, selten Verben enthalten, die mit den Gefühlssubstantiven frequent sind. Es handelt sich fast immer um triviale Kombinationen von Verben mit Substantiven: "18187979: *Ela não* ***pensa*** *na* ***<admiração>*** *do viajante ...* " (Cetempúblico) ('Sie denkt nicht an die Verwunderung des Reisenden'). Der Großteil der möglichen Kollokate für das lexikalische Feld der Substantive der Gefühle scheint von den lemmatisierten Verben abgedeckt zu werden.
Trotzdem ist nicht daran zu rütteln, dass durch die Lemmatisierung von nur 226 Verben ca. 40% der Fundstellen verloren gehen - auch wenn es sich dabei überwiegend um freie Wortkombinationen handelt. Ebenso hat der Anspruch, nur richtige Extraktionsergebnisse zu liefern, durchaus seinen Preis. Im Durchschnitt werden ca. 25% der syntaktisch korrekten Fundstellen nicht extrahiert, weil die Verben außerhalb des Suchraums stehen. Auch scheint es möglich, dass Verben überhaupt nicht in der Rankingliste vorkommen, weil sie nur außerhalb des Suchraums mit dem Substantiv kollokieren (dieses Verhalten wurde bei mehrmaligem Vorkommen des Verbs mit dem Substantiv aber nicht beobachtet). Wiederum wird deutlich, wie wichtig ein linguistisch annotiertes Corpus ist, aus dem man über die POS-Informationen alle verbalen Kollokate eines Substantivs extrahieren kann. Werden im Corpus zusätzlich Phrasenstruktruen annotiert, entfällt die Beschränkung der Akquisition auf einem bestimmten Suchraum. Die Differenzierung der Substantiv-Verb Kombinationen nach der grammatischen Funktion des Nomens ist nur mit einem vollständig geparsten Text zu leisten, in dem auch diese Relationen verzeichnet sind. Die relativ freie Wortstellung im Portugiesischen eignet sich nicht für die Simulation syntaktischer Strukturen über die Reihenfolge der Wörter in der Anfrage.
Dass die Ergebnisse eines Programms, welches mit einem linguistisch nicht annotierten Corpus arbeitet, trotz der aufgezählten Schwierigkeiten durchaus lexikografisch zu verwerten sind, wird das nächste Kapitel zeigen. Dort wird auch noch einmal klar, warum die Fundstellen der Nomina nach deren Numerus separat behandelt werden und dass die beiden Formen mitunter ein deutlich unterschiedliches Kollokationsverhalten aufweisen. Der Vorteil des Perl-Programms, das mit einem Fenster um das Substantiv herum arbeitet, besteht in der Einstellungsmöglichkeit der Parameter, so dass Extraktionsergebnisse erzielt werden können, die eine sehr hohe Precision aufweisen. Im Gegensatz zu den aufwendigen Parsing-Verfahren werden hier wenige Fehler gemacht, die die Qualität der Ergebnisse betreffen. Der große Nachteil dieser Sicherheit sind die schlechten Recall-Werte. Die beste Lösung für diese Probleme wären manuell korrigierte Parsing-Ergebnisse in Form einer Baumbank, die eine geeignete Größe zur Kollokationsextraktion hat.

6. PECCIs Anwendung in der Lexikografie am Beispiel der Substantive der Gefühle

Durch die Corpuswahl ergibt sich eine klare diastratische Beschränkung der Ergebnisse auf die schriftliche Standardsprache. In diatopischer Hinsicht werden zwei Varietäten der portugiesischen Sprache, die Portugals und die Brasiliens, berücksichtigt, deren Verhältnis zueinander schon in Kapitel 4 Thema ist. Ein akkurater Vergleich der beiden Varietäten ist aufgrund der unterschiedlichen Corpusgrößen nicht möglich, die Corpusgröße von *Cetenfolha* beträgt nur 13,5% von *Cetempúblico.* In diesem Zusammenhang wird auch deutlich, dass ein Corpus in der Größe von *Cetenfolha* mit knapp 23.5 Millionen Wörtern nicht ausreicht, um einen Überblick über potentielle Kollokationskandidaten zu geben. In *Cetempúblico* findet man für *admiração* ('Bewunderung') bei einer Signifikanzgrenze von 1 (d.h. Kookkurrenzen werden auch bei einmaligem Vorkommen in die Rankinglisten aufgenommen) immerhin 63 verschiedene Verben, die innerhalb der Kollokationsspanne vorkommen, bei einer Signifikanzgrenze von 2 stehen noch 43 Kandidaten zur Verfügung. *Cetenfolha* liefert bei einer Signifikanzgrenze von 1 nur 29 Verben, müssen diese mindestens 2 mal erscheinen, bleiben nur noch 11 Kandidaten übrig.
Aus diesem Grund wird die Darstellung der Kollokationen zunächst regional beschränkt. Als Basis der lexikografischen Darstellung dienen zunächst nur die Extraktionsergebnisse aus *Cetempúblico.* Dass die Differenzen in der Standardsprache Portugals und Brasiliens auch die Wahl der Kollokate betreffen, wird im letzten Teil dieses Kapitels dargestellt. Dort werden auffällige Abweichungen im Kollokationsverhalten zwischen den beiden Varietäten explizit verzeichnet. Weitere Varietäten werden nicht berücksichtigt. Dies hat zum einen den Grund, dass in der *Linguateca* aus den portugiesischsprachigen Ländern Afrikas keine Corpora vorhanden sind. Hingegen haben die transkribierten Corpora gesprochener Sprache oder belletristische Corpora keine geeignete Größe zur Kollokationsextraktion.
Die Signifikanzgrenze wird hier in der Auswertung mit 1 festgelegt, da im Hausmannschen Sinne eine Kollokation nicht unbedingt frequent sein muss und da sich in portugiesischen Wörterbüchern genannte Kollokationen mitunter nur einmal im Corpus befinden. Die Assoziationsmaße werden als Grundlage gewählt um dem Lexikografen eine statistisch vorsortierte Liste zu bieten, sie sollen nicht einer Kollokationsidentifikation dienen. Ihre Werte geben im Vergleich eher Aufschluss über das Verhalten einer Kookkurrenz als eine einfache Rankingliste. Darum gibt es in PECCI die Möglichkeit, die Werte von zwei verschiedenen Assoziationsmaßen anzuzeigen. Bekannt ist, dass der t-Test die frequenten Kookkurrenzen am höchsten bewertet. Da diese in der lexikografischen Arbeit meist auch einen relevanten Stellenwert einnehmen, wird der t-Test in der Default-Einstellung der Assoziationsmaße als erstes statistisches Maß gewählt. Nach dem ersten Assoziationsmaß wird die Rankingliste sortiert. Bei der Wahl der Mutual Information als Default-Einstellung für den zweiten statistischen Test, wird die Korrelation seltener Paare im Corpus hoch bewertet.
Wie schon in Kapitel 2.1 erwähnt, wird die gleichzeitige Darstellung der Werte der vier implementierten statistischen Tests als unübersichtlich empfunden. Die Werte von log-likelihood und χ^2 bewegen sich im Rankingverhalten fast immer zwischen den Werten von t-score und Mutual Information, weshalb sich bei der Auswertung der beiden Extreme die Information von log-likelihood und χ^2 in diesem Zusammenhang als redundant erweist.

Interessiert man sich aber speziell für die Ergebnisse des Likelihood-Ratio-Tests oder des χ^2 -Tests, kann man über die Benutzerschnittstelle auch sie als statistischen Test wählen.
Die Effizienz für den Lexikografen aus der Anzeige der Werte der beiden statistischen Tests soll folgende Analyse zeigen. Im *PONS Standardwörterbuch* (2002) wird unter dem Eintrag von *infundir* ('einflößen') nur ein Beispiel für den Gebrauch gegeben: *infundir respeito* ('Respekt einflößen'). Unter dem Eintrag von *respeito* fehlt die Angabe jeglichen kollokationalen Verhaltens mit einem Verb. In keinem der anderen Wörterbücher wird diese Kollokation genannt. Betrachtet man nun die nach dem t-score sortierte Rankingliste entdeckt man *infundir* erst auf Platz 17. Dies liegt daran, dass das Vorkommen der Kombination mit *respeito* relativ selten ist - *infundir respeito* hat nur 8 Okkurrenzen (der Anführer der Rankingliste *merecer respeito* ('Respekt verdienen') bringt es auf 166 Okkurrenzen). Bei der Mutual Information erhält hingegen die Kombination *infundir respeito* den höchsten Wert und stünde hier an Platz 1.[85] Das seltene Gesamtvorkommen im Corpus von *infundir* mit nur 47 Okkurrenzen wird bei der Mutual Information stärker bewertet als beim t-Test (*merecer* kommt auf 16.937 Okkurrenzen im Gesamtcorpus). Mit einem Blick auf die Werte des zweiten Assoziationsmaßes würde der Lexikograf *infundir respeito* schnell als relevanten Kollokationskandidaten identifizieren und somit den Wörterbucheintrag verifizieren können.[86]

6.1. Aspekte der Kollokationsbeziehungen und computationelle lexikografische Darstellung von Kollokationen

6.1.1. Interne Kollokationsrelationen

Die Kollokationstheorie eröffnet ein sehr viel weiteres Spektrum an aufzunehmenden externen Daten und internen Kollokationsbeziehungen, als die Darstellungsform von Kollokationen in den Print-Wörterbüchern vermuten ließe. Typische Einträge von Kollokationen in Print-Wörterbüchern wurden in Kapitel 3.2 vorgestellt. Je nach Güte und Ausführlichkeit der Nachschlagewerke erscheinen die Kollokationen mehr oder weniger strukturiert. Werden die Kollokationen von den weiteren Angaben im Eintrag des Lemmas getrennt dargestellt, ist die Unterteilung der Kollokationen nach den an ihnen beteiligten Wortarten inzwischen Usus, und auch die Gruppierung nach (quasi)synonymen Kollokaten geht häufig in die Kollokationsbeschreibung mit ein. Dazu im Gegensatz sind selten explizite Informationen zur Valenz (der Gesamtstruktur), morphosyntaktische Präferenzen, diasystematische Marker oder Frequenzangaben verzeichnet. Die aufgezählten Spezifikationen sind typische Vertreter externer Kollokationsinformationen, die für eine präzise Beschreibung unabdingbar sind.
Interne Kollokationsrelationen sind im Besonderen das Thema von Kapitel 2.4.1 und Kapitel 3.2.1. Die bekannteste Bezeichnung der internen Kollokationsbeziehung ist das Konzept von Basis und Kollokator. Darüber hinaus sind nur wenige Versuche bekannt, diese Beziehung weitergehend zu differenzieren, d.h. die kollokationale Relation, die zwischen zwei Wörtern besteht, systematisch zu präzisieren. Die Kollokationssystematik (Abb. 6) zeigte die unterschiedliche Kollokationsterminologie verschiedener Autoren und deren Definitionskriterien im Vergleich. Die Kriterien dienen der Abgrenzung der Kollokationen

85 Beim Likelihood-Ratio-Test liegt *infundir respeito* auf Platz 12, beim χ^2 -Test auf Platz 2.
86 Wie in Kapitel 3.1 beschrieben und erläutert, sind die Werte der Mutual Information für sehr niedrig frequente Kookkurrenzen (ca. 1-4 Fundstellen) wegen deren Überbewertung mit Vorsicht zu verwenden.

gegenüber den Idiomen und den freien Wortkombinationen, sowie der Analyse der Kollokationsrelationen.
Die Semantik des Kollokats ist ausschlaggebend für die Zuordnung zu den opaken Kollokationen oder Teilidiomen. In diesem Fall weicht die Bedeutung des Kollokats stark von seiner primären Bedeutung ab, zudem ist die Verbindung mit der Basis sehr restriktiv. Die Bedeutung des Kollokats ist daher aus dem Vorkommen mit weiteren Wörtern nicht zu erschließen, und die Gesamtbedeutung der Kollokation ohne Kenntnis der spezifischen Bedeutung des Kollokats nicht zu ermitteln. Als weitere Unterscheidungskriterien werden Idionsynkrasie (inter- und intralingual), Üblichkeit und semantische Motiviertheit genannt. Doch entbehrt die bloße Aufzählung dieser Kriterien bei verschiedenen Autoren eigentlich jeglicher Systematik. Neben dem Aspekt der grammatischen Verbundenheit der Kollokationen, der sich in der Kollokationsstruktur zeigt, stellt sich daher die Frage, wie die Beziehung, die zwischen den beteiligten Wörtern besteht, zu beschreiben ist, denn auch innerhalb einer Kollokationsstruktur liegen Kombinationen mit ganz unterschiedlichen internen Relationen vor.
Eine formal bestechende Version der Systematisierung der internen Kollokationsbeziehung gibt Mel'čuk (1998). Ohne auf den ausführlichen phraseologischen Rahmen einzugehen, in dem die Kollokationen bei Mel'čuk stehen[87], mögen hier nur die Kollokationsdefinition und die weitere Differenzierung wiedergegeben werden:

> A COLLOCATION **AB** of language **L** is a semantic phraseme of **L** such that its signified 'X' is constructed out of the signified of one of its two constituent lexemes - say, of **A** - and a signified 'C' ['X' = 'A ⊕ C'] such that the lexeme **B** expresses 'C' only contingent on **A**.
>
> (Mel'čuk 1998: 30)

Die Kollokationen werden in vier Typen unterteilt:

> 1. EITHER 'C' ≠ 'B', i.e. **B** does not have (in the dictionary) the corresponding signified;
> AND [(a) 'C' is empty, that is, the lexeme **B** is, so to speak, a semi-auxiliary selected by **A** to support it in a particular syntactic configuration;
> OR (b) 'C' is not empty but the lexeme **B** expresses 'C' only in combination with **A** (or with a few other similar lexemes)];
> 2. OR 'C' = 'B', i.e. **B** has (in the dictionary) the corresponding signified;
> AND [(a) 'B' cannot be expressed with **A** by any otherwise possible synonym of **B**;
> OR (b) 'B' includes (an important part of) the signified 'A', that is, it is utterly specific, and thus **B** is 'bound' by **A**].
>
> (Mel'čuk 1998: 30-31)

Zu 1(a) zählen die FVG, 1(b) besteht aus opaken Kollokationen wie *schwarzer Kaffee*, 2(a) umfasst Kollokationen mit "Intensivieren" wie *starker Raucher* oder *tief bewegt* und 2(b) bilden Kombinationen wie *das Pferd wiehert* oder *ranzige Butter*. Im Falle der Kollokationen von 1(a), 2(a) und 2(b) ist die Bedeutung des Kollokats transparent - auch wenn die Bedeutung der Verben in den FVG von der Normalbedeutung der Verben abweicht, bereiten diese aufgrund ihres großen Kollokationsradius keine Dekodierungsprobleme. Die Wahl der

87 Vgl. Mel'čuk (1998): 25-31. Für die Abgrenzung der Kollokationen von anderen Wortkombinationen ist bei Mel'čuk die Saussuresche Unterscheidung von Signifikat (signified) und Signifikant (signifier) ausschlaggebend, für die Kollokationstypen spielt diese Dichotomie jedoch keine Rolle, daher wird im Weiteren im Deutschen allgemein von Bedeutung gesprochen.

Kollokate geschieht in 1(a), 1(b) und 2(a) idiosynkratisch, was sie von den Kollokationen in 2(b) unterscheidet. Die Kombinationen in 2(b) werden üblicher-weise als lexikalische Solidaritäten bezeichnet, und genau genommen ist diese Gruppe von Kombinationen nicht mit der Kollokationsdefinition von Mel'čuk zu vereinen, denn die aufgezählten Kollokate haben außer in den Kollokationen keine weiteren Bedeutungen, zudem geschieht ihre Auswahl nach semantisch vorgegebenen Kriterien.

Die Typisierung der Kollokationen bei Mel'čuk wird hier erwähnt, um zu verdeutlichen, dass spezifische interne Kollokationsrelationen bestehen und diese durchaus systematisch zu erfassen sind, auch wenn bei Mel'čuk wiederum verschiedene Aspekte fehlen, wie ein Vergleich mit der Kollokationssystematik in Abbildung 6 zeigt, und auch sicherlich zwischen den vier Typen mitunter fließende Übergänge auszumachen sind. Die Typisierung der Kollokationen findet sich in den lexikalischen Funktionen wieder. Für das *Keyword*, hier das Substantiv, wird für eine bestimmte lexikalische Funktion ein Feld lexikalischer Ausdrücke bestimmt, dessen Werte in Abhängigkeit vom *Keyword* eine spezifische Bedeutung ausdrücken (vgl. Kapitel 2.4.1). Die semantisch entleerten Funktionsverben, gehören je nach ihren tiefensyntaktischen Beziehungen, in denen sie zum Substantiv stehen, zu den lexikalischen Funktionen 'Oper', 'Func' oder 'Labor', die semantisch reicheren Verben zu den lexikalischen Funktionen 'Real', 'Fact' oder 'Labreal'. Für sehr spezifische und opake Kollokationen stehen Nicht-Standard Funktionen bereit. In einer separaten Zone gespeichert werden die Idiome. Die lexikalischen Solidaritäten fließen trotz ihrer semantischen Vorhersagbarkeit nahtlos in dieses Konzept mit ein.

Semantische Motiviertheit stellt bei Mel'čuk generell kein Ausschlusskriterium für die Aufnahme in das Lexikon unter den syntagmatischen lexikalischen Funktionen dar. Wie schon in Kapitel 1 erwähnt, kommt Mel'čuk in der Praxis zu einem sehr weiten Kollokationskonzept. Auch diejenigen Kombinationen, die semantisch transparent sind, die einen geringen Fixiertheitsgrad aufweisen und, in denen die Kookkurrenz vorhersehbar ist, gehören durch Analogie mit restringierten Fällen zu den Phrasemen (Mel'čuk 1998: 42). Die Prämisse der idiosynkratischen Wahl des Kollokats tritt in den Hintergrund, da die lexikalischen Funktionen als Interlingua dienen, aus der die abstrakten Konzepte einzelsprachlich zu verwirklichen sind. Freie Kombinationen, Einwortlexeme oder nicht kollokationale Konstruktionen der einen Sprache sind Kollokationen in einer weiteren Sprache und daher alle als Versprachlichungsmethoden eines bestimmten Konzepts zu speichern.

Eine Äquivalenz findet die Theorie der lexikalischen Funktionen im Kollokationskonzept von Hausmann und der Aufteilung der Kollokationen in Basis und Kollokator. Hausmann jedoch betont gerade die idiosynkratische Wahl des Kollokats (vgl. Kapitel 3.1.3). In späteren Artikeln (1997, 2005) nimmt Hausmann eine Zweiteilung der lexikalisch offenen Klassen des Wortschatzes in Auto- und Synsemantika vor. Die entsprechende Definition wurde schon in Kapitel 3.1.3 wiedergegeben und in Kapitel 3.2.1 teilweise kritisiert. Genauere Ausführungen Hausmanns zur Theorie der Autosemantika und Synsemantika sind nicht bekannt.

Hier soll nur ein weiteres Beispiel angeführt werden, das zeigt, dass die Determination polysemer Wörter durchaus in der Richtung vom Kollokat zur Basis verlaufen kann. Betrachtet man die Kollokation *mit der Maus klicken* bzw. *auf die Maus klicken*, geschieht die Wahl des Kollokats idiosynkratisch, *klicken* wird von der *Maus* selegiert. Die Kombination *auf die Maus tippen* wäre zwar von der Semantik des Verbes durchaus denkbar,

ist jedoch im Sprachgebrauch nicht üblich. *Klicken* trägt aber trotz der kollokationalen Abhängigkeit von der Basis zur Disambiguierung des polysemen Substantivs *Maus* bei. Durch das Verb wird deutlich, dass es sich hier nicht um ein Tier, sondern nur um ein Peripheriegerät des Computers handeln kann. Auch dieses Beispiel zeigt, dass das Konzept der Semiotaxis, wie von Hausmann dargestellt, so nicht bestehen kann, denn auch viele Nomina sind nicht autonom, sondern wie die Verben (auch) über ihren Kontext zu beschreiben.
Problematisch am Konzept der lexikalischen Funktionen und den zugehörigen Wörterbucheinträgen ist die Einseitigkeit der Beziehungsrichtung, denn sie setzt eine disambiguierte Bedeutung des Lemmas schon voraus. Präzise zu unterscheiden sind zwei Arten von Relationen, die zwischen den Lexemen im Satz bestehen, die bidirektionalen Beziehungen sind nicht durch ein einfaches regelhaftes Schema zu beschreiben. In sprachverarbeitenden Systemen ist die Analyse- und Generierungsseite explizit zu trennen. Die Kollokate können wie die Basen bei der Sprachanalyse zur Disambiguierung polysemer Wörtern dienen, für die kollokationale Wahl bei der Generierung ist die Basis entscheidend.
An anderer Stelle präzisiert Hausmann den scheinbaren Widerspruch der Abhängigkeit des Verbs vom Substantiv innerhalb der Kollokation und der Determination der Satzstruktur durch die Valenz des Verbs. Hausmann findet unter dem Stichpunkt "Collocation et valence" eine evidente Erklärung für die vermeintliche Kontradiktion. Handelt es sich um Kollokationen, gibt es in einem Satz nicht nur das syntaktische Modell der Valenz, sondern ebenso phraseologische Zwänge. Der Sprecher wählt zunächst das Substantiv und in Abhängigkeit davon das Verb mit seinen spezifischen Valenzeigenschaften: "Il n'y a par conséquent aucune contradiction entre le fait, d'une part, que dans la collocation le verbe soit sélectionné **après** le substantif et **en fonction du** substantif et le fait, d'autre part, qu'une fois sélectionné, ce soit le verbe que impose les modalités syntaxiques que le relient au substantif" (Hausmann 2005: 3).
Der Ansatz Hausmanns weist auf eine bisher in der Kollokationstheorie nicht weiter behandelte Diskrepanz hin und kann in verschiedenen Richtungen zu weiteren Überlegungen führen. Zum einen ist es nicht immer nur das Verb, das für die Satzstruktur ausschlaggebend ist, auch das Substantiv und weitere Wortarten erfordern mitunter Aktanten. Kurz angedeutet wurde das Problem bei der Beschreibung der FVG im Paradigma der Kollokationen in Kapitel 3.1.3. In diesen Wortkombinationen werden die Valenz- und Rektionseigenschaften durch die Substantive bestimmt, Selektionsrestriktionen, die beide Mitspieler betreffen, gehen vom Nomen aus, und von vielen Autoren wird das Nomen als Prädikatskern nicht als Aktant aufgefasst, sondern ihm wird selbst eine prädikative Funktion zugeschrieben. Aber auch in den Kollokationen, in denen die Verben ihre Semantik weitgehend behalten oder eine andere semantisch ausgeprägte Bedeutung erhalten, kann die Valenz des Substantivs eine Rolle spielen. Dies wurde in Kapitel 3.3 bei der Übersetzungsproblematik, die mit den Kollokationen einhergeht, gezeigt, gleichzeitig wurde aber auch verdeutlicht, dass in den Lexikoneinträgen der vorliegenden Arbeit eine einheitliche Darstellung in der Nomenklatur gewahrt werden soll und aus diesem Grund auch die Nomina der FVG ihren Aktantenstatus behalten.[88]

88 Eine Zusammenfassung von Lösungsvorschlägen in der Computerlinguistik für den Umgang mit FVG wird u.a. von Böhmer (1994: 326-404) mit Bezug auf Grammatiktheorien und die Maschinelle Übersetzung gegeben. Im Rahmen des FrameNet-Projekts (framenet.icsi.berkeley.edu/) wird die lexikografische Darstellung der FVG ausführlich behandelt, die semantischen Rollen der Komplemente werden hier nicht vom Funktionsverb induziert, sondern bezogen auf das prädikative Nomen kodiert.

Eine weitere Frage, die weder Hausmann noch Mel'čuk aufwerfen oder beantworten, betrifft die Wahl des Kollokats durch das Substantiv im Satz. Es ist nämlich keineswegs von vornherein klar, welches Substantiv im Satz für die Selektion des Kollokats verantwortlich ist. In dem Satz *die Maus frisst Käse*, selegiert das Subjekt das Verb, im Sinne Coserius handelt es sich hier um den Fall einer Affinität im Paradigma der lexikalischen Solidaritäten (vgl. 3.1). In dem Satz *die Maus schlürft Austern* ist es hingegen das Objekt, das für die Wahl der Prädikats ausschlaggebend ist, ob es sich bei der Maus um ein Tier, beispielsweise in der Darstellung eines Comics handelt, oder um ein humanes Subjekt, für das *Maus* als Koseform bzw. abwertende Bezeichnung gebraucht wird, ist ohne den weiteren Kontext nicht zu bestimmen. *Austern schlürfen* wird als usuelle Verbindung im Gegensatz zu *die Maus frisst* zu den Kollokationen zählen, von Coseriu werden sie beide lexikalische Solidaritäten genannt.
Ohne dass die vorliegende Arbeit selbst eine befriedigende Antwort auf die Probleme und Mehrdeutigkeiten bieten könnte, sollen diese Beispiele verdeutlichen, wie viel Arbeit auf dem Gebiet der Kollokationen noch zu leisten ist. Neben den externen Kollokationsinformationen sind auch Angaben über interne Kollokationsrelationen im Wörterbuchartikel wünschenswert. Eine präzise Beschreibung der Kollokation in allen Bereichen stellt vor allem für die Computerlinguistik eine unabdingbare Prämisse dar, denn im Gegensatz zum Menschen ist für ein maschinelles sprachverarbeitendes System die Dekodierung des komplexen Zusammenhangs, der zwischen den Lexemen besteht, anhand des Sprachgefühls nicht zu leisten.

6.1.2. Kollokationen in Online-Wörterbüchern

Im Weiteren soll zunächst der Umgang mit Kollokationen in elektronischen Wörterbüchern, die über das Internet öffentlich zugänglich sind, kurz beschrieben werden, um darüber die Darstellungsform der Kollokationen in der vorliegenden Arbeit zu motivieren. Ziel der Arbeit ist eine möglichst exakte Beschreibung der Kollokationen, die sowohl über externe Kollokationsinformationen wie auch über interne Kollokationsrelationen Auskunft gibt. Im Gegensatz zu den Print-Medien spielt der verfügbare Platz in den WorldWideWeb-Ressourcen nur eine untergeordnete Rolle. Zu erwarten wären gerade hier ausführliche Informationen zu den Kollokationen, doch wird diese Erwartung nur in einigen Fällen bestätigt. Gemeinsam ist den elektronischen Wörterbüchern mit institutionellem Hintergrund ihre Entstehung auf der Basis umfangreicher Corpora.
Das *Digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts*[89] von der Berlin-Brandenburgischen Akademie der Wissenschaften stellt für jedes Lemma den Eintrag aus dem *Wörterbuch der deutschen Gegenwartssprache*, Textbeispiele aus dem *DWDS* Kerncorpus, automatisch berechnete semantische Relationen und ein viertes Feld mit automatisch berechneten Kollokationen bereit. Die häufigsten Kollokationen werden unabhängig von ihrer Wortart in der Form eines gerichteten Graphen gezeigt, desweiteren kann man über einen komfortablen Benutzerdialog die "Kollokationsstatistik" für ein Lemma zusammenstellen. Die zu durchsuchenden Corpora, die Fenstergröße, das statistische Maß, das Ausgabeformat und einige weitere Parameter sind vom Benutzer einstellbar. Die Kollokationen werden in Form von Rankinglisten angezeigt. Abgesehen von "Stoppwörtern" gelten alle Wortarten als Kollokate auch die einzelnen Konjugationsformen der Verben werden als unterschiedliche Kollokate verarbeitet. So befindet sich beispielsweise *Freude macht* in der

89 http://www.dwds.de/

Rankingliste, die auf log-likelihood basiert, auf Platz 5, während man *Freude gemacht* auf Platz 13 findet. Als Basis der Kollokation kann jedes beliebige Wort dienen.
Am Institut für Deutsche Sprache gibt es in der Abteilung Lexik das laufende Projekt *elexiko*[90], das ein "lexikalisch-lexikologisches korpusbasiertes Informationssystem des IDS" werden soll. Begonnen wurde es 2004 und der Demonstrationswortschatz beläuft sich auf 243 Lemmata. Lesartenübergreifende Angaben bestehen aus Informationen zu Ortografie, Wortbildung, nationaler Verteilung sowie Herkunft und Wandel. Die Kollokationen sind unter den lesartenbezogenen Angaben enthalten, die außerdem Bedeutungserläuterungen, sinnverwandte Wörter, Grammatik und Besonderheiten des Gebrauchs umfassen. Zu finden sind die Kollokationen unter dem Stichwort "semantische Umgebung und lexikalische Mitspieler". Hier sind die Kollokationen neben anderen häufigen Wörtern zu finden, die in einer bestimmten grammatischen Relation zum Lemma stehen. Dient als Lemma *Frage* erscheinen unter "wie ist eine Frage?" die Kollokate *bang, berechtigt, bohrend,* etc. in alphabetischer Reihenfolge, unter "wer stellt eine Frage?" die Nomina *Journalisten* und *Richter*. Kollokationen werden daher wie andere Wörter behandelt, die zwar in einer bestimmten grammatischen Relation zum Lemma stehen *(Journalist* = Subjekt*, Frage* = Objekt), deren grammatische Relation aber keiner Kollokationsstruktur entspricht. Auch genauere Kollokationsinformationen werden hier nicht aufgeführt. Unter dem Stichwort "Typische Verwendungen" sind teilweise die gleichen Kollokationen zu finden, diesmal mit typischen Präpositionen und Artikeln, sowie Tripelkollokationen wie *eine Frage in den Raum werfen*, daneben weitere Kollokationen, die im Schema der Erfragbarkeit der semantischen Umgebung keinen Platz finden, wie *eine Frage des Geldes*, und auch wieder Wortkombinationen, die nicht in das übliche Schema der Kollokationsstrukturen passen, wie *eine Diskussion über Fragen der/des.*
Für das Französische gibt es das elektronische Wörterbuch *Le Trésor de la Langue Française informatisé*[91], das im Aufbau einem typischen Print-Wörterbuch entspricht. Doch bietet sich hier die Möglichkeit, die Kollokationen, die über den Wörterbuchartikel verteilt erscheinen, unter dem Stichwort "Syntagme" farblich hervorzuheben. Für das Französische steht auch ein spezielles elektronisches Kollokationswörterbuch zur Verfügung, das *Dictionnaire des Collocations*[92]. Als Lemmata sind Substantive, Verben und Adjektive vorgesehen, doch dient hier als Grundlage ein nicht näher erläutertes Kollokationskonzept, das als Kollokate der Adjektive geläufige Substantive und als Kollokate der Verben verschiedene Komplemente bietet. Allein unter den Substantiven sind Kollokationen im üblichen Sinne zu finden, hier werden Adjektive und Verben alphabetisch sortiert ohne weitere Informationen dargeboten.
Auf dem Wortschatz-Portal der Universität Leipzig[93] gibt es die Möglichkeit monolinguale corpusbasierte Wörterbücher in 17 Sprachen zu konsultieren. Der Eintrag des Lemmas enthält zum einen die üblichen lexikografischen Daten zum gesuchten Wort, darüberhinaus vor allem Kookkurrenzangaben verschiedener Art. Unter dem Abschnitt "Teilwort von" werden Kollokationen mit vielfältigen Strukturtypen gezeigt, hier sind *Prinzip Hoffnung, neue Hoffnung, keine Hoffnung haben, eine Hoffnung zu Grabe tragen, Hoffnung einflößen, Wechsel von Furcht zu Hoffnung* und viele weitere Wortkombinationen in scheinbar beliebiger Reihenfolge vertreten. Weiter unten werden satzinterne signifikante Kookkur-

90 http://www.elexiko.de/
91 http://atilf.atilf.fr/tlf.htm
92 http://www.tonitraduction.net/
93 http://wortschatz.uni-leipzig.de/

renzen mit Frequenzangaben genannt, und in zwei weiteren Abschnitten die Wörter, die jeweils direkt rechts bzw. links vom Lemma stehen. In diesen Abschnitten wird jede Wortform einzeln aufgezählt, man kann dem Eintrag entnehmen, dass *schöpften* 34 mal und *schöpfte* 11 mal als direkter Nachbar rechts von *Hoffnung* steht. Begründet wird dieses Vorgehen von Projektleiter Quasthoff damit, dass die Sammlung von Vollformen relativ einfach ist, und die Reduktion auf die Grundformen eine zusätzliche Fehlerquelle birgt. Auch sind durch diese Methode Aussagen über das Nichtvorkommen bestimmter flektierter Formen möglich. Als Nachteil sieht auch Quasthoff eine gewisse Redundanz in der Materialsammlung (1998: 94). Hierzu sei zum einen angemerkt, dass die Extraktion von Kookkurrenzlisten ungeachtet der Wortart der Kollokate sowohl für den Lexikografen als auch für den Benutzer eines Wörterbuchs erhebliche Nachteile bei der Konsultation desselben mit sich bringt, interessiert man sich für das Vorkommen des Lemmas mit einer spezifischen Wortart. Zudem macht das Aufführen der einzelnen Wortformen einen Überblick über die tatsächliche Kookkurrenz des Lemmas mit dem betreffenden Kollokat nahezu unmöglich, auch statistisch aussagekräftige Ergebnisse, die ein bestimmtes Kollokat betreffen, lassen sich auf diese Art nicht erzielen. Zum anderen kann man natürlich auch aus einem lemmatisierten Text die einzelnen Wortformen bei Bedarf einzeln extrahieren oder eine Kookkurrenzliste für alle Kollokate erstellen. Grundsätzlich ist für lexikografische Zwecke die Extraktion der Kollokate sowohl sortiert nach ihrer Wortart als auch reduziert auf ihre Stammform der Extraktion aller benachbarter Wortformen eines Lemmas vorzuziehen, vorausgesetzt dass man einen Corpus mit Lemmaangaben und POS-Tags zur Verfügung hat.

Auf demselben Portal kann man ein Deutsch-Englisches Wörterbuch konsultieren, in dem zahlreiche Kollokationen des Lemmas ihre äquivalente Übersetzung im Englischen finden. In die Darstellung (der Stammformen der Kollokate) fließt bei den Substantiv-Verb Kollokationen die Verwendung des Artikels mit ein, außerdem werden die Frequenzangaben für die einzelnen Wörter angezeigt sowie die Anzahl der Belege der Übersetzung im englischen Corpus. Da die Datenerfassung durch Mitarbeiter aus den verschiedensten Fachrichtungen auf freiwilliger Basis geschieht und die Qualitätskontrolle ebenfalls dezentral organisiert wird, indem bereits erfasstes Material zur Diskussion gestellt wird und Änderungen vorgeschlagen werden können, sind die Einträge mit einer gewissen Vorsicht zu genießen. So ist z.B. die Übersetzung von *voll Freude* mit *beaming* nicht ganz einsichtig. Ein Vorteil dieses Vorgehens ist die ungeheuer große Datenmenge, die das Wortschatz-Portal zur Verfügung stellt.

Im bilingualen Bereich existieren weitere Online-Wörterbuch für die Sprachrichtungen deutsch-englisch, deutsch-französisch und deutsch-spanisch der LEO GmbH[94], die wie das *Dictionnaire des Collocations* auf eine Privatinitiative zurückzuführen sind. Durch das Einbinden externer Ressourcen können beispielsweise Informationen zur Wortdefinition, Aussprache, Flexion oder Etymologie zentral abgefragt werden. Der Beitrag der LEO GmbH am Wörterbuchartikel beschränkt sich im Wesentlichen auf die Wiedergabe der möglichen Übersetzungsäquivalente der Lemmata und der Kollokationen, die mit ihnen gebildet werden. Wie die Anfragestatistik zeigt, liegen an einem Wochentag fast 8 Millionen Anfragen vor, woraus sich schließen lässt, dass ein Bedarf an Online-Wörterbüchern tatsächlich vorhanden ist. Die Kollokate der Substantive sind bei LEO in drei Bereiche gegliedert. Unter der Rubrik "Verben und Verbzusammensetzungen" findet man die verbalen

94 http://dict.leo.org/

Kollokate in beliebiger Reihenfolge - der Artikelgebrauch, die Präpositionen und obligatorische Aktanten sind weitgehend verzeichnet. Die Rubriken "Wendungen und Ausdrücke" und "Zusammengesetzte Einträge" führen die Kollokationen mit Adjektiven, weiteren Substantiven, sowie typische Präpositionen für das Lemma auf.
Das ELDIT-Programm[95] ist eine Online-Plattform für Sprachenlerner des Deutschen oder Italienischen und deckt einen Grundwortschatz von ca. 3000 Wörtern ab. Neben dem elektronischen Lernerwörterbuch Deutsch-Italienisch/Italienisch-Deutsch bietet es Texte zu verschiedenen Themenbereichen, Kurzgrammatiken und Übungen zu beiden Sprachen. Anhand des Programms, soll es möglich sein, sich gezielt auf die Zweisprachigkeitsprüfung in Südtirol vorzubereiten. Auf einem sehr übersichtlichen Interface sind die detaillierten lexikografischen Angaben zu finden. Unter dem Stichwort "combinazioni" bzw. "Verwendung" werden bei den Substantiven die verbalen Kollokate aufgeführt. Präferenzen im Numerusgebrauch, die Artikelwahl, notwendige Aktanten und übliche Modifikatoren werden in Form einer fortlaufenden Phrase aufgeführt (*jemdn. erfüllt bitterer, blinder, tiefer ... Hass (gegen/auf jemdn.)*). In der gesuchten Sprachrichtung wird die Verwendung der Kollokation durch einen Beispielsatz untermalt. Die Adjektiv-Kollokate werden davon abgesetzt durch den Satz "wie der 'Hass' sein kann" eingeleitet. Eine Sortierung der Kollokate findet weder unter den Verben noch den Adjektiven nach alphabetischen oder semantischen Kriterien statt.

In keinem der beschriebenen elektronischen Wörterbüchern werden die externen Kollokationsinformationen, wie sie in Abbildung 7 zusammengefasst sind, auch nur annähernd exhaustiv verzeichnet. Obwohl für das elektronische Medium das Platzproblem der Print-Medien entfällt, enthalten etliche der in Kapitel 3.2.3 vorgestellten Papier-Wörterbücher mehr Informationen und sind außerdem systematischer aufgebaut. Gerade in den Online-Wörterbüchern wäre es möglich, die Kollokationen in übersichtlicher Form zunächst als einfache Wortkombinationen anzuzeigen, um dann über einen weiteren Link die gewünschten zusätzlichen Informationen zu bieten. Für den Wörterbuchbenutzer von Vorteil wäre sicher eine primäre Anordnung nach der Wortart der Kollokate. Ob diese alphabetisch sortiert oder in absteigender Reihenfolge nach den Werten eines der Assoziationsmaße erscheinen, oder nach Synonymen gruppiert werden, ist die Entscheidung der Lexikografen. Auch über die internen Kollokationsrelationen sind in keinem der Wörterbücher im WWW Angaben zu finden. Zudem ist eine Weiterverarbeitung der Kollokationsdaten mit anderen sprachverarbeitenden Systemen auf der Grundlage der repräsentierten Datenformate nicht möglich. Die elektronischen Wörterbücher bieten den kostenlosen Zugang zu Sprachdaten, doch schöpfen sie in Bezug auf die Kollokationen die Möglichkeiten der elektronischen Datenverarbeitung nicht aus.

Eine sehr ausführliche Darstellung der Kollokationen erfolgt in zwei Online-Wörterbüchern, die auf der Theorie der lexikalischen Funktionen beruhen, dem *DiCouèbe (Dictionnaire en Ligne de Combinatoire du Français)* und dem *DiCE* (*Diccionario de Colocaciones del Español*). Erwähnt wurden sie bereits in Kapitel 2.4.1, wo auch die lexikalischen Funktionen erklärt sind und sich Lexikoneinträge deutscher Gefühlssubstantive im Format des *DEC* (*Dictionnaire explicatif et combinatoire du français contemporain*) befinden, das die lexikografische Realisierung der *Meaning-Text Theory* in Papierform darstellt. Die formale Datenbasis *DiCo (Dictionnaire de Combinatoire)* ist die elektronische Weiter-entwicklung des *DEC*, sie ist als Grundlage von NLP-Systemen und von Wörterbüchern konzipiert. Auf

95 http://dev.eurac.edu:8081/MakeEldit1/Eldit.html

der Webseite des *Observatoire linguistique Sens-Texte*[96] an der Universität Montreal gibt es ein Handbuch, in dem Prinzipien, Datenstruktur und Implementierung ausführlich erläutert sind. In der vorliegenden Arbeit wurden Ausschnitte aus *DiCo* in Kapitel 3.2.3 wiedergegeben, zusammen mit einem Artikel aus *LAF (Lexique actif du français),* der Informationen zu Kollokationen in einer einfach verständlichen und gut strukturierten Form für den Sprachenlerner bietet. Das *LAF* soll direkt aus der Datenbank *DiCo* kompiliert werden. Ein entsprechender Platz für das *LAF* ist auf der Webseite des *Observatoire linguistique Sens-Texte* unter dem Stichpunkt "Teaching/Learning the Lexicon" bereits vorgesehen, doch befinden sich hier noch keine Einträge.
Das *DiCouèbe* als Interface erlaubt Anfragen auf eine kompilierte Version von *DiCo* im Internet, die Datenbank besteht derzeit aus 519 Vokabeln des Grundwortschatzes. Im Gegensatz zu den weiter oben beschriebenen Online-Wörterbüchern, bietet sich dem Benutzer bei der Auswahl aller verfügbaren Parameter eine wahre Flut an Informationen dar. Die Regulierung der angezeigten Ergebnisse ist über das Interface möglich, auch die Benutzung des Interface ist im Handbuch ausführlich beschrieben. Die Struktur des Anfrageformulars gliedert sich im Expertenmodus in 7 Bereiche (Lexik, lexikalische Funktionen, semantische Aktanten, syntaktische Aktanten, (idiomatische) Wendungen, Beispielsätze, Metainformationen zur Artikelerstellung), deren Unterpunkte wiederum individuell auszuwählen und bei Bedarf spezifisch einzuschränken sind. So kann man beispielsweise nur nach bestimmten semantischen Aktanten oder lexikalischen Funktionen suchen. Es lassen sich auf diese Weise sehr präzise Ergebnisse mit den gewünschten Informationen zum Lemma erzielen. Dargestellt werden die Ergebnisse im Tabellenformat. Es können alle verschiedenen Lesarten des Lemmas untereinander angezeigt werden, oder nur spezifische vom Benutzer selektierte Lesarten. Für das Lemma *espérance* ('Hoffnung') erhält man bei der Wahl aller Parameter ohne Einschränkungen eine Darstellung mit 1579 Ergebnissen in Zeilen, die 42 Spalten entsprechen den 7 Bereichen des Anfrageformulars mit ihren Unterpunkten. Das *DiCouèbe* ist für den linguistisch versierten Benutzer, der mit der Theorie der lexikalischen Funktionen vertraut ist, sicherlich von Nutzen, überfordert jedoch den durchschnittlichen User eines Wörterbuchs durch das vorausgesetzte Wissen über eine sprachwissenschaftliche Theorie bei der Abfrage und der darauf basierenden Darstellung der Ergebnisse.
Das *DiCE* der Universität Coruña bietet eine geringere Anzahl an untersuchten Lemmata, 10 Gefühlssubstantive, und legt seinen Schwerpunkt auf die Darstellung der syntagmatischen lexikalischen Funktionen. Für die verschiedenen Lesarten der Substantive werden zunächst die generische Bedeutung, die propositionale Form, ein Beispielsatz, Quasisynonyme und -antonyme angegeben. Unter jeder Lesart kann der Benutzer wählen, ob in einem neuen Fenster die Attribute der Partizipanten, die Kombinationen mit Adjektiven, Substantiven oder Verben angezeigt werden. Die Darstellung der Substantiv-Verb Kollokationen ist nach der Subjekt- oder Objektfunktion des Substantivs untergliedert, die beiden zugeordneten Klassen werden beispielsweise "admiración'+verbo" und "verbo+'admiración'" genannt.[97] Anders als im *DiCouèbe* wird der Schwerpunkt nicht auf eine exhaustive Beschreibung der

96 http://www.olst.umontreal.ca/dicoeng.html

97 Ziel dieses Vorgehens ist wohl die einfache Verständlichkeit für den Benutzer des Interfaces, für den das Wörterbuch damit auch ohne linguistisches Wissen zugänglich ist. Tatsächlich ist aber im Spanischen, wie auch im Portugiesischen, die Subjektfunktion des Substantivs nicht immer an die Satzstellung vor dem Verb gebunden, wie der Beispielsatz unter *admiración+envolver* zeigt: "Le envuelve la admiración de sus conciudadanos".

Lemmata im Rahmen der *Meaning-Text Theory* gelegt, das Wörterbuch wendet sich vermutlich an Sprachenlerner, denn neben dem Wörterbuch bietet das Projekt auch umfangreiche Übungen mit Kollokationen im Bereich der Sprachproduktion und des Sprachverständnisses an (explizit wird die Zielgruppe jedoch nirgends genannt).
Hat der Benutzer im *DiCE* den Typ der Wortkombination gewählt, zu dem er genauere Informationen wünscht, erscheinen die Kollokate semantisch geordnet, nach der Zugehörigkeit zu lexikalischen Funktionen, und innerhalb einer lexikalischen Funktion in alphabetischer Reihenfolge. Die lexikalischen Funktionen werden im *DiCE* durch Glossen umschrieben, denn für den Benutzer, der die *Meaning-Text Theory* nicht kennt, stellt sich die Benennung der Relationen, die zwischen den Kollokationspartnern bestehen, mit lexikalischen Funktionen in höchstem Maße kryptisch dar. Die lexikalische Funktion '$Oper_1$' wird beispielsweise durch die Glosse *SENTIR* wiedergegeben, '$Caus_2Func_1$' durch *CAUSAR A EN ALGUIEN*. Die Umschreibung der lexikalischen Funktionen mit popularisierten Ausdrücken wird auch im *LAF* verwirklicht, die Glosse als Meta-Sprache ist auch für den Sprachenlerner gut zu verstehen. Im *DiCouèbe* kann man die lexikalischen Funktionen ebenfalls mit Glossen anzeigen lassen, in einer zusätzlichen Tabellenspalte. Das *DiCE* bietet wiederum die Möglichkeit, außer der Glosse auch den Namen der lexikalischen Funktion durch Mouseover zu betrachten. Im *DiCE* wird für jede Kollokation ein Beispielsatz gezeigt, sowie in eckigen Klammern Angaben zur Rektion, die sich auf den Gebrauch des Artikels, der typischen Präposition zwischen Kollokat und Basis, und des direkt auf das Substantiv folgenden Aktanten belaufen. Im *DiCE* findet der Benutzer somit einen Teil der externen Kollokationsinformationen in einer systematischen Form, es fehlen jedoch pragma-tische Angaben zum Usus (Frequenz und Diasystematik), zum Numerus der Nomina, ausführliche Informationen zur Nomen- und zur Verbvalenz, sowie spezifische Modifika-toren der Nomina. Informationen zu internen Kollokationsrelationen sind implizit in der Zuordnung zu bestimmten lexikalischen Funktionen enthalten. Passen die Kollokate nicht zu den vorgegebenen lexikalischen Standardfunktionen, weil sie sehr spezifisch sind, erfolgt die Umschreibung mit gebräuchlichen Synonymen. Zusätzlich bietet das *DiCE* für jedes Kollokat einen Link, der Aufschluss gibt über alle untersuchten Substantive, mit denen das Kollokat kombiniert.

6.1.3. Darstellung der verbalen Kollokationen der Gefühlssubstantive

6.1.3.1. Form der Wörterbuchartikel

In den folgenden Wörterbuchartikeln der portugiesischen Gefühlssubstantive richtet sich das primäre Ordnungskriterium innerhalb eines Wörterbuchartikels nach der Semantik der Kollokate. Die Auflistung der Kollokate in alphabetischer Reihenfolge, nach ihrer Frequenz oder den Werten eines der statistischen Assoziationsmaße, wird nicht als optimal empfunden. Entsprechend der Strukturierung der Kollokate im *Oxford Dictionary of Collocations* und etlichen weiteren Print-Wörterbüchern wird eine Gruppierung nach den semantischen Beziehungen der Kollokate angestrebt. Diese Vorgehensweise bietet dem menschlichen Benutzer den Vorteil, in einem monolingualen Wörterbuch über bekannte Wörter in der gleichen Gruppe auf die Bedeutung des Kollokats zu schließen und gleichzeitig sein Wissen in der Fremdsprache über die synonymen Kollokate zu erweitern. Auch den NLP-Systemen, die auf ein elektronisches Wörterbuch zugreifen, wird eine Auswahlmöglichkeit an Lexemen geboten, um Wiederholungen bei der Textgenerierung zu vermeiden. Die Strukturierung in Kollokatsbereiche erfolgt durch die generalisierende

Bedeutungsangabe in Form von Glossen. In einem Bereich von Quasisynonymen werden die Kollokate nicht alphabetisch sortiert, sondern entsprechend ihrem Platz in der nach t-score berechneten Rankingliste verzeichnet. Dadurch bekommt der Benutzer auf den ersten Blick eine Übersicht über die Etabliertheit der Kollokationen innerhalb eines syntaktisch und semantisch beschränkten Bereichs.
Die Wörterbucheinträge sind in zwei Bereiche unterteilt (vgl. Kapitel 6.2). Im oberen Bereich sind die Informationen zu finden, die einen Wörterbuchbenutzer interessieren, der einen Text in der Fremdsprache verfassen will, und der dazu neben den passenden Kollokaten Angaben zur (morpho)syntaktischen Realisierung der Kollokation braucht. In einer elektronischen Form des Wörterbuchs wäre die Darstellung weiterer spezifischer Informationen über einen Link zu realisieren, der sie mit dem betreffenden Kollokat oder der Glosse verbindet. In der hier vorliegenden Papierform erscheinen diese detaillierten Informationen zu den Kollokaten und Glossen eines Substantivs im unteren Bereich des Wörterbuchartikels, der durch zwei Querlinien abgesetzt ist.
Die Zielgruppe der vorliegenden Wörterbuchartikel sind deutschsprachige Benutzer, die geeignete verbale Kollokate und präzise Kollokationsinformationen im Portugiesischen suchen. Ein bilingualer Wörterbuchaufbau in beiden Sprachrichtungen bietet sich aufgrund der fehlenden (morpho)syntaktischen Informationen für deutsche Substantiv-Verb Kollokationen nicht an. Da als Grundlage des Wörterbuchs nur die Extraktion von Kollokationen aus portugiesischen Corpora dient, fehlen auch die portugiesischen verbalen Äquivalente von deutschen Substantiv-Verb Kollokationen (*enfurecer* 'in Wut geraten', *preocupar-se com alguém* 'sich um jdn Sorgen machen').
Das Format der Wörterbucheinträge der Gefühlssubstantive bietet die Möglichkeit, Übersichtlichkeit mit Ausführlichkeit zu verbinden. Die Vordergrundinformationen sind für einen Wörterbuchbenutzer bestimmt, der (morpho)syntaktische Informationen bei der Textproduktion benötigt. Verwendbar sind die Wörterbuchartikel auch im umgekehrten Fall, wenn sich Verständnisschwierigkeiten bei einer portugiesischen Kollokation einstellen. Zu finden sind die Kollokationen in einer elektronischen Form des Wörterbuchs über ein Eingabefeld, in welches das deutsche oder portugiesische Kollokat und das Substantiv einzufügen sind. Eine zusätzliche Option ist denkbar, über die der Benutzer wählen kann, ob er sich nur die betreffende Kollokation, die weiteren Kollokate der Glosse, oder den gesamten Wörterbuchartikel des Substantivs anzeigen lassen will. Im Papierformat sind die geeigneten Kollokate im Portugiesischen über die generalisierende Bedeutungsangabe der Glossen zu finden. Die erste Glosse unter den Gefühlssubstantiven vereint die Kollokate, die die Empfindung des Gefühls beschreiben, die zweite Glosse diejenigen, die die Ursache benennen. Danach variiert die Reihenfolge der Glossen je nach Substantiv, wobei die Glossen mit den frequenten Kollokaten am Anfang stehen, um dadurch für den Großteil der Kollokate ein schnelles Auffinden zu erleichtern, und die kollokationalen Spezifika der Substantive zu zeigen. Grundsätzlich werden zuerst die Kollokationen genannt, in denen das Gefühlssubstantiv die Objektposition einnimmt, darauf folgen die Kollokationen, in denen das Gefühlssubstantiv in der Position des Subjekts steht.
Die Kollokation wird unter der Basis verzeichnet. Dieses Vorgehen entspricht dem Charakter der Kollokation, da die Wahl der Kollokate (häufig) idiosynkratisch durch die Basis erfolgt, und sich daher in diesem Bereich die meisten Fragen stellen. Zum anderen tragen die Kollokate, und mitunter auch spezifische Subkategorisierungseigenschaften der Verben, zur Bedeutungsdifferenzierung der Basis bei, wie Kapitel 6.3 ausführlich zeigt. Es

wird daher von einer doppelten Funktion der Wörterbucheinträge ausgegangen, sie erlauben das Auffinden der passenden Kollokate und bieten eine Disambiguierungshilfe bei polysemen Substantiven. Wünschenswert wäre ebenso ein Eintrag der gebräuchlichen Substantive unter den Verben, was in Kapitel 3.2.1 schon diskutiert wurde. Eine Strukturierung von Wörterbuchartikeln, in denen Verben als Lemmata fungieren, wäre nach dem Modell von Cowie (Kapitel 3.1.1) oder Welker (Kapitel 4.1) möglich. Da das Wortfeld der Substantive der Gefühle zu homogen ist, um Unterschiede in der Verbbedeutung aufzuzeigen, beschränkt sich die Darstellung der Kollokationen in den folgenden Kapiteln auf Wörterbuchartikel mit substantivischen Lemmata.

6.1.3.2. Vom PECCI-Output zum Wörterbucheintrag

Nach einem Programmdurchlauf von PECCI liegt eine Rankingliste der Substantiv-Verb Kookkurrenzen sortiert nach den Werten eines statistischen Assoziationsmaßes vor. Die primäre Ausgabedatei zeigt, da sie der lexikografischen Weiterverwertung dienen soll, die verbalen Kollokate separat für jedes untersuchte Gefühlssubstantiv. Weitere Ausgabeformate, die die Kookkurrenzen sortiert nach den Verben oder für das gesamte Wortfeld enthalten, stehen ebenfalls zur Verfügung. Zu Beginn des Kapitels wurde noch einmal die Zweckmäßigkeit der Darstellung mit den beiden statistischen Assoziationsmaßen t-score und MI erläutert, die Auswahl von log-likelihood oder χ^2 ist ebenfalls möglich, und auch eine Datei, die die Werte der vier implementierten Assoziationsmaße anzeigt, wird erzeugt. Die automatische Extraktion der Kookkurrenz zweier Lexeme bietet heutzutage keine Probleme mehr. Auch die maschinelle Ausgabe der Frequenzen und der Werte der statistischen Assoziationsmaße ist ohne größeren Aufwand zu erreichen. Die Kollokationen in der vorliegenden Untersuchung sind homogen hinsichtlich der an ihnen beteiligten Wortarten und durch die Corpuswahl regional und stilistisch einheitlich zu bewerten. Diasystematische Markierungen könnten aber auch in einem heterogenen Corpus durch die Auswertung der Metadaten automatisch erzeugt werden.
Die Größe des Corpus mit ca. 200 Millionen Wörtern erlaubt es, auch seltenere Kollokationen mit mehreren Fundstellen auszuweisen, doch liegen andererseits für die gebräuchlichen Kollokationen Exzerptionsdateien für eine Substantiv-Verb Kollokation mit mehreren Tausend Sätzen vor, was eine exhaustive Auszählung, beispielsweise um prozentuale Angaben über den Artikelgebrauch oder bestimmte Subkategorisierungsrahmen zu bieten, nahezu unmöglich macht. Um diese Arbeit zu erleichtern, wird von PECCI für jede Kollokation eine zusätzliche Datei generiert, die für jede Fundstelle der Kollokation nur das direkt links neben der Kollokation stehende Wort, die zwischen Basis und Kollokat befindlichen Wörter und sechs Wörter rechts davon anzeigt. Hier erhält man einen schnellen Überblick über den Kontext der Kollokation. Außerdem werden einige morphosyntaktische Eigenschaften der Kollokation mit numerischen Werten belegt. Darunter fällt der Gebrauch des Artikels und der Präpositionen zwischen den Kollokationspartnern. Automatische Angaben zu spezifischen (beispielsweise adjektivischen) Modifikatoren wären nur in einem Corpus mit POS-Tags zu erzielen (oder mit einem beträchtlichen zusätzlichen Implementierungsaufwand). Numerische Angaben zu Subkategorisierungseigenschaften sind nur in der Form enthalten, als dass häufig gebrauchte Konjunktionen und Präpositionen, die der Kollokation folgen, gezählt werden. Dies gibt natürlich keinen Aufschluss darüber, ob bestimmten Präpositionen ein ganzer Nebensatz oder nur ein Nomen folgt, über den Aufbau des Nebensatzes oder den Kasus des Nomens. Auch über die Struktur des gesamten Satzes,

in dem die Kollokation vorkommt, können keine Aussagen getroffen werden. Diese Informationen können nur vom Menschen aus den Exzerptionsdateien ausgelesen werden. Für die automatische Ausgabe der Subkategorisierungs- oder gar Valenzeigenschaften einer Kollokation müsste das Corpus vollständig geparst sein.
Weil vollständig geparste Corpora mit einer geeigneten Größe zur Kollokationsextraktion heute noch nicht existieren, können präzise Kollokationsinformationen auf der syntaktischen Ebene nur vom Menschen zusammengestellt werden. Dadurch ist auch zu erklären, warum die Online-Wörterbücher, die genauere Informationen bieten, entweder nur wenige Lemmata mit sehr vielen Kollokaten (*DiCE, DiCouèbe, elexiko*), oder viele Lemmata mit nur wenigen Kollokaten (*LEO, ELDIT*) enthalten. Auch in der Darstellung der Kolloka-tionen in den Print-Wörterbüchern sind genaue Kollokationsinformationen nur selten zu finden. Da die bloße Auflistung der Kollokate weder für einen Wörterbuchbenutzer noch für die Weiterverarbeitung mit NLP-Systemen ausreichend ist, werden im folgenden Kapitel die Kollokationen zwar mit präzisen Informationen, aber nur für einige wenige Gefühls-substantive dargestellt. Wie wichtig genaue Kollokationsinformationen gerade im Bereich der Übersetzung sind, wurde in Kapitel 3.3 skizziert.
Die (morpho)syntaktischen Informationen für die Zielsprache werden in einer linearen, leicht verständlichen Notation für jede Kollokation angegeben, um die Kollokationspartner in der für sie üblichen Umgebung zu zeigen. Aufgeführt werden neben dem Artikelgebrauch und dem Numerus des Substantivs auch Präpositionen, die zwischen Basis und Kollokat stehen, sowie spezifische Modifikatoren. Sind der bestimmte sowie der unbestimmte Artikel gebräuchlich, bezeichnet DET die Wahlmöglichkeit; ist die Kollokation auf einen Artikel festgelegt, wird dieser aufgeführt; ist es zudem möglich, dass kein Artikel steht, wird das Nicht-Vorkommen des Artikels durch ein zusätzliches Zeichen markiert ('a | uma | Ø'), die Reihenfolge richtet sich nach der Häufigkeit des Auftretens im Corpus; wird zwischen Basis und Kollokat grundsätzlich kein Artikel verwandt, fehlt jegliche Notation.
Die Kollokation wird mit der Numerusform des Substantivs verzeichnet, die häufiger in dieser Wortkombination im Corpus auftritt. Da dies allein wenig Aufschluss darüber gibt, ob nur diese Numerusform gebräuchlich ist, oder ob sie in der Wortkombination nur leicht präferiert ist, ist für jede Kollokation die Frequenzangabe für die Singular- und Pluralform der Kookkurrenz über einen Link zu finden, der diese Zusatzinformationen mit der Glosse verbindet. Begleitet wird die Frequenz von den Werten des t-score und der MI, die jeweils für Singular und Plural des Substantivs separat berechnet sind. Fakultative und obligatorische Aktanten der Basis und des Kollokats werden wiederum im sichtbaren Teil des Wörterbuchartikels verzeichnet, zusammen mit den von der Kollokation abhängigen Nebensatzstrukturen. Zu jeder Kollokation wird ein Beispielsatz geboten für die häufigste (morpho)syntaktische Struktur. Mitunter liegen für eine Substantiv-Verb Kollokation stark unterschiedliche syntaktische Realisierungsmöglichkeiten vor, in diesem Fall werden diese einzeln dargestellt und jeweils durch einen Beispielsatz belegt.
Die Darstellung weiterer spezifischer Informationen zu jeder Kollokation wäre wieder über einen Link zu realisieren, der sie mit der betreffenden Kollokation verbindet. Denkbar wären prozentuale Angaben zum Artikelgebrauch oder zu den einzelnen Subkategorisierungs-rahmen. In der vorliegenden Arbeit werden in diesem Teil des Wörterbuchartikels die Kollokationen gezeigt, die in portugiesischen und portugiesisch-deutschen Wörterbüchern aufgeführt werden. Dadurch ergibt sich die Möglichkeit, die Ergebnisse der corpusbasierten lexikalischen Akquisition mit den Angaben in den Print-Wörterbüchern zu vergleichen. Die

Kennzeichnung der portugiesischen Kollokationen, die in einem der Wörterbücher erwähnt werden, erfolgt durch Unterstreichung. Es fließen die Einträge folgender Nachschlagewerke mit ein:

AU - *Novo Dicionário Aurélio da Língua Portuguesa.* (1986)
DC - *Dicionário Contextual Básico da Língua Portuguesa.* (2000)
IDP - *Idiomatik Deutsch-Portugiesisch.* (2002)
LS - *Langenscheidts Taschenwörterbuch Portugiesisch. P-D, D-P* (2001)
PO - *Pons Standardwörterbuch. Portugiesisch-Deutsch,Deutsch-Portugiesisch.* (2002)

Bei einem Vergleich der Angaben zu den Kollokationen in den Wörterbüchern mit den Daten der Corpusextraktion wird offensichtlich, dass sowohl in dem Kollokationswörterbuch wie auch in den idiomatischen oder den allgemeinen Wörterbüchern nur ein kleiner Teil der aus dem Corpus extrahierten Kollokationen zu finden ist, wobei die Entscheidung über die Auswahlkriterien der aufgeführten Kollokationen in den Wörterbüchern nicht immer nachzuvollziehen ist. Nur für ein Nomen (*esperança*) werden zahlreiche verbale Kollokate verzeichnet, daher beschränkt sich die Darstellung bei vielen der portu-giesischen Gefühlssubstantiven auf die Kollokationsdaten aus der lexikalischen Akquisition.

In den Wörterbüchern werden die Kollokationen mitunter beim Verb aufgeführt. Da es nicht möglich ist, die Einträge aller Verben durchzusehen, wurden nur die Verben kontrolliert, die häufig mit den Gefühlssubstantiven kookkurrieren. Erfolgt die Eintragung der Kollokation im Wörterbuchartikel des Verbs, wird dies durch ein Subskript am entsprechenden Wörterbuch angezeigt (PO_V). Häufig wird eine Kollokation nur im portugiesischen oder deutschen Teil des Wörterbuchs genannt. Bestimmte Kriterien, die zur Wahl einer Sprachrichtung führen, sind nicht ersichtlich. Erfolgt die Nennung der Kollokation in beiden Sprachrichtungen, wird dies durch das nachgestellte Symbol <-> gekennzeichnet. In den Wörterbüchern fehlen viele der frequenten Kollokationen aus den Rankinglisten von PECCI, sowohl diejenigen, die interlingual idiosynkratisch sind, als auch diejenigen, mit einem wörtlichen Übersetzungsäquivalent. Dazu im Gegensatz sind mitunter Kollokationen zu finden, die keine einzige Belegstelle im portugiesischen oder brasilianischen Corpus aufweisen, wie die Übersetzung von *schüren* mit *atiçar* (*Eifersucht*), in der vollständig neu entwickelten Ausgabe des *PONS Standardwörterbuchs* (2002). Tatsächlich ist *atiçar* nur mit einigen wenigen anderen Substantiven gebräuchlich. Die "Kollokationen" der Wörterbücher ohne Corpusbelege sind unter den Stichpunkt "Anmerkungen" im unteren Teil des Wörterbuchs zu finden.

Für die Aufnahme einer Substantiv-Verb Kookkurrenz aus den Rankinglisten von PECCI in die Wörterbuchartikel der Gefühlssubstantive sind verschiedene Faktoren ausschlaggebend. Der Usus in den Kollokationswörterbüchern im Vergleich zu den Definitionskriterien aus Abbildung 6, ist Thema von Kapitel 3.2.4. Dort wird auch die Motivation für die Aufnahme etablierter und typischer Kollokationen diskutiert. Die folgenden Punkte präzisieren die Aufnahmekriterien, die sich aufgrund der Corpusannotation von *Cetempúblico* und *Cetenfolha,* der Programmstruktur von PECCI und dem Vergleich mit existierenden Wörterbüchern ergeben:

1. Analog zum Usus in den Kollokationswörterbüchern ist die Gebräuchlichkeit einer Substantiv-Verb Kombination das Kriterium, das zur Aufnahme in die Wörterbucheinträge führt. Um der extrem unterschiedlich ausfallenden Frequenz einzelner Substantive (*asco(s)* 29, *pena(s)* 26088 in *Cetempúblico*) gerecht zu werden, werden Schwellenwerte für die absolute Frequenz der Kookkurrenz und für die Werte von t-score

und MI angesetzt. Hohe Werte der MI deuten oftmals auf wenig frequente Wörter hin, die sehr viel häufiger miteinander kombinieren, als dies aufgrund der Unabhängigkeitsannahme zu erwarten ist (vgl. Kapitel 2.1). Auch diese Wortkombinationen sind etabliert, gemessen am seltenen Auftreten der einzelnen Wörter. Das einmalige Vorkommen einer Kookkurrenz gilt jedoch nicht als Indikator, die Wortkombination muss mindestens zweimal im Corpus vertreten sein und einen MI-Wert größer als 3 aufweisen, um als Kollokationskandidat zu gelten. Der Mindestwert des t-score einer Kookkurrenz wird mit 2 festgelegt, die absolute Frequenz mit 5 Vorkommen einer Kookkurrenz. Die anhand der numerischen Werte in Frage kommenden Kookkurrenzen bedürfen der kritischen Durchsicht durch einen Lexikografen.[98]

2. Die Fundstellen, in denen das Gefühlssubstantiv als Bestandteil eines Nomenkompositums vorliegt, sind in den von PECCI generierten Exzerptionsdateien enthalten, denn im Portugiesischen kombinieren die beteiligten Substantive nicht zu einem Wort wie im Deutschen (*esperança de vida* - 'Lebenserwartung'). Lexikalisierte Nomenkomposita unterscheiden sich meist deutlich in ihrem Kollokationsverhalten gegenüber dem Kopflexem (vgl. Zinsmeister/ Heid 2004), dies ist auch bei *esperança de vida* der Fall. In den Wörterbuchartikeln der Gefühlssubstantive werden Nomenkomposita, die ein Gefühlssubstantiv enthalten, generell nicht verzeichnet, das Aussortieren der Fundstellen ist nur manuell zu leisten.
3. In die Wörterbuchartikel aufgenommen werden Substantiv-Verb Kombinationen, die fixierte Redewendungen (*vale a pena* 'es lohnt sich/die Mühe') bilden, sie werden zusammen mit bedeutungsähnlichen Kollokationen unter einer Glosse verzeichnet. Hinter den Glossen gibt es einen Bereich, in dem auch Idiome oder Sprichwörter stehen können, wenn diese das betreffende Substantiv und ein lemmatisiertes Verb enthalten. Im Gegensatz zu den Kollokationen und vielen fixierten Redewendungen ist bei ihnen die Gesamtbedeutung nicht aus der Bedeutung der einzelnen Bestandteile herzuleiten. Daher werden sie unter dem Stichpunkt "Idiome" verzeichnet, denn eine Gruppierung nach Glossen ist hier nicht möglich, die Übersetzung muss für jedes Idiom individuell erfolgen.
4. Kollokationen, die in den zum Vergleich untersuchten (deutsch-)portugiesischen Wörterbüchern stehen, die sich aber nicht im Corpus befinden, werden nur unter dem Link verzeichnet, der zusätzliche Informationen zu jeder Glosse zeigt. Die "Kollokationen" der Print-Wörterbücher ohne Fundstellen im Corpus sind entweder veraltet, oder werden bei Nachfrage auch von einem portugiesischen Muttersprachler als unüblich empfunden. Die in den Wörterbüchern genannten Kollokationen, die im Corpus mindestens eine Fundstelle aufweisen, stehen im vorderen Teil des Wörterbuchartikels zusammen mit den Kollokationen, die die numerischen Kriterien zur Aufnahme erfüllen.

Neben den präzisen (morpho)syntaktischen Information für die portugiesischen Substantiv-Verb Kollokationen sind im vorderen Teil des Wörterbuchs die Übersetzungen der portugiesischen Kollokate ins Deutsche zu finden. Ein NLP-System, das auf ein Wörterbuch in elektronischer Form zugreift, benötigt für die Verarbeitung der Kollokate eine formale Definition ihrer Bedeutung. Als linguistisches Modell zur Bedeutungsangabe der Kollokate

98 Die gewählten Schwellenwerte für MI, t-score oder absolute Frequenz, die eine Substantiv-Verb Kookkurrenz erreichen muss, um als möglicher Aufnahmekandidat in das Wörterbuch zu gelten, wurden empirisch festgelegt. Mathematisch fundierte und anhand einer umfangreichen Evaluierung motivierte Werte sind in Evert (2005a) zu finden.

dienen im folgenden die lexikalischen Funktionen aus der *Meaning-Text Theory*. Wie bereits in Kapitel 2.4.1 ausführlich erläutert, spielen die lexikalischen Funktionen innerhalb eines maschinellen Übersetzungssystems die Rolle einer Interlingua auf der Ebene einer syntaktischen Tiefenstruktur. In diesem Ansatz sind präzise monolinguale Eintragungen der Kollokationen ausreichend: "it suffices to reduce the source-language collocation to its LF-representation, then translate the keyword only, and, finally, to select the value of the LF for the equivalent of the keyword in the target language" (Mel'čuk 1998: 45). Doch wird die folgende Darstellung der Gefühlssubstantive die Möglichkeiten der *Meaning-Text Theory* nicht voll ausschöpfen. Es sollen Einträge entstehen, die für einen Menschen leicht zu interpretieren sind, und in denen die lexikalischen Funktionen gruppierend wirken. Dass eine Strukturierung des Wörterbuchartikels anhand der lexikalischen Funktionen auch für den menschlichen Benutzer von Vorteil ist, demonstrieren die Beispiele aus *LAF* (Kapitel 3.2.3) und *DiCE* (Kapitel 6.1.2).

6.1.3.3. Semantische Beschreibung der Kollokationen

Ein Nachteil der Reduktion einer natürlichen Sprache auf eine Interlingua besteht darin, dass gewisse Nuancen der Sprache in der verallgemeinernden Bedeutungsangabe verloren gehen. Bei bildlichen Kollokaten im Portugiesischen, die in der wörtlichen Übersetzung im Deutschen nicht üblich sind, wird daher in den Lexikoneinträgen der Gefühlssubstantive die primäre Verbbedeutung als Übersetzung ebenfalls verzeichnet. Diese Kollokate werden mit einer '1' für kollokationale Divergenzen in der Übersetzung gekennzeichnet. Ein Beispiel für ein portugiesisches Kollokat, das nicht wörtlich ins Deutsche übersetzt werden kann, ist *acalentar* ('aufwärmen') in der Kombination mit *esperança* (in Abb. 8 wurden weitere Beispiele gezeigt), ein im Portugiesischen nicht gebräuchliches Kollokat ist *schöpfen* in der Kombination mit *Hoffnung*. Meistens haben diese Kollokate nur einen sehr engen Kollokationsradius, d.h. sie kombinieren nur mit einem oder wenigen Gefühlssubstantiven. Kollokate, die mit mehreren Substantiven aus dem Wortfeld auftreten, haben häufig eine bildliche Entsprechung im Deutschen wie *alimentar esperança* ('Hoffnung nähren').
Informationen über den Kollokationsradius der Verben erhält man in der von PECCI generierten Datei, in der die Kollokationen nach Verben sortiert erscheinen. Zusätzliche Angaben über auffällige Kollokate enthält die Ausgabedatei, die die Samplerelevanz der Verben anzeigt (den prozentualen Anteil der Kookkurrenzen eines Verbs mit den Substantiven des untersuchten Samples am Gesamtvorkommen des Verbs im Corpus), hier können auf einen Blick die Kollokate erfasst werden, die auffällig häufig mit den Wortfeldteilnehmern kombinieren. Ausschnitte aus beiden Dateien befinden sich im Anhang. In diesem Zusammenhang wird der Vorteil einer Untersuchung deutlich, die den Wortschatz in Wortfelder unterteilt. In der Datei mit den Daten der Samplerelevanz können signifikante Kollokate für Substantive, die bestimmte semantische Eigenschaften teilen, ermittelt werden. Kombiniert ein Kollokat sehr viel häufiger als zu erwarten mit den Gefühlssubstantiven als mit anderen Wörtern kann dies an vier verschiedenen Faktoren liegen:

- das Kollokat kombiniert restriktiv mit nur einem Gefühlsnomen des Wortfeldes,
- das Kollokat kombiniert mit mehreren, innerhalb des Wortfeldes semantisch ähnlichen, Substantiven,
- das Kollokat kombiniert mit mehreren, innerhalb des Wortfeldes semantisch stark verschiedenen, Substantiven,
- das Kollokat kombiniert mit dem gesamten Wortfeld.

Die Unterteilung des Wortschatzes nach untersuchten Wortfeldern in Form von Samples, wie sie in PECCI praktiziert wird, wäre in einem Corpus, das semantische Annotationen enthält, redundant. Durch die Integration semantischer Informationen in das Corpus, die beispielsweise EuroWordNet (vgl. Kapitel 2.4.2) bereitstellt, wäre es möglich, mit gezielten Anfragen, die Angaben zu Basis- und Topkonzepten enthalten, ausschließlich Substantive mit den gewünschten semantischen Eigenschaften zu extrahieren. Doch stehen Corpora mit semantischen Annotationen heute noch nicht zur Verfügung.
Die ursprüngliche Motivation für eine Untersuchung des Corpus anhand von Wortfeldern bestand darin, einen Vergleich zwischen den von Mel'čuk und Wanner (1994) manuell erstellten Daten für deutsche Gefühlssubstantive, die eine Korrelation zwischen semantischen Merkmalen der Substantive und der restringierten lexikalischen Kookkurrenz zeigen, und einer automatischen Zuordnung der Substantive anhand ihrer verbalen Kollokate mit Clusterverfahren durchzuführen. Die Ergebnisse werden in Kapitel 7 erläutert. Daneben ergibt sich nach einem Programmdurchlauf von PECCI mit dem Wortfeld der Substantive der Gefühle auch die Möglichkeit, die Daten von Mel'čuk und Wanner mit den Ergebnissen einer portugiesischen Corpusauswertung zu vergleichen, sowohl die Angaben für die Substantive, als auch die für die Verben. In diesem Kapitel werden die Einträge der Gefühlssubstantive bei Mel'čuk und Wanner dazu dienen, die deutschen Verben, die bei ihnen unter den lexikalischen Funktionen aufgeführt werden, als Übersetzungsäquivalente für die entsprechenden portugiesischen Werte der lexikalischen Funktion in dem Teil des Wörterbuchartikels zu zeigen, der mit der Glosse verlinkt ist. Auch die portugiesischen Kollokationen werden in der vorliegenden Arbeit lexikalischen Funktionen zugeordnet und sind dort zu finden.
Um die Kollokate der Gefühlssubstantive bei Mel'čuk und Wanner zu ermitteln, muss man zunächst die semantischen Bedingungen, die jedes Kollokat an eine potentielle Basis stellt, aus dem generischen Lexikoneintrags von 'Gefühl' mit den Angaben zu den semantischen Eigenschaften der Substantive vergleichen (vgl. Kapitel 2.4.1). Nach dem Vergleich verbleiben einige Verben aus dem generischen Eintrag von 'Gefühl', andere Verben werden entsprechend des individuellen Lexemeintrags ersetzt, getilgt und neue hinzugefügt. Diejenigen Kollokate, die nicht aus dem generischen Lexikoneintrag von 'Gefühl' stammen, sondern unter dem spezifischen Nomen individuell genannt sind, werden im folgenden Kapitel kursiv gedruckt, um kollokationale Besonderheiten zu verdeutlichen.
Die *Meaning-Text Theory* beschreibt die verschiedenen Probleme, die bei der Übersetzung einer Kollokation auftreten können, durch einen Formalismus, dessen Notation für die Anwendung in der Maschinelle Übersetzung geeignet ist (vgl. Kapitel 2.4.1). Da die folgenden Wörterbuchartikel jedoch zunächst für Menschen als Wörterbuchbenutzer konzipiert sind, wird weitgehend auf eine formale Darstellung verzichtet. Kategoriale Divergenzen werden in der *Meaning-Text Theory* z.B. mittels der paradigmatischen lexikalischen Funktionen S_0, V_0, A_0 und Adv_0 beschrieben. Werden diese lexikalischen Funktionen auf ein Lexem angewandt, erhält man ein semantisches Derivat. Im Falle der Fusion eines Funktionsverbgefüges zu einem Vollverb würde folgende Regel greifen: $'Oper_1(V_0(admiração))'$ (*DEC* III: 38-39). Auf kategoriale Divergenzen wird in den Wörterbuchartikeln der portugiesischen Gefühlssubstantive mit einer '2' hingewiesen, und die Wortkombination im Deutschen zusätzlich genannt. Syntaktische Divergenzen werden mit einer '3' gekennzeichnet. Auf die Abbildung von Standardtransformationen für die Ersetzung einer lexikalischen Funktion in der Ausgangssprache durch eine andere in der Zielsprache wird verzichtet.

Die genaueste Darstellungsform für Kollokationen bieten heute die lexikalischen Funktionen. Mit ihnen ist es möglich, interne Kollokationsrelationen und externe Kollokationsinformationen präzise zu beschreiben. Der obere Bereich des Wörterbuchartikels, der sich in einer elektronischen Form des Wörterbuchs dem Benutzer als grafische Oberfläche zeigt, gibt allerdings nur einen stark reduzierten Kanon der lexikalischen Funktionen wieder. Die Wörterbucheinträge orientieren sich eher an den systematisierenden Eigenschaften, die mit dem lexikografischen Konzept der lexikalischen Funktionen verbunden sind, und schöpfen die Möglichkeiten der Bedeutungserklärung und - differenzierung nicht voll aus. Dies hat unterschiedliche Gründe:

1. Die formalen Eigenschaften der lexikalischen Funktionen bilden auf der einen Seite die Grundlage für eine Verarbeitung mit NLP-Systemen, doch übersteigen sie die Kompetenz eines Wörterbuchbenutzers, der mit der *Meaning-Text Theory* nicht vertraut ist, und bieten damit keine geeignete Form der Bedeutungsangabe der Kollokate. Weder mit den lateinischen Bezeichnungen noch mit den Subskripten, die die tiefensyntaktische Beziehung der Verben zu ihren Argumenten darstellen, kann dieser etwas anfangen. Die Umschreibung der lateinischen Bezeichnung mit Glossen stellt ein adäquates Mittel zur Bedeutungsangabe dar. Die Glossen variieren im Gegensatz zum gleich bleibenden Namen der lexikalischen Funktionen mit der Bedeutung des Substantivs, auf das sie Bezug nehmen. Eine mögliche Paraphrase für 'Oper1' mit dem Wortfeld der Gefühlssubstantive ist 'empfinden', während die Paraphrase mit Substantiven, die eine Tätigkeit bezeichnen, 'machen' ist. In diesem Sinne verliert der Formalismus der lexikalischen Funktionen durch die Benutzung von Glossen an Einheitlichkeit. Die paraphrasierenden Verben, die als Glosse dienen, sollen in einem einsprachigen Wörterbuch möglichst generell sein und aus dem Grundwortschatz stammen. In einem zweisprachigen Wörterbuch hingegen erscheint es angemessen, als Glosse neben dem sehr allgemeinen Kollokat auch übliche Kollokate in der Ausgangssprache zu zeigen, dies erleichtert dem Wörterbuchbenutzer die Zuordnung seines zu übersetzenden Kollokats zu einem Bereich von Quasisynonymen. Innerhalb des syntaktisch und semantisch bestimmten Bereichs von Quasisynonymen gibt es für die Übersetzung verschiedene Auswahlmöglichkeiten. In den deutsch/portugiesischen Wörterbüchern findet man als Übersetzung eines portugiesischen Kollokats mitunter verschiedene deutsche Verben, die aber als Werte derselben lexikalischen Funktion auftreten. Ist die Semantik des Kollokats spezifisch, lässt sich genau bestimmten, ob es ein semantisch bzw. bildlich äquivalentes Kollokat in der anderen Sprache gibt, ob man auf semantisch entleerte Funktionsverben zurückgreifen kann, oder die Übersetzung des Kollokats mit einem bildlich unterschiedlichen Kollokat angemessen erscheint. Zu jeder Glosse werden über den Link (hier dargestellt unterhalb des Wörterbucheintrags), die lexikalischen Funktionen angezeigt, die zur Glosse gehören. Die Werte werden einmal gefüllt mit den deutschen Verben, die Mel'čuk und Wanner (1994) als Werte der äquivalenten deutschen Substantive angeben, und zum Anderen durch portugiesische Kollokate, die den lexikalischen Funktionen aufgrund der Auswertung der Exzerptionsdateien, zugewiesen sind.
2. Die automatische Zuordnung einer Kookkurrenz zu einer bestimmten lexikalischen Funktion ist weder mit einem getaggten noch mit einem vollständig geparsten Corpus zu leisten. In einem geparsten Text könnte man die Fundstellen nach Subjekt- und Objektpositionen des Substantivs unterscheiden und Subkategorisierungsrahmen automatisch extrahieren, doch beinhalten die lexikalischen Funktionen mehr Informa-

tionen, die nur in einem semantisch annotierten Corpus enthalten sind. Eine semantische Annotation müsste zum einen die Abbildung der Verb- und Substantivargumente auf die syntaktisch realisierten Komplemente näher bestimmen, wofür sich eine Annotation im FrameNet-Format eignet. Zum anderen sind semantische Angaben, wie sie WordNet bietet, notwendig, um die Verben über ihre Bedeutungsangaben zu Quasisynonymen zusammenzufassen, oder sie auf die lexikalischen Standardfunktionen abzubilden, die in Kapitel 2.4.1 näher erläutert sind. Die manuelle Zuordnung zu lexikalischen Funktionen bedeutet einen erheblichen Arbeitsaufwand, die Klassifizierung nach Glossen, die mitunter einige ähnliche lexikalische Funktionen vereinen, ist für den Lexikografen sehr viel leichter. Sollen die Wörterbuchartikel zusätzlich als lexikografische Basis für ein maschinelles Übersetzungssystem dienen, sind jedoch auf jeden Fall Angaben zu den Kollokationen notwendig, die äquivalent sind zum Informationsgehalt der lexikalischen Funktionen, die zusätzliche Explikation ist dann auf jeden Fall zu leisten.

3. Die Exzerptionsdatei einer Kollokation kann hinsichtlich der syntaktischen Realisierungsmöglichkeiten der Kookkurrenz homogen oder heterogen ausfallen; dies bedeutet, dass die Vorkommen einer Substantiv-Verb Kollokation manchmal verschiedenen lexikalischen Funktionen zuzuordnen sind. Die syntaktische Variationsmöglichkeit einer Substantiv-Verb Kollokation ist ein Aspekt, der auch in den Wörterbüchern, die mit den lexikalischen Funktionen arbeiten, nicht immer präzise dargestellt ist. Beispielsweise gehört *provocar ciúmes* ('Eifersucht erregen') je nachdem, ob in dem realisierten Satz erwähnt ist, bei wem die Eifersucht erregt wird, oder nicht, zu der lexikalischen Funktion '$Caus_2Func_1$' oder '$Caus_2Func_0$'. Andere Kollokationen hingegen sind immer eindeutig, *pregar um susto* ('einen Schreck einjagen') verlangt immer die Realisierung des 3. Verbarguments als Aktanten. In einem vollständig geparsten Corpus könnten die Exzerptionsdateien automatisch nach den vorkommenden Subkategorisierungsrahmen unterteilt werden. Es stellt sich jedoch die Frage, ob es sinnvoll ist, dem Wörterbuchbenutzer die Informationen in dieser Ausführlichkeit mit Hilfe der lexikalischen Funktionen zu zeigen. Syntaktische Variationsmöglichkeiten, die die Anzahl der Aktanten betreffen, werden in den folgenden Wörterbucheinträgen der Gefühlssubstantive meist unter einer Glosse zusammengefasst. Durch die lineare Annotation der gebräuchlichen Subkategorisierungsrahmen erhält der Benutzer einen Überblick über die Verwendungsweise der Kollokation. Können bestimmte Komplemente bzw. Nebensätze stehen, oder auch nicht, wird die optionale Struktur durch runde Klammern gekennzeichnet. Das *DEC* wird den syntaktischen Möglichkeiten dadurch gerecht, dass bestimmte Kollokate innerhalb des Eintrags eines Substantivs mehrfach aufgeführt sind. Im *DiCE* hingegen wird für jede Substantiv-Verb Kollokation nur eine lexikalische Funktion angegeben, und mitunter stimmen die Beispielsätze nicht mit der lexikalischen Funktion überein, die der Kollokation zugeordnet ist, so wird für die Kollokation *provocar alegría* mit der lexikalischen Funktion '$Caus_2Func_1$' als Beleg 'Ver la media tostada provoca la alegría cordial' genannt. Das Zusammenfassen mehrerer lexikalischer Funktionen, die sich nur in ihren Subskripten unterscheiden, zu einer Glosse wird im *DiCE* und im *LAF* praktiziert, die für einen menschlichen Wörterbuchbenutzer konzipiert sind.
4. Unter einer Glosse zusammengefasst werden in den folgenden Wörterbucheinträgen der Gefühlssubstantive mitunter auch diejenigen lexikalischen Funktionen, die sich bezüglich des Handlungsverlaufs unterscheiden. Mel'čuk bezeichnet die lexikalischen Funktion 'Incep', 'Fin' und 'Cont' als "Phasals", sie kommen fast immer in Kombination mit

weiteren lexikalischen Funktionen vor. Die Verben, die den Handlungsverlauf bestimmen, werden unter der jeweiligen Glosse mit '+begin', '+end', '+continue' markiert. Dieses Vorgehen erscheint angebracht, wenn nur wenige Kollokate für eine Glosse existieren, die die gleiche tiefensyntaktische Relation zwischen Basis und Kollokat wiedergibt. Liegen hingegen zahlreiche Kollokate für die gleiche Phase einer Glosse vor, werden diese nach Glossen getrennt zusammengefasst, und die Bedeutung in der deutschen Benennung der Glosse zum Ausdruck gebracht. Diese Art der Unterteilung soll vor allem benutzerfreundlich sein, denn der Aufbau des Wörterbuchartikels wird auf diese Weise nicht von Glossen dominiert, unter denen sich nur ein Kollokat befindet. Auch die Unterscheidung Mel'čuks in "Auxiliaries" und "Realizations" wird aufgegeben. Sowohl im *DiCouèbe* als auch im *DiCE* werden Verben wie *nourrir* ('(er)nähren'), *perder* ('verlieren'), *matar* ('töten') oder *exultar* ('jubeln') mit den lexikalischen Funktionen 'Oper' und 'Func' bezeichnet, die eigentlich den Funktionsverben vorbehalten sind. Die beiden Kategorien "Auxiliaries" und "Realizations" werden unter einer Glosse zusammengefasst; handelt es sich um semantisch sehr ausdrucksvolle Verben, deren Übersetzung sich in den beiden Sprachen unterscheidet, wird dies (wie schon oben erwähnt) durch eine '1' gekennzeichnet, und das Übersetzungsäquivalent im Deutschen zusätzlich verzeichnet. Markiert werden diese Verben auch bei einer äquivalenten Übersetzung im Deutschen durch ein vorangestelltes '+intense'. Bringt ein Kollokat die Wiederholung einer Handlung zum Ausdruck, wird '+repetitiv' vorangestellt.

5. Das Hauptargument, das gegen die Anwendung der lexikalischen Funktionen spricht, ist jedoch die Inkonsistenz bei ihrer Verwendung. Die Zuordnung der Kollokate zu bestimmten lexikalischen Funktionen variiert in den Wörterbüchern verschiedener Autoren, sowie innerhalb desselben Wörterbuchs, auch wenn die Beziehung vom Verb zum Substantiv identisch ist, und selbst bei Mel'čuk unterliegt sie einem Wandel durch die Zeit. Im *DiCE* wird z.B. die Kollokation *morir de celos* ('vor Eifersucht sterben') als 'Magn+Oper$_1$' klassifiziert, im *DEC* hingegen ist *mourir de peur* ('vor Angst sterben') unter 'Degrad(*organisme*) - Sympt$_{23}$' verzeichnet, in beiden Kollokationen ist das Vorkommen des 3. Aktanten fakultativ. Es gibt im *DEC* und im *DiCouèbe* nur ein Gefühlssubstantiv, das in beiden Wörterbüchern verzeichnet ist, das Lemma *admiration.* Anhand eines Vergleiches der lexikalischen Funktionen und der ihnen zugeordneten Kollokate sind zahlreiche Unterschiede zwischen den Wörterbüchern festzustellen, von denen hier nur einige zitiert werden: *provoquer* und *susciter* stehen im *DEC* unter 'Caus$_{(2)}$Func' - im *DiCouèbe* unter 'IncepOper$_3$'; *frapper* wechselt von 'Magn+Labor$_{21}$' zu 'Magn.IncepLabor$_{31}$'; *saisir* wird im *DEC* unter 'IncepFunc$_1$' genannt, im *DiCouèbe* unter 'Magn.IncepLabor$_{31}$'. Die Einordnung der Kollokate als Werte verschiedener lexikalischer Funktionen in beiden Wörterbüchern ist nicht genau nachzuvollziehen, da detaillierte theoretische Informationen, wie sie im Vorwort des *DEC* gegeben werden, auf der Webseite des *DiCouèbe* fehlen. Auch der Grund für die Nennung zahlreicher Kollokate im *DiCouèbe*, die sich nicht im *DEC* befinden, und für die Streichung vieler Kollokate aus dem *DEC*, wird nicht angegeben.

   Die aufgezeigten Differenzen deuten darauf hin, dass das Format der lexikalischen Funktionen nicht nur in der Dechiffrierung dem ungeübten Leser Schwierigkeiten bereitet, sondern dass sich auch die Spezialisten in diesem Bereich bei der Zuordnung von Werten zu den Funktionen nicht immer einig sind. Zusätzlich besteht das Problem, dass einige der lexikalischen Funktionen das gleiche bedeuten, wie 'IncepPredMinus' und

'$FinFunc_0$' oder 'CausPredPlus' und '$CausFunc_0$', die beiden lexikalischen Funktionen werden jeweils auch im DEC als Äquivalent genannt.
Der ambigen und unpräzisen Natur einiger lexikalischer Funktionen widmet Fontenelle[99] ein ausführliches Kapitel und kommt zu folgendem Ergebnis: "Lexical functions are undoubtedly a very powerful descriptive device, but the small number of tests which can lead to their correct assignment clearly indicates that the theory in general needs more operational definitions than the somewhat fuzzy paraphrases we are used to finding in the MTT literature, especially if one wishes to employ the mechanism of lexical functions in an NLP perspective" (1997: 205).

Auch wenn die aufgeführten Punkte gegen eine Verwendung der lexikalischen Funktionen sprechen, bieten sie doch bis heute die einzige Möglichkeit die verschiedenen Aspekte, die die Beziehung zwischen Basis und Kollokat betreffen, in einem formalen System zu beschreiben. Eine Zuordnung der Kollokate zu den lexikalischen Funktionen zwingt den Lexikografen, die verschiedenen Arten der Kollokationsinformationen präzise zum Ausdruck zu bringen. Eventuell ist es nicht nur im Hinblick auf Wörterbuchbenutzer, sondern ebenso für die Weiterverarbeitung der Daten mit sprachverarbeitenden Systemen, sinnvoll, die verschiedenen linguistischen Ebenen, die die lexikalischen Funktionen bilden, wieder zu separieren - dadurch würde die Definition der Kollokationen weniger kryptisch erscheinen, die Überlappung bestimmter lexikalischer Funktionen ließe sich vermeiden, und sowohl für den Lexikografen als auch mit einem automatisierten Verfahren wäre die Bezeichnung der internen Kollokationsrelationen und externen Kollokationsinformationen leichter zu generieren. Voraussetzung hierfür wäre ein genau ausgearbeitetes formales System, das als Grundlage dient, und das für jede linguistische Ebene, die für die Kollokationsbeschreibung relevant ist, alle Parameter bietet, die notwendig sind, um die Kollokation exakt zu beschreiben. Genau zu bestimmen wäre die Semantik der Verben und der Substantive, die häufig in wechselseitiger Abhängigkeit steht, - durch eine präzise Bedeutungsangabe der einzelnen Komponenten ließe sich auch die interne Kollokations-relation definieren -, ebenso ist die Abbildung der Argumente auf die Aktanten relevant, sowie die konkreten syntaktischen Realisierungsmöglichkeiten im Satz, morphosyntaktische und pragmatische Präferenzen.
Unter dem Link, der jede Glosse und jede Kollokation mit weiteren Informationen verbindet, könnten sehr viel mehr Angaben enthalten sein, als die im folgenden Kapitel dargestellten. Hier wäre der geeignete Platz um numerische Daten bereitzustellen, wie beispielsweise Prozentangaben zum Artikelgebrauch und den Subkategorisierungsrahmen, oder Informationen, die die Abbildung der Argumente auf die Aktanten betreffen, sowie ausführliche semantische Definitionen und Angaben über die Gebräuchlichkeit des Verbs mit weiteren Substantiven. Lägen die Informationen in einem genau definierten Datenformat vor, könnten die Wörterbucheinträge in elektronischer Form als Grundlage für ein maschinelles Übersetzungssystem dienen, und außerdem dem Wörterbuchbenutzer, der sich dafür interessiert, zusätzliche Erklärungen bieten. Inwieweit eine automatische Generierung der Glossen und Zuordnung der Kollokate zu verwirklichen ist, wäre erst zu entscheiden, wenn es möglich ist, die vollständigen Angaben zu den internen Kollokationsrelationen und externen Kollokationsinformationen maschinell aus linguistisch aufbereiteten elektronischen Corpora zu extrahieren.

99 Fontenelle, Thierry (1997): *Turning a Bilingual Dictionary into a Lexical-Semantic Database*. Tübingen, Max Niemeyer.

Ein Problem bei den durch PECCI generierten Daten ist die Begrenzung der Kollokate auf 226 lemmatisierte Verben, sie spiegeln nur ein eingeschränktes Abbild der sprachlichen Realität wieder. Das Vorkommen bestimmter Glossen und deren Belegung durch eine bestimmte Anzahl von Verben sowie das Fehlen anderer Glossen ist daher nicht unbedingt repräsentativ, sondern geprägt durch die Selektion der Verben. Durch die Verwendung und Verarbeitung eines mit POS-Tags annotierten Corpus wären sehr viel bessere Ergebnisse zu erzielen. Da die lemmatisierten Verben zum Großteil aufgrund ihrer allgemeinen Gebräuchlichkeit mit den Gefühlssubstantiven gewählt wurden, sind vor allem restriktive Wortverbindungen stark unterrepräsentiert.
Für die Substantive *alegria, ciúme, esperança, inveja* und *susto* werden im folgenden Kapitel Wörterbuchartikel in dem beschriebenen ausführlichen Format gezeigt. Zusätzlich gibt es Wörterbucheinträge für *medo* und *ódio*, deren Kollokate sich in einem Vergleich gegenüberstehen. Der Sinn des Vergleichs ist es, auf einen Blick die kollokationalen Differenzen für die gleichen Glossen bei verschiedenen Substantiven zu zeigen. Ausgewählt wurden hier jeweils die ersten 30 Kollokate aus der nach log-likelihood sortierten Rankingliste der Singularform des jeweiligen Substantivs. Bei den durch den likelihood-Ratio Test ermittelten Werten wird die absolute Frequenz der Kookkurrenz im Vergleich zum t-Test weniger stark gewichtet, ein größeres Gewicht spielt bei der Berechnung hingegen die Häufigkeit der Kookkurrenz im Verhältnis zur Unabhängigkeits-annahme. Die Verwendung der Werte des likelihood-Ratio Tests ist zu empfehlen, wenn bei der lexikografischen Auswertung der Kookkurrenzdaten nur die Ergebnisse von einem statistischen Test benutzt werden (vgl. Kapitel 2.1). Die Kollokationsinformationen sind in den Einträgen von *medo* und *ódio* weniger ausführlich, lediglich das Vorkommen von Determinanten vor dem Gefühlssubstantiv, der Subkategorisierungsrahmen und eine mögliche deutsche Übersetzung für das Kollokat werden verzeichnet, sowie der Wert von log-likelihood und die Frequenz der Kookkurrenz der Singularform des Substantivs.

6.2. Wörterbucheinträge portugiesischer Gefühlssubstantive

alegria

Sg: 4775, Pl: 513

Freude, Fröhlichkeit, Heiterkeit (LS,PO)

X empfindet Freude (über Y)

	ter	[a ~$_{acc}$ (de + INF) | [uma ~$_{acc}$ Adj]	haben

2900845: A Itália vibrou com o seu « Giro » e teve a <alegria> de assistir ao nascer de uma nova grande estrela , o jovem Massimiliano Lelli , de 23 anos , que se guindou à terceira posição final , a 6m 56s do vencedor .
5308507: Feliz , sempre , o ministro tem uma <alegria> leve .

	sentir	[~$_{acc}$]	verspüren, empfinden

1349531: « Sinto uma grande <alegria> quando vejo que há certos conceitos introduzidos no discurso oficial .

	viver	[a ~$_{acc}$]	erleben
		[com ~]	leben mit

71895074: « É gratificante , quer dizer que , se eles estão a viver esta <alegria> , também serve para lhes dar força nas suas provas .
171728810: Penso que vivem com muita <alegria> , mas não com muito dinheiro .

+repetitiv	*recuperar*	[a ~ (de + INF)]	wiedererlangen

3795048: Pensou que no futebol inglês podia recuperar a <alegria> .

Y verursacht Freude (in X)

1 *dar* [uma | ∅ ~$_{acc}$ a N$_{dat}$] geben -> bereiten

10197544: « As crianças dão <alegria> e enriquecem a vida » , dizem os grandes anúncios afixados no metro de Berlim , no âmbito de uma campanha lançada pelo Ministério da Família e Seniores e destinada a combater o envelhecimento contínuo da população .

3 *fazer* [a ~N$_{acc}$ de N$_{dat}$] die Freude von jdm machen -> jdm Freude machen

376725: Esta iniciativa , que , « numa segunda fase , vai chegar também às escolas secundárias » , segundo Tó-Pê , animou também os jogadores que nesta manhã de quinta-feira fizeram a <alegria> dos pequenotes .

trazer [a | ∅ ~$_{acc}$ | ~s (a N$_{dat}$)] bringen

57262384: « Sem dúvida , o sudário traz-nos <alegria> .

transmitir [uma | ∅ ~$_{acc}$ a N$_{dat}$] übertragen

89306085: « Nos meus filmes , exprimo tanto aquilo que me preocupa como o que me traz felicidade , tentando transmitir a minha <alegria> aos espectadores . »

provocar [~$_{acc}$] erregen

43640235: Mas para o presidente do Benfica , outros números há que lhe provocam « grande <alegria> e optimismo quanto ao futuro » .

+intense *espalhar* [~$_{acc}$] verbreiten

168325139: Ela era muito alegre , cheia de vida , espalhava <alegria> e vida por onde quer que passasse .

causar [grande ~N$_{acc}$ a N$_{dat}$] verursachen

54391011: E o estudo do seu trabalho desenvolvido por Teresa Casendo , aluna da Universidade Nova e membro do Coro da Academia de Amadores de Música de Lisboa , tem causado grande <alegria> ao compositor .

X bringt die Freude zum Ausdruck

não esconder [a (sua) ~$_{acc}$] nicht verstecken

43539650: Marília Raimundo , não escondendo a sua <alegria> , chama o carro do som e agradece a recepção sempre « maior que a de 87 » , como afirma .

manifestar [a (sua) | ∅ ~$_{acc}$] zeigen, zum Ausdruck bringen

4699642: Na Cidade do Panamá , os automobilistas fizeram soar as buzinas em ar de celebração da pena anunciada e pequenos grupos juntaram-se para acenar lenços brancos e por outras formas manifestar a sua <alegria> .

exprimir [a (sua) | uma | ∅ ~$_{acc}$] ausdrücken

130385299: O centro de Copenhaga tornou-se um caos de cerveja e gente -- mas gente pacífica e que exprimia uma <alegria> sem limites .

mostrar [a (sua) | uma | ∅ ~$_{acc}$] zeigen

77665834: Confrontado com as críticas ou pura indiferença , Diepgen já tomou uma decisão : pelo menos a cinzenta cidade tem de mostrar <alegria> .

+antonym *disfarçar* [a ~$_{acc}$] verhehlen

137437692: Schwarz disfarçou a <alegria> .

expressar [a (sua) | ∅ ~$_{acc}$] ausdrücken

57382388: « Nós tentamos lutar sempre contra a apatia e o conformismo , criando algumas situações que expressem <alegria> , mas infelizmente estamos sozinhos neste esforço » , lamenta Natércia Pedroso .

demonstrar [a (sua) | ∅~$_{acc}$, ~ demonstrada] demonstrieren

10006405: Sousa Cintra teve ontem um desses dias , pelo menos a julgar pela <alegria> demonstrada ao longo da tarde .

testemunhar [a | uma ~$_{acc}$] bekunden

31179450: Os comentários recolhidos nesta zona abastada da capital testemunhavam uma grande <alegria> :

X bringt seine Freude körperlich zum Ausdruck

chorar [de ~] vor Freude weinen

36322078: « Até chorei de <alegria> » , disse Maria Rosa .

saltar [de ~] vor Freude hüpfen

2722223: Segundos depois , saltou de <alegria> ao saber que o jogo de Londres tinha acabado .

exultar [de ~] vor Freude jubeln

1544801: Mas houve casos em que tal não foi preciso e os socialistas exultaram de <alegria> .

Y erfüllt X mit Freude

encher [de ~ N_{acc}] mit Freude erfüllen

17011892: Dois golos de Hassan encheram de <alegria> o Estádio de São Luís ,

X verliert die Freude (an Y)

perder [a ~$_{acc}$ (de NS)] verlieren

10697649: Afinal , dizem , « a sua vida artística inspirava-se no optimismo e agora perderam a <alegria> de viver » .

X empfängt Y mit Freude

receber [N_{acc} com ~] (meist Passiv) empfangen mit Freude

119944136: A reacção do Presidente russo , Boris Ieltsin , não terá deixado de ser recebida com <alegria> pelos sérvios bósnios em Pale .

Y nimmt X die Freude

tirar [a ~ a N_{dat}] nehmen

32489924: Mas , para os estreantes da selecção nacional , não há frio nem cansaço que lhes tire a <alegria> .

Es gibt Freude

1 *haver* [~] es hat -> es gibt

2149041: Se a barriguinha não estiver cheia , garantia o António , « não há <alegria> que chegue » .

+antonym *faltar* [(a) ~] fehlen

3091264: Faltou a <alegria> do improviso característico daquelas gentes .

+begin *chegar* [(a) ~] aufkommen

78397291: A <alegria> chegou ao fim

X empfindet Freude (über Y)

ter	6.6	1.1	98	4.5	2.0	27
sentir	6.5	3.6	45	0.8	2.1	1
viver	4.8	2.8	27	1.6	2.8	3
recuperar	2.6	2.6	8			

$Oper_1$ empfinden, fühlen [~$_{acc}$]; haben [~$_{acc}$]

$Oper_1$ ter, sentir, viver
$Rep+Oper_1$ recuperar

Y verursacht Freude (in X)

dar	11.1	3.3	133	5.5	4.1	32
fazer	5.7	1.4	57	3.5	2.3	15
trazer	4.5	3.4	22	4.1	5.4	17
transmitir	3.0	3.1	10			
provocar	2.6	2.0	9	0.8	2.0	1
espalhar	2.1	3.3	5			
causar	2.0	2.4	5	0.9	3.0	1

(Für causar gibt es viel mehr Fundstellen. Weil die in dieser Kollokation am häufigsten verwendete Verbform, die 3.Pers.Sg. Präsens, homograph mit dem Substantiv *causa* ist, werden diese jedoch Fundstellen nicht extrahiert)

$Caus_2Func_1$ machen [N_{dat} ~$_{acc}$], hervorrufen [bei N_{dat} ~$_{acc}$], erregen [in N_{dat} ~$_{acc}$]

$Caus_2Func_1$ dar, transmitir, causar
$Caus_2Func$ provocar, espalhar
$Caus_2Func_{(1)}$ trazer
$Caus_2Oper_1$ fazer

dar

seine Freude haben an jm./etw. - alg vai/há-de dar alegrias a alg (IDP)

jdm eine Freude machen - dar uma alegria a alguém (PO)

causar

causar - *Freude* bereiten (LS)

X zeigt seine Freude

não esconder	5.5	4.3	31	0.9	3.1	1
manifestar	5.4	3.6	31			
exprimir	3.5	4.5	13			
mostrar	3.5	2.3	15			
disfarçar	3.1	4.8	10			
expressar	3.1	4.0	10			
demonstrar	2.4	2.7	7			
testemunhar	1.9	4.0	4	0.9	4.8	1

$Caus_1Manif$ não esconder, manifestar, exprimir, mostrar, disfarçar, expressar, demonstrar, testemunhar

X bringt die Freude zum Ausdruck

chorar	5.0	5.6	26
saltar	5.0	5.1	26
exultar	3.4	7.2	12

$Magn+Caus_1Manif$ chorar, saltar, exultar

chorar
chorar de ~ (DC)

saltar
saltar de ~ (DC)

Y erfüllt X mit Freude

encher	5.2	4.9	28

$Caus_2Oper_1$ versetzen [N_{acc} in ~$_{acc}$]

MagnCaus2Oper1 encher

X verliert die Freude (an Y)

perder	4.1	2.3	21	0.7	1.5	1

$FinOper_1$ perder

X empfängt Y mit Freude

receber	3.1	1.8	14

$Oper_2$ receber

Y nimmt X die Freude

tirar	2.2	2.7	6

$LiquFunc_2$ tirar

Es gibt Freude

haver	6.4	1.6	62	1.9	1.5	6
faltar	2.4	2.6	7			
chegar	0.7	0.3	5			

$IncepFunc_1$ aufkommen [in N_{dat}]

$Func_0$ haver

¬ $Func_0$ faltar

$MagnFunc_0$	espalhar
$IncepFunc_1$	chegar

Anmerkungen

Für folgende lexikalische Funktionen befinden sich unter den lemmatisierten portugiesischen Verben keine Kollokate:

IncepPredMinus	nachlassen
Magn + $IncepOper_1$	geraten [in ~$_{acc}$]
$FinFunc_0$	sich legen
fast $FinFunc_0$	verfliegen
$Liqu_1Func_0$	überwinden [$PRON_{poss}$ / DET ~$_{acc}$]
Magn + $IncepFunc_1$	erfassen [N_{acc}]
Magn + fast $IncepFunc_1$	packen [N_{acc}]
$Liqu_1Fact_0$	unterdrücken [$PRON_{poss}$ / DET ~$_{acc}$]
Magn + $IncepFact_1$	überkommen, überwältigen [N_{acc}]

Für die folgenden Einträge der untersuchten Wörterbüchern gibt es keine Fundstellen im Corpus:

dar gritos de DC
jauchzen/(jubeln) vor Freude *path* dar gritos/pulos de alegria (IDP)
fast außer sich geraten vor Freude - ficar doido de alegria *fam* (IDP)
jm. in Freud' und Leid zur Seite stehen *path* - estar/ficar ao lado de alg na alegria e na tristeza (IDP)
(Das Substantiv *gritos* (Schreie) ist in der Exzerptionsdateien von *der alegria/alegrias* nicht in der Wortkombination *dar gritos de alegria* zu finden, auch *doido* und *ao lado* befinden sich in der Exzerptionsdatei von *ficar/estar* und *alegria/alegrias* nicht in der angegeben Wortkombination.)

Die Kollokate folgender Einträge der untersuchten Wörterbüchern sind nicht lemmatisiert:
Freud' und Leid mit jm./miteinander teilen *path* - partilhar alegrias e tristezas com alg (IDP)

ciúme(s)

Sg: 308, Pl: 437

Eifersucht (LS,PO)

X empfindet Eifersucht (auf Y)

2 *ter* [~s$_{acc}$ (de N_{dat})] haben -> eifersüchtig sein (auf)
115655811: -- A minha mulher tinha <ciúmes> , ciúmes doentios .

sentir [~$_{acc}$ (de N_{dat})] verspüren
145678480: Neil , que sentia um crescente <ciúme> de Lionel , podia ou não ter evitado o acidente ?

+intense
+continue 1 *roer* [ficar roído de ~s] zernagt bleiben vor -> platzen vor
25135091: d) Fica roída de <ciúmes> ?

Y verursacht Eifersucht (in X)

provocar [o ~$_{acc}$ (de N_{dat})], [~s$_{acc}$ (a | em | entre N_{dat})] provozieren
14188796: Como não podiam dar os bom-bons a toda a população carenciada , optaram naturalmente por contemplar as « somalis mais giras » , o que provocou o <ciúme> dos homens , explicou o militar francês .
162499443: Para provocar mais <ciúmes> a Francisquita , bem entendido .

motivar [motivado por ~s$_{acc}$] motiviert durch
9189298: O autor do crime , de aparência jovem , terá agido motivado por <ciúmes> , uma vez que não houve roubo nem de dinheiro nem do veículo do oficial assassinado .

causar [~s$_{acc}$ (a N_{dat})] verursachen
52574924: Aceitar significa causar <ciúmes> a Santos , que enviará um guarda para o pôr a dormir .

suscitar [~s$_{acc}$ (de N_{dat})] erregen
83477809: A própria CIP tentou dar o braço à CGTP para se chegar a um acordo bilateral , o que -- segundo Nogueira Simões e a CGTP -- suscitou muitos « <ciúmes> » , « bocas à parte » por parte da CCP e da UGT e « algum nervosismo » do Governo .

desencadear [~s_{acc} (de N_{dat})] auslösen

165400075: E continua a insistir-se nas relações « amigais » entre ela e o vizinho e que teriam desencadeado os <ciúmes> do filho .

2 *fazer* [~s_{acc} a N_{dat}] machen -> jdn eifersüchtig machen

3194791: Se era apenas para fazer <ciúmes> a Santana Lopes , presidente do Sporting -- que deu nove dos 11 titulares à selecção de sub-21 -- , o risco pode ser demasiado : vamos ver se , na retribuição , Madaíl não será convidado para almoçar no Kentucky Fried Chicken de Palermo .

X handelt aus Eifersucht

matar [a N_{dat} por ~s] jdn aus Eifersucht töten

116176996: Um caso passado no distrito de Coimbra há alguns anos -- um indivíduo proxeneta chamado Maximino matou por <ciúmes> uma mulher e cortou-a aos bocados , metendo-a em malas e atirando-a para o rio Mondego .

Es gibt Eifersucht

1 *haver* [~] es hat -> es gibt

139566173: Que não seja necessário (...) Portugal e Angola amam-se , mas não haja <ciúme> e não haja negócio , haja respeito mútuo .

X empfindet Eifersucht (auf Y)

ter	1.3	0.8	5	7.0	2.9	55
sentir	1.6	3.7	3			
roer				1.4	7.3	2

$Oper_1$ empfinden, fühlen [~$_{acc}$];

$Oper_1$ ter, sentir
$Magn+ContReal_1$ ficar roído de

ter

ter ~/~s (de, por alg.) (DC)
ter ~s de - eifersüchtig sein auf (LS, PO)

sentir

sentir ~s: *Ela não pode deixar de sentir ~s sempre que é o seu marido com outra mulher.* (DC)

* Im Corpus wird sentir innerhalb des Suchraums von ciúmes nicht gefunden, aber es gibt zwei Vorkommen außerhalb des Suchraums.

Y verursacht Eifersucht (in X)

provocar	2.2	4.2	5	**4.1**	**5.0**	**17**
motivar				2.6	5.8	7
causar				2.2	4.8	5
suscitar				1.4	4.4	2
desencadear				1.4	4.6	2
fazer				1.9	1.5	6

$Caus_2Func_1$ *wecken [in N_{dat} ~$_{acc}$],* erregen [in N_{dat} ~$_{acc}$]

$CausFunc_1$ motivado por
$Caus_{(2)}Func_1$ provocar, causar, suscitar, desencadear
$Caus_2Func_1$ fazer

X handelt aus Eifersucht

matar				1.7	4.4	3

$Magn+Labreal_{(1/3)}$ matar

Es gibt Eifersucht

haver	2.1	2.0	6	2.0	1.7	6

$Func_0$	haver

Anmerkungen

Für folgende lexikalische Funktionen befinden sich unter den lemmatisierten portugiesischen Verben keine Kollokate:

IncepPredMinus	nachlassen
$CausContFunc_1$	*schüren [in N_{dat} ~$_{acc}$]*
$FinFunc_0$	sich legen
$fastFinFunc_0$	verfliegen
$Liqu_1Func_0$	überwinden [$PRON_{poss}$ / DET ~$_{acc}$]
$IncepFunc_1$	aufkommen [in N_{dat}]
$Magn+IncepFunc_1$	erfassen [N_{acc}]
$Magn+fastIncepFunc_1$	packen [N_{acc}]
$Liqu_1Fact_0$	unterdrücken [$PRON_{poss}$ / DET ~$_{acc}$]
$Magn+IncepFact_1$	überkommen [N_{acc}]

esperança (em)

Sg: 11113, Pl: 4900

Hoffnung, Erwartung (LS), Hoffnung (*auf* em) (PO)

X hat/empfindet Hoffnung (auf Y)

	ter	[(a) ~$_{acc}$((de)(que) +NS, em N)]	haben

82474915: R.-- Férias , Verão , tenho sempre a <esperança> que este país continue a ser o que é porque assim há sempre temas .

83533808: Para ter <esperança> numa resposta de esquerda , do lado do meu coração .

+continue	*manter*	[a ~$_{acc}$((de)(que) +NS, em N)]	behalten

9876265: Apesar da negativa ministerial , a Resistência considera ter razões para manter a <esperança> de que Jacarta acabe por ceder .

+continue	*ficar*	[com a ~ ((de)(que) +NS, em N)]	bleiben mit

49049265: E , até ao fim , fiquei com a <esperança> de poder sair de Lisboa , para ir presidir a esse Conselho .

	viver	[na ~$_{acc}$((de)(que) +NS)], [sem/com ~]	leben in/ohne/mit

40532840: Vivia na <esperança> de atingir o meu objectivo e , ao mesmo tempo , no receio de não o conseguir .

62920804: « Vocês sabem lá como é triste viver sem <esperança> » , cantava , convidando com o gesto ao aplauso .

+continue	*continuar*	[na/com a ~ ((de)(que) +NS, em N)]	behalten

141218578: « Continuo com a <esperança> de que o atleta continue no Sporting » , disse .

+continue	*conservar*	[a ~$_{acc}$((de)(que) +NS, em N)]	bewahren

143337758: Muitos ainda conservam a <esperança> de regressar às suas casas .

+repetitiv	*recuperar*	[(a) ~$_{acc}$((de)(que) +NS)]	wieder erlangen

102170448: Quem recuperou a <esperança> foram os adeptos do Manchester United , que já fizeram um apelo público a Ince para que fique .

+continue	*guardar*	[a ~$_{acc}$((de)(que) +NS, em N)]	bewahren

101037328: As que não têm notícias guardam a <esperança> de que os seus filhos estejam « nas mãos do inimigo » .

	estar 3	[com ~ ((de)(que) +NS)]	mit Hoffnung sein -> Hoffnung haben
		[na ~ (de)(que) +NS]	in der Hoffnung sein, dass

99676466: Mas estou com <esperança> .

104061083: Estava na <esperança> que pudesse haver mais público , havia jogos de bom nível ,

	sentir	[(DET) ~$_{acc}$((de) (que) +NS, em N)]	fühlen

142276321: Na final tudo pode acontecer e , neste momento , sinto uma <esperança> maior , porque ele tem de entender que é candidato a uma medalha » , acentuou Ferreira da Silva .

+begin 1 *ganhar* [~s (em N)] gewinnen -> schöpfen
84605843: Ganharam novas <esperanças> quando , em Junho deste ano , ouviram Marçal Grilo dizer na Assembleia da República que o problema estaria resolvido até 1 de Setembro .

Z verursacht Hoffnung (in X, auf Y)

+continue *alimentar* [a~$_{acc}$ | ~s$_{acc}$ (de+NS, a/em N)] nähren
11319483: Mas não vamos alimentar grandes <esperanças> » .

+continue 1 *acalentar* [a~$_{acc}$ | ~s$_{ac}$ (de+NS, em N)] erwärmen -> hegen
1055044: Passou à meia-maratona com 1h13m17s e , nessa altura , acalentava <esperanças> de bater o seu recorde pessoal (2h28m11s , de 1989) .

1 *dar* [(DET) ~$_{acc}$ (a N$_{dat}$)] geben -> machen, geben
7273481: Recorrendo a uma imagem metafórica da Associação da Cova da Moura , é necessário dar <esperança> às « viúvas dos vivos » .
30l4412: Esta situação dá a <esperança> de uma melhoria dos órgãos atingidos que são assistidos por aparelhos . »

+begin *trazer* [(DET) ~$_{acc}$ (a/de N$_{dat}$)] bringen
45560904: O referendo é um voto que realmente conta e traz mudanças , não como o voto eleitoral , que traz uma <esperança> de mudança .

deixar [a alg. poucas/muitas ~s ((de)(que) +NS, em N)] lassen
9052255: Já soma sete derrotas e os três pontos de diferença que vai ter relativamente ao Benfica não deixam muitas <esperanças> aos portistas de se sagrarem campeões nacionais .

+begin *suscitar* [~s$_{acc}$ ((de)(que) +NS, em N)] wecken
44398677: Primeiro , o Governo está a pagar as promessas da campanha presidencial , que tinham suscitado <esperanças> mesmo nos sindicatos como a Força Operária .

+begin 1 *criar* [~s$_{acc}$ ((de)(que) +NS, em N)] erzeugen, züchten -> erwecken
28345436: A associação de residentes de Garvaghy Road disse ontem que as negociações de Trimble e Mallon criaram « falsas <esperanças> » .

+repetitiv *restaurar* [(a) ~$_{acc}$ (de/em N)] wieder herstellen
8313035: Num país como Angola , afundado na ruína e na miséria , essa ainda é uma das poucas imagens capazes de restaurar a <esperança> .

+continue *nutrir* [~s$_{acc}$ ((de)(que) +NS, em N)] nähren
112455055: É verdade que o clube francês e a imprensa em geral nutriam grandes <esperanças> para a partida .

+begin *levar* [(DET) ~$_{acc}$ (a N$_{dat}$)] bringen
144830355: 5 . Continua a representar para mim esse marco assinalável , que conseguiu mudar o nosso tipo de relações com povos por nós colonizados , por vezes de forma quase escravizante , e levar uma <esperança> de libertação a muitos povos da Terra .

gerar [~$_{acc}$ ((de)(que) +NS, em N)] erzeugen
21002817: A <esperança> gerada na altura das eleições esfumou-se .

+begin *despertar* [~$_{acc}$ ((de)(que) +NS, em N)] wecken
46629656: « E Rocha Vieira , que tanta <esperança> despertou por ser diferente dos outros ? »

+continue *nutrir-se* [a ~$_{acc}$ (de+NS)] nähren
21418306: Para o de Torres Vedras , iniciado em Março , nutre-se a <esperança> que fique concluído este mês .

+begin *semear* [a ~$_{acc}$ (de/em N)] säen
59194058: Penso que é possível dar os factos com rigor , objectividade e veracidade , mas simultaneamente de forma atraente , divertida , alegre e semeando <esperança> .

infundir [~$_{acc}$ a/em N] einflößen
14484702: Se , ao entrar num governo chefiado por Netanyahu , ele puder mudar significativamente a política de Netanyahu ; se , ao fazer isso , ele puder infundir nova <esperança> ao lado palestiniano -- então que se danem tais considerações .

1 *lançar* [alguma ~$_{acc}$ (em N)] werfen, schleudern -> einflößen
18450843: O aparecimento deste novo teste , ainda que necessitando de estudos mais completos , vem lançar alguma <esperança> na comunidade médica .

+continue 1 *fomentar* [a ~$_{acc}$ (de+NS)] ankurbeln, fördern -> schüren
123516361: ... , implicando-o numa teia de aventuras sem sucesso , fomentando a <esperança> , infundada , de poder um dia trazer a viagem a bom porto .

inspirar [~s_{acc}] einflößen

25910630: Todavia , a actual situação económica e social do distrito não inspira grandes <esperanças> , a acreditar num documento revelado pelo PUBLICO na última quinta-feira , no qual a USP expressa as suas reticências quanto ao futuro .

X äußert Hoffnung (auf Y)

manifestar [a ~$_{acc}$ | ~s_{acc} ((de)(que) +NS, em N)] zeigen, zum Ausdruck bringen

4044243: Franco Baresi , futebolista italiano do AC Milão , manifestou a <esperança> de vir a jogar na Liga Japonesa na próxima temporada , conforme noticia a imprensa nipónica de ontem .

exprimir [a ~$_{acc}$ ((de)(que) +NS, em N)] ausdrücken

expressar [a ~$_{acc}$ ((de)(que) +NS, em N)] ausdrücken

+antonym *esconder* [a ~$_{acc}$ ((de)(que) +NS, em N)] verstecken

confessar [a ~$_{acc}$ ((de)(que) +NS, em N)] zugeben

mostrar [a ~$_{acc}$ ((de)(que) +NS, em N)] zeigen

revelar [a ~$_{acc}$ ((de)(que) +NS, em N)] offenbaren

proclamar [a ~$_{acc}$ ((de)(que) +NS, em N)] verkünden

X setzt seine Hoffnung auf Y

1 *depositar* [~s_{acc} em N)] einzahlen / abgeben -> setzen auf

172246068: Queriam ver o Presidente , um homem de quem gostam , em quem depositam tantas <esperanças> para que a Bulgária ferida volte de novo a exibir o seu orgulho .

colocar [~s_{acc} em N] setzen auf

11145708: Dos seis candidatos anunciados à compra da companhia , que se encontra em situação de falência técnica , resta agora apenas um , no qual o ISP parece colocar todas as <esperanças> .

pôr [~s_{acc} em N] setzen auf

85433346: Às vezes , quase que percebo porque é que as pessoas põem as suas <esperanças> no futebol .

X verliert die Hoffnung (auf Y)

perder [a ~$_{acc}$ (de+INF, em N)] verlieren

588545: Mas pode estar aqui a resposta para aqueles que já perderam a <esperança> de ser papás .

1 *abandonar* [a ~$_{acc}$ ((de)(que) +NS)] verlassen -> aufgeben

35163796: Os alunos que vão recorrer a esta segunda hipótese são alunos que apenas pretendem concluir o ensino secundário e outros que já abandonaram a <esperança> de entrar no ensino superior no contingente normal .

+intense *enterrar* [as ~$_{acc}$ ((de)(que) +NS)] begraben

9334540: Se o país tivesse caucionado a sua deserção , teria igualmente enterrado qualquer <esperança> de regeneração do poder político por iniciativa da sociedade civil .

Y repräsentiert die Hoffnung von X

representar [~$_{acc}$ de, para N] repräsentieren

4331640: Tínhamos entrado num período em que o design representava uma <esperança> para os artistas .

Z vergrößert die Hoffnung (in X,auf Y)

aumentar [a ~$_{acc}$ ((de)(que) +NS, em N)] vergrößern

9810435: « Mas só no Verão passado é que as escavações viriam a confirmar a existência de um recinto megalítico e aumentar as <esperanças> de termos descoberto um observatório de astros » , contava Eduardo Silva .

*Sg, 12 mal aumentar a esperança de vida - 'Lebenserwartung erhöhen'

Z nimmt (X) die Hoffnung (auf Y)

1 *retirar* [a ~$_{acc}$ ((a N) (de+INF, de/em N)] zurücknehmen -> nehmen

73566948: Resta apenas a ilusão , vital para Bertolucci :« Viver uma ilusão é muito importante porque , tal como a utopia , a ilusão contém uma grande esperança e se me retirarem a <esperança> não serei capaz de fazer mais filmes » .

desfazer [as ~s_{acc} (de N)] zunichte machen

46477879: A partir daí , Edberg pôs ainda mais pressão no saque adversário , chegando rapidamente a 4-0 e desfazendo todas as <esperanças> de Sampras .

+intense *matar* [as ~s_{acc} (que+NS, de N)] töten

61967772: Ontem , « matou » as <esperanças> dos portugueses com aquele golo magnífico a abrir a segunda parte .

destruir [as ~s_{acc} (de +INF, de/em N)] zerstören

143860840: « O que se passa -- acrescentou -- é que Milosevic , com a sua política de ziguezague , a única coisa que conseguiu foi destruir todas as <esperanças> do povo sérvio de manter uma Jugoslávia unida ao Governo de Belgrado .

dissipar [as ~s ((de) que + NS, de N)] zerstreuen

56675950: Num pequeno e superpovoado país do interior de África , o Ruanda , falhou um novo cessar-fogo , voltando a dissipar as <esperanças> de uma solução política para uma guerra iniciada em Outubro de 1990 e que já vitimou alguns milhares de pessoas e provocou um milhão de refugiados .

1 *afastar* [as ~s_{acc} ((de)(que)+NS, de/em N)] entfernen -> ersticken

91039907: A falta de experiência da equipa francesa nesta prova é tal que o próprio Jean Todt já afastou quaisquer <esperanças> na vitória .

tirar [a ~$_{acc}$ (a N)] nehmen

157507687: A Eulália não deixou que o desemprego aos 34 anos lhe tirasse a <esperança> , e voltou à escola .

Y erfüllt X mit Hoffnung

encher [de ~$_{acc}$ a N] erfüllen

43213632: A ocasião encheu de <esperança> a comunidade internacional .

Y erfüllt die Hoffnungen von X

cumprir [as ~$_{acc}$ (de N)] erfüllen

79043593: « Portugal talvez tenha encontrado no actual primeiro-ministro um líder capaz de cumprir as <esperanças> constantemente defraudadas dos nossos idealistas . »

Es gibt Hoffnung

1 *haver* [~ ((de)(que) +NS, de, para, em N)] es hat -> es gibt

3307742: Portanto , há <esperança> e há riscos .

existir [~ ((de)(que) +NS, em N)] existieren

17950154: Mas para os adeptos ainda existe uma ténue <esperança> de que a equipa consiga evitar a descida .

1 *estar* [a ~ estar em N] sein in -> liegen auf

83891740: A sua última <esperança> está agora na indústria .

die Hoffnung bleibt

restar [a ~$_{acc}$ | ~s_{acc} ((de)(que) +NS, de N)] bleiben

2463084: Resta então a <esperança> da avaria no automóvel a algum viajante que esteja em circuito pela França .

permanecer [(em N)] bleiben

39312737: Mas , apesar da tensão , a <esperança> permanecia e levava uma jovem a afirmar :

ficar [a ~$_{acc}$ ((de)(que) +NS, em N)] bleiben

42877480: Fica a <esperança> , até porque seria mau para o negócio que assim não fosse , de que não acabem , também , as viagens aos estúdios .

Hoffnung kommt wieder

renascer [a ~ ((de)(que) +NS, de/em N)] wieder aufleben

3579601: E agora renasce a <esperança> da população de flamingos subir para os dois milhões , como na década de 60 .

Hoffnung beginnt

1 *nascer* [~] geboren werden -> aufkeimen

1996842: Conjugando-se com o aumento das vagas nasce uma ténue <esperança> de que seja mais fácil entrar em Medicina .

surgir [~ (entre/para N)] sprießen

127680994: Pouco tempo depois da queda do Muro , veio a União Monetária e , com a reunificação dos dois Estados alemães , surge uma nova <esperança> .

die Hoffnung wächst (in Y,auf Z)

crescer [a ~$_{acc}$ ((de)(que) +NS, de N)] wachsen

81268570: Mas se os balladurianos concentram agora a « artilharia pesada » em Chirac é porque no campo do presidente da Câmara de Paris cresce a <esperança> de uma segunda volta entre Chirac e Jospin .

die Hoffnung nimmt ab (in Y,auf Z)

diminuir [a ~$_{acc}$ ((de)(que) +NS, em N)] nachlassen

72344656: Foram entretanto recuperados partes de corpos , documentos e bagagens dos passageiros mas diminuem as <esperanças> de se encontrarem intactos a maioria dos corpos das vítimas do acidente .

die Hoffnung schwindet

+intense *morrer* [a ~ (em N)] sterben

118691612: Todos os meses vai lá , ao sítio onde a <esperança> morre lentamente : ontem à Croácia , amanhã à Bósnia , a Gaza um outro dia .

desvanecer-se [as ~s ((de) que + NS, de N)] sich zerstreuen

13266207: No entanto , os índices económicos actuais ainda parecem ser piores do que o eram então , tendo-se desvanecido todas as <esperanças> que existiam no início da década passada quanto às grandes potencialidades da Nigéria , que não as soube aproveitar em benefício da generalidade dos seus cidadãos .

+fast *evaporar-se* [as ~s ((de) que + NS, de N)] verfliegen

24594603: Gorbatchov já viu evaporarem-se as <esperanças> de que o tratado possa ser assinado a tempo de coincidir com a visita que este mês fará a Londres , onde tentará persuadir os líderes do Grupo dos Sete países capitalistas mais industrializados de que a URSS é um parceiro económico estável .

X hat/empfindet Hoffnung (in Y, auf Z)

ter	26.9	2.4	870	19.1	2.5	429	
manter	10.1	3.4	110	7.6	3.6	61	
ficar	5.6	1.5	50*	1.4	0.6	9*	
viver	5.3	2.2	36	0.7	0.5	3	
continuar	4.1	1.3	30*	2.8	1.4	14*	
conservar	3.0	3.7	10	1.6	3.4	3	
recuperar	2.7	2.0	10	0.9	1.2	2	
guardar	2.6	2.8	8	2.5	3.5	7	
estar	2.1	0.3	62*	2.0	0.4	31*	
sentir	1.0	0.6	5				
ganhar	-0.4	-0.2	3	1.4	1.1	5	

$Oper_1$ hegen, empfinden, ?fühlen [~$_{acc}$]; haben [~$_{acc}$]

$IncepOper_1$ bekommen [~$_{acc}$]

$Oper_1$ ter, viver em, sentir, estar com/em

$IncepOper_1$ ganhar

$ContOper_1$ manter, ficar, continuar, conservar , guardar

$Rep+Oper_1$ recuperar

ter

ter ~: *Apesar da gravidade da doença continua a ter ~.* (DC)

sich (keine) Hoffnungen machen - (não) ter ~(s) de (IDP)

esperançar-se em - Hoffnungen haben auf (ac) (LS)

sich Hoffnungen machen auf - ter ~s em (LS)

Hoffnung haben - ter ~ (PO)

estar

sich (keine) Hoffnungen machen - (não) estar com esperança(s) de (IDP)

ganhar

schöpfen - (*Hoffnung*) ganhar (PO)

Y verursacht Hoffnung (in X, auf Z/dass Z)

alimentar	12.9	5.4	169	13.4	6.3	181
acalentar	8.3	7.5	70	10.9	8.9	119
dar	**9.6**	**2.3**	**114**	**7.9**	**2.7**	**72**
trazer	6.2	3.2	42	4.5	3.4	22
deixar	5.0	1.6	39	5.7	2.4	40
suscitar	1.9	2.1	5	4.8	4.5	24
criar	2.7	1.2	14	4.2	2.5	21
restaurar	3.8	4.6	15			
nutrir				2.8	6.2	8
levar	2.4	0.9	16	0.1	0.1	3
encher	2.2	2.5	6	2.1	3.2	5
gerar	2.1	2.0	6	1.1	1.7	2
despertar	2.1	2.8	5	1.9	3.4	4
nutrir-se	1.7	4.4	3			
semear	1.6	3.5	3	0.9	3.2	1
infundir	1.4	6.5	2	1.0	6.6	1
lançar	1.4	0.7	7	1.2	1.0	4
fomentar	0.8	1.9	1			
inspirar				0.8	1.6	1

$Caus_{(2)}Func_1$ *machen [N_{dat} ~$_{acc}$], wecken [in N_{dat} ~$_{acc}$]*
$Caus_2Func_1$ *einflößen [N_{dat} ~$_{acc}$]*

$IncepCaus_2Func_{(1)}$	trazer, suscitar, criar, levar, despertar, semear
$Caus_2Func_{(1)}$	dar, deixar, lançar, gerar, inspirar
$ContCaus_2Func_{(1)}$	alimentar, acalentar, nutrir, nutrir-se, fomentar
$Caus_2Func_1$	infundir
$Rep+Caus_2Func_{(1)}$	restaurar

alimentar
alimentar a ~ (DC),
jm Hoffnungen machen - alimentar ~s a alg (IDP)

acalentar
acalentar a ~ (DC)
acalentar - *Hoffnung* hegen (LS)

dar
jm Hoffnungen machen - dar ~s a alg (IDP)
Hoffnung machen - dar ~ (LS)
esperançar - Hoffnung machen (LS)
dar (falsas) ~s a alguém - jdm (falsche) Hoffnungen machen (PO)

nutrir
jm Hoffnungen machen - nutrir ~s a alg (IDP)
nutrir - *Hoffnung* hegen (LS)
nutrir ~s - Hoffnungen nähren (LS)
hegen Hoffnung ~ nutrir esperanças (PO)

fomentar
fomentar a ~ (DC)

X äußert die Hoffnung (auf Y)

manifestar	20.1	5.3	411	6.2	3.9	41
exprimir	6.7	4.9	46	1.9	3.3	4
expressar	4.1	3.7	18	2.1	3.2	5
esconder	3.2	2.5	12	0.6	0.9	1

confessar	2.7	2.4	9			
mostrar	2.3	1.2	11	0.3	0.3	2
revelar	2.3	1.1	11	0.8	0.7	3
proclamar	2.1	3.0	5			

$Caus_1Manif$ manifestar, exprimir, expressar, confessar, mostrar, revelar, proclamar
$Liqu_1Manif$ esconder

X setzt seine Hoffnung auf Y

depositar	8.2	5.3	68	18.2	7.7	334
colocar	0.2	0.1	4	4.0	2.5	19
pôr	2.1	1.2	9	3.7	2.6	16

colocar
seine (ganzen) Hoffnungen auf jn./etw. setzen - *colocar (todas) as suas esperanças em alg/qc* (IDP)

pôr
seine (ganzen) Hoffnungen auf jn./etw. setzen - *pôr (todas) as suas esperanças em alg/qc* (IDP)

X verliert die Hoffnung (auf Y)

perder	14.6	3.8	223	12.0	4.3	150
abandonar	2.5	1.7	10	2.7	2.4	9
enterrar	1.2	1.9	2	2.9	4.2	9

$FinOper_1$ perder, abandonar
$FinReal_1$ enterrar

perder
die Hoffnung aufgeben - perder a ~ (DC, LS, PO)

enterrar
(seine) Hoffnung zu Grabe tragen - *enterrar as (suas) esperanças* (IDP)

abandonar
(seine) Hoffnung zu Grabe tragen - *abandonar as (suas) esperanças* (IDP)

Y repräsentiert die Hoffnung von X

representar	6.1	2.7	43	2.3	1.8	8

$Caus_2Manif$ representar

Z vergrößert die Hoffnung (in X,auf Y)

aumentar	4.6	2.4	26	4.5	3.1	23

CausPredPlus aumentar

Z nimmt X die Hoffnung (auf Y)

retirar	2.8	2.1	10	2.8	2.8	9
desfazer	1.6	2.6	3	3..9	5.1	16
matar	1.8	1.7	5	3.5	3.5	13
destruir	1.2	1.2	3	3.2	3.4	11
dissipar				2.2	5.1	5
afastar	1.2	1.0	4	1.9	2.0	5
tirar	1.8	1.6	5	1.8	2.2	4

$Liqu_{(2)}Func_1$ retirar, desfazer, matar, destruir, dissipar, afastar, tirar

destruir
jm. die Hoffnung nehmen - *destruir a esperança a alg.* (IDP)

tirar
jm. die Hoffnung nehmen - *tirar a esperança a alg.* (IDP)

Y erfüllt X mit Hoffnung

encher	2.2	2.5	6	2.1	3.2	5

$Caus_2Oper_1$ encher

Y erfüllt die Hoffnungen von X

cumprir	0.4	0.3	3	1.8	1.6	5

$Real_2$ cumprir

Es gibt Hoffnung

haver	16.6	2.5	328	7.3	1.8	76
existir	7.3	2.5	63	4.4	2.4	24
estar	2.1	0.3	62*	2.0	0.4	31*

$Func_0$ haver, existir
$Func_1$ estar com/em

die Hoffnung bleibt

restar	12.1	5.0	149	3.1	3.2	11
permanecer	2.0	1.8	6	1.1	1.5	2
ficar	5.6	1.5	50*	1.4	0.6	9*

$ContFunc_{(1)}$ restar, permanecer, ficar

Hoffnung beginnt

renascer	6.3	6.1	41	3.8	5.9	15
nascer	3.9	2.4	19	2.3	2.2	7
surgir	3.6	1.8	19	1.9	1.4	6

$IncepFunc_{(1)}$ aufkommen [in N_{dat}]

Rep+IncepFunc renascer
$IncepFunc_0$ nascer
$IncepFunc_{(1)}$ surgir

die Hoffnung wächst (inY,auf Z)

crescer	3.1	2.3	12	3.1	3.0	11

die Hoffnung nimmt ab (inY,auf Z)

diminuir	2.1	2.1	6	3.2	3.5	11

IncepPredMinus *nachlassen*
IncepPredMinus diminuir

die Hoffnung schwindet

morrer	3.0	1.8	13	2.0	1.9	6
desvanecer-se	0.9	3.3	1	2.6	6.1	7
evaporar-se				1.4	5.7	2

$fastFinFunc_0$ verfliegen

$FinFunc_0$ desvanecer-se
$FinFunc_{0[0/1)}$ morrer [a ~ (em N)]
$fastFinFunc_0$ evaporar-se

desvanecer
desvanecer - *Hoffnung* auslöschen; zunichte machen, ~-se verfliegen (LS)

Anmerkungen

Für folgende lexikalische Funktionen befinden sich unter den lemmatisierten portugiesischen Verben keine Kollokate:

$Liqu_1Fact_0$ unterdrücken [$PRON_{poss}$ / DET ~$_{acc}$]

Für die folgenden Einträge der untersuchten Wörterbüchern gibt es keine Fundstellen im Corpus:

in die Hoffnung kommen *form veraltend selten* - ficar de esperança *pop* (IDP)
estar de esperanças - guter Hoffnung sein (LS)

Die Kollokate folgender Einträge der untersuchten Wörterbüchern sind nicht lemmatisiert:

sich (da) keine falschen Hoffnungen machen - não tecer esperanças vãs (IDP)
js. Hoffnungen erfüllen/erfüllen sich - as esperanças de alg concretizam-se (IDP)
erfüllen - *Hoffnung* concretizar-se (LS)
(wieder) (neue) Hoffnung schöpfen - (voltar a) encontrar (nova) esperança (IDP)

inveja

Sg: 1179, Pl: 159

Neid (LS,PO)

X empfindet Neid (auf Y)

2 *ter* [~$_{acc}$ (de N_{dat})] haben -> neidisch sein (auf)
170967669: Não vejo , portanto , por que razão Pinto da Costa pode ter <inveja> .

+intense 1 *roer-se* [de ~$_{dat}$] sich zernagen vor -> platzen vor
6017970: Roa-se de <inveja> : o consumo anda à volta dos três decilitros aos ...

olhar [com ~$_{dat}$ $N_{acc)}$] schauen mit
146932716: Desde 1990 que a população da Culatra olhava com <inveja> a vizinha ilha da Armona .

+continue 2 *ficar* [com ~ (de N_{dat})] bleiben mit -> weiterhin neidisch sein (auf)
158027950: Felizmente , os clientes do Windows 3.1 não precisam ficar com <inveja> .

+intense 1 *morrer* [de ~$_{dat}$] sterben vor -> vergehen vor
29769915: As outras comunas morrem de <inveja> .

sentir [~$_{acc}$ de N_{dat}] empfinden
408871: E , como tantas vezes aconteceu nos seus anos da ressurreição , até os deuses sentiram <inveja> dos homens .

Y verursacht Neid (in X)

2 *fazer* [~$_{acc}$ (a N_{dat})] Neid machen -> jdn neidisch machen
11069173: Um dinâmico grupo de pequenas e médias empresas revelam uma capacidade de resistência de fazer <inveja> aos grandes grupos .

1,2 *fazer corar* [fazer N_{akk} corar de ~$_{dat}$]
jdn rot vor Neid machen -> jdn vor Neid erblassen lassen
13702487: Efeitos de luz , neblinas matinais , um detalhe no solo que faz corar de <inveja> alguns simuladores com placas 3D , está lá tudo .

1,3 *causar* [~$_{acc}$ (a N_{dat})] jdn ~ verursachen -> den Neid erregen von
76434745: Os jornais descobrem todos os dias esquemas de corrupção dentro do Governo capazes de causar <inveja> à máfia italiana .

2 *fazer ficar verde* [fazer N_{akk} ficar verde de ~$_{dat}$]
machen, dass jmd grün vor Neid bleibt -> vor Neid grün werden
18002511: Um aspecto de relevância suficiente para fazer qualquer lisboeta ficar verde de <inveja> .

1,2 *meter* [~$_{acc}$ (a N_{dat})] Neid hineinstecken -> jdn neidisch machen
16508330: Tenho aqui umas cerejas de meter <inveja> .

provocar [~$_{acc}$] provozieren
2596786: A voz de Walker não é voz para provocar muitas <invejas> .

despertar [~$_{acc}$] wecken

38526129: De tal modo que continua a despertar cobiças , <invejas> e acusações de se aproveitar de situações em benefício pessoal .

suscitar [as ~s$_{acc}$ (de/em N$_{dat}$,)] erregen

140546649: Ganham muito mais dinheiro que o comum dos portugueses , pelo que suscitam a <inveja> dos simples .

gerar [~$_{acc}$] erzeugen

125957320: Só que , pelo caminho , gerou rivalidades , <invejas> e raivas surdas .

X leidet unter dem Neid von Y

3 *sofrer* [a ~$_{acc}$ de N$_{dat}$] erleidet den Neid von -> leiden unter

154616164: A esposa , Albertina Alves , afirma que o marido « nunca fez mal a ninguém , pelo contrário » , mas que sempre sofreu a <inveja> do seu agressor .

Es gibt Neid

haver [~] es hat -> es gibt

24822992: Há muita <inveja> e ciúme por quem já teve sucesso neste país .

X empfindet Neid (auf Y)

ter	6.7	2.0	60	1.1	1.0	3
roer-se	6.0	9.3	37			
olhar	3.5	5.3	13	1.4	5.4	2
ficar	3.1	2.4	12(7)			
morrer	2.9	3.7	9			
sentir	2.9	3.5	9			

Oper$_1$ empfinden, fühlen [~$_{acc}$]

Oper$_1$ ter, sentir
ContOper$_1$ ficar com
Magn+Real$_1$ roer-se de, morrer de

ter
ter ~ de alguém - auf jdn neidisch sein (PO) = invejar (LS)

roer-se
vor Neid erblassen - *form* - *path* - roer-se de ~ *fam* (IDP)
roer-se de ~ *F* - vor Neid platzen (LS) vor Neid vergehen (PO)

morrer
vor Neid erblassen *form* - *path* - morrer de ~ *fam* (IDP)
grün/gelb/blaß vor Neid werden *path* - até morrer de ~ *fam* (IDP)
vor Neid platzen - morrer de ~ (PO)

Y verursacht Neid (in X)

fazer	18.3	4.6	343*			
corar	4.0	7.8	16			
causar	3.8	4.9	15			
ficar verde	3.1	2.4	12(3)			
meter	2.8	4.8	8			
provocar	2.5	3.2	7	2.6	5.2	7
despertar	2.2	5.1	5	2.5	7.3	6
suscitar	1.9	4.1	4	2.5	6.6	6
gerar	1.7	3.6	3	2.2	6.1	5

Caus$_2$Func$_1$ *hervorrufen [bei N$_{dat}$ ~$_{acc}$], wecken [in N$_{dat}$ ~$_{acc}$], erregen [in N$_{dat}$ ~$_{acc}$]*

CausContFunc$_1$ *schüren [in N$_{dat}$ ~$_{acc}$]*

$CausFunc_1$	provocar, despertar, gerar
$Caus_{(2)}Func_1$	fazer, causar, meter, suscitar
$Magn+Caus_2Fact_1$	fazer corar de, fazer ficar verde de

causar
Causar grande inveja a (AU)
Neid erregen - causar inveja (LS)

X leidet unter dem Neid von Y

sofrer 1.6 2.6 3

$Oper_2$ sofrer de

Es gibt Neid

haver 1.5 0.9 7 2.2 2.7 6

$Func_0$ haver

Anmerkungen

Für folgende lexikalische Funktionen befinden sich unter den lemmatisierten portugiesischen Verben keine Kollokate:

fastFinFunc$_0$	*verfliegen*
$Liqu_1Func_0$	überwinden [$PRON_{poss}$ / DET ~$_{acc}$]
$IncepFunc_1$	aufkommen [in N_{dat}]
$Magn+IncepFunc_1$	erfassen [N_{acc}]
$Liqu_1Fact_0$	unterdrücken [$PRON_{poss}$ / DET ~$_{acc}$]
$Magn+IncepFact_1$	überkommen [N_{acc}]

Für die folgenden Einträge der untersuchten Wörterbüchern gibt es keine Fundstellen im Corpus:

Matar de inveja (AU) (töten)
alg até parece que estala de inveja *fam* (IDP) (ausbrechen)
schüren - (*Eifersucht*) atiçar (PO)

Die Kollokate folgender Einträge der untersuchten Wörterbüchern sind nicht lemmatisiert:

grün/gelb/blaß vor Neid werden *path* - morder-se de inveja *fam* (IDP) (beißen)
alg até parece que estoira de inveja *fam* (IDP) (platzen)

susto

Sg: 1263, Pl: 225

Schreck(en) (LS), Schreck (PO)

X bekommt/hat/erleidet einen Schrecken

+begin 1 *apanhar* [um ~$_{acc}$] ernten, pflücken -> bekommen
1105653: Malaquias apanhou um <susto> .

+begin *sofrer* [o | um ~$_{acc}$] erleiden
91072140: O próprio presidente sofreu um <susto> quando uma manhã se dirigia à Sala Oval .

1 *não passar* [de um ~]
hinausgehen über -> mit dem Schreck davonkommen (im Port. nur unpersönlich)
129320503: Afinal , tudo não passou de um <susto> e o pequeno incêndio que criou tanto alarido foi circunscrito a uma parte da carpintaria , situada na cave , e que consumiu apenas algumas madeiras .

+intense *morrer* [de ~] sterben vor
43715178: Mas quando Camilo editasse « O Regicida » , o grande megalómano , que Santana é , morreria de <susto> .

passar [por um ~, (por) os ~s] durchmachen
76402798: Para isso era preciso ousar mais , fazer os checos passarem mais alguns <sustos> , pelo menos , mas o jogo acabava por esgotar-se entre as duas áreas .

viver [o | um ~$_{acc}$] erleben

46731841: Mas as oportunidades de golo não surgiam e só aos 70 ' se viveu o primeiro <susto> : um livre de Breheme fez a bola bater na barreira e sair a escassos centímetros do poste da baliza do atento Pudar .

ter [um ADJ ~] haben

68618747: Não houve feridos , mas Ribeiro confessava que teve um grande <susto> .

Y verursacht einen Schreck (in X)

1 *pregar* [um ~$_{acc}$ a N$_{dat}$] einschlagen -> einjagen

662853: « Das pessoas que eu conhecia , era ele aquele que eu pensei que me poderia ajudar , pregando um <susto> ao indivíduo que me ficou com o dinheiro » .

provocar [um ~ (a/em N$_{dat}$)] provozieren

151078723: O paraguaio mandou para canto e a jogada terminou com um remate de Swierczewski (56 ') a provocar grande <susto> .

causar [um ~ (a/em N$_{dat}$)] verursachen

133019298: Por sorte não transitava na artéria qualquer veículo e a derrocada apenas causou um valente <susto> numa mulher que passava perto .

2 *dar* [um ~ (a N$_{dat}$)] geben -> jdn erschrecken

103656129: Quero lhes dar um <susto> .

X überwindet den Schreck nicht

1 *não ganhar* [para o ~$_{acc}$]
gewinnt nicht für -> nicht über den Schreck hinwegkommen

30739470: Os oito países mais industrializados do mundo não ganharam para o <susto> , ontem , em Birmingham , na última sessão da cimeira do G8 , quando o chanceler alemão Helmut Kohl gelou a sala com o anúncio de um teste atómico paquistanês :

não livrar-se [de um ~] sich nicht befreien von

124015030: Logo atrás de si seguia Schumacher , que não se livrou de um <susto> .

X überwindet den Schreck

recuperar [do ~$_{dat}$] sich erholen von

36442564: As crianças recuperam do <susto> , enquanto aguardam por eventuais notícias dos familiares que deixaram em Bissau .

livrar [do ~$_{dat}$] sich befreien von

171599963: Apenas um arranhão mas que não chegou para o livrar do <susto> ...

Es gibt Schrecken

haver [um ~, ~s]
es hat Schreck -> im Deutschen auch mit geben (es gibt Schreck) nicht gebräuchlich

57866201: Houve <sustos> , há sempre armas psicológicas que se usam .

Der Schreck lässt nach

passar vorübergehen

118769833: Mas o <susto> passou e ontem de manhã , em Quarteira , Acácio apresentou-se à partida da segunda-etapa , em que conquistou a nona posição , com o mesmo tempo do vencedor .

X bekommt/hat/erleidet einen Schrecken

apanhar	11.4	7.3	131		4.4	7.1	20
sofrer	3.9	4.2	16		1.7	4.3	3
não passar	4.4	3.1	22(16)				
morrer	3.0	3.7	10				
passar	4.4	3.1	22(2)		1.9	3.1	4
viver	1.4	1.9	3		1.6	3.6	3
ter	2.1	0.7	17(3)		0.9	0.7	3

$Oper_1$	empfinden, fühlen [~$_{acc}$]
$IncepOper_1$	*bekommen [DET ~$_{acc}$]*

$Oper_1$	ter
$IncepOper_1$	apanhar
$ContOper_1$	passar por
$Oper_2$	sofrer, não passar de
$Magn+Real_1$	morrer de

apanhar
einen Schreck bekommen - apanhar um ~ (LS, PO)
bekommen - *(Schreck)* apanhar (LS,PO)

morrer
der Schreck(en) fährt j. in die/alle Glieder/Knochen *path* - (quase) morrer de susto/medo (IDP)

Y verursacht einen Schreck (in X)

pregar	8.0 8.7 65	3.3 8.7 11
provocar	5.1 4.4 27	1.9 4.3 4
causar	2.7 4.2 8	
dar	2.1 1.7 7	1.2 2.2 2

$Caus_2Oper_1$	*versetzen [N_{acc} in ~$_{acc}$]*
$Caus_2Func_1$	hervorrufen [bei N_{dat} ~$_{acc}$]

$Caus_{(2)}Func_1$	provocar, causar, dar
$Caus_2Func_1$	pregar

pregar
jm. (mit etw.) einen Schrecken einjagen *path* - pregar um ~ a alg (com qc) (IDP)
jdm einen Schreck einjagen - pregar um ~ a alguém (LS, PO_{N+V}), <->

X überwindet den Schreck nicht

não ganhar	10.5 5.5 112
não livrar-se	1.9 5.9 4(3)

¬ $Liqu_1Func_0$	não ganhar para

X überwindet den Schreck

recuperar	3.4 4.3 12
livrar	1.9 5.9 4(1)

$Liqu_1Func_0$	überwinden [$PRON_{poss}$ / DET ~$_{acc}$]
$Liqu_1Func_0$	recuperar de, livrar de

recuperar
der Schreck(en) sitze/steckt j. (noch) in allen/den Gliedern *path* - alg ainda não conseguiu recuperar-se do susto (que apanhou) (IDP)

Es gibt Schrecken

haver	1.2 0.6 6 2.9 2.9 10

$Func_0$	haver

Der Schreck lässt nach

passar	4.4 3.1 22(4)

$FinFunc_0$	sich legen
IncepPredMinus	nachlassen
fast $FinFunc_0$	verfliegen

$FinFunc_0$	passar

Anmerkungen

Für folgende lexikalische Funktionen befinden sich unter den lemmatisierten portugiesischen Verben keine Kollokate:

$Magn + IncepFunc_1$	erfassen [N_{acc}]
$Magn + fast\ IncepFunc_1$	packen [N_{acc}]
$Liqu_1Fact_0$	unterdrücken [$PRON_{poss}$ / DET ~$_{acc}$]
$Magn + IncepFact_1$	überkommen [N_{acc}]

medo (Sg: 12288, Pl: 989)

Angst (*de* vor) (LS), Angst (PO)

X hat / empfindet Angst (vor Y) (log-like, Kookkurrenz)

ter	~ (de)	haben	(25159, 4055)
sentir	~ (de)	spüren	(431, 88)
estar	com ~ (de)	sein mit	(367, 236)
tremer	de ~	zittern vor	(287, 31)
viver	com / em ~ (de)	leben mit / in	(273, 72)
ficar	com ~	bekommen	(132, 69)
morrer	de ~	sterben vor	(85, 27)
borrar -se	de ~	sich beklecksen -> in die Hose machen vor	(54, 5)
andar	com ~	gehen mit -> haben	(38, 14)

Y verursacht Angst (in X)

meter	~ (a)	hineinstecken -> einjagen	(3324, 323)
provocar	o ~	verursachen	(145, 72)
inspirar	~ (a)	einflößen	(60, 13)
causar	~	verursachen	(60, 17)
fazer	~ (a)	machen	(59, 89)
gerar	o ~	erregen	(68, 17)
instilar	o ~ (em)	einträufeln	(45, 4)

ódio (Sg: 2347, Pl: 659)

Hass (LS), Hass (*por* auf) (PO)

X empfindet Hass (auf Y) (log-like, Kookkurrenz)

sentir	~ (por)	empfinden	(142, 25)
ter	um ~ (a)	haben->hegen	(26, 40)

Y verursacht Hass (in X)

incitar	ao ódio (ADJ / contra)	anstacheln -> schüren	(250, 21)
atiçar	o ódio	schüren	(176, 12)
destilar	~ (sobre)	destillieren -> schüren	(174, 13)
nutrir	um ódio ADJ	nähren	(168, 13)
alimentar	um | o ódio	nähren	(108, 15)
fomentar	o ~	fördern -> schüren	(108, 11)
instigar	(a)o ~	anzetteln, anstiften -> schüren	(103, 8)
semear	o ~ (em)	säen	(69, 7)
desencadear	o | ∅ ~ (de/contra)	entfesseln	(45, 7)
motivado	pelo ~	motiviert von	(44, 7)
gerar	o | ∅ ~	erzeugen	(38, 7)
espalhar	o ~	verbreiten	(38, 6)
suscitar	o | ∅ ~ (entre)	erregen	(35, 6)

				pregar	o ~ (a/contra)	einschlagen -> schüren	(34, 4)
				provocar	o | ∅ ~ (de/ADJ)	provozieren	(27, 8)
				cultivar	o ~	kultivieren	(20, 3)
				criar	o | ∅ ~ (a/entre)	erzeugen	(17, 7)
				despertar	~	wecken	(17, 3)
				inspirar	o | ∅ ~	einflößen -> verursachen	(15, 3)
X verliert seine Angst (vor Y)							
perder	o ~ (de/a)	verlieren	(352, 91)				
X überwindet seine Angst (vor Y)							
vencer	o ~	besiegen	(139, 40)				
superar	o ~	überwinden	(40, 9)				
X bringt seine Angst (vor Y) zum Ausdruck							
confessar	o (seu) | ∅ ~	eingestehen	(122, 25)	exprimir	o | ∅ ~	ausdrücken	(16, 3)
mostrar	~	zeigen	(57, 26)	manifestar	o | um | | ∅ ~	zeigen	(16, 5)
revelar	o | um | /∅ ~	offenbaren	(56, 26)	demonstrar	o | um | /∅ ~	demonstrieren	(15, 4)
X versteckt seine Angst (vor Y)							
esconder	o ~ (de)	verstecken	(93, 22)	esconder	o ~ (a)	verstecken	(43, 8)
				disfarçar	o | um | /∅ ~	verbergen	(19, 3)
X tritt seiner Angst gegenüber							
enfrentar	o ~	gegenübertreten	(39, 12)				
				X handelt aus Hass			
				matar	por ~	aus Hass töten	(38, 6)
				X unterdrückt seinen Hass (auf Y)			
				reprimir	o ~	unterdrücken	(25, 3)

Es gibt Angst

dominar	o ~	herrschen	(172, 35)
haver	~	haben -> geben	(108, 102)
reinar	o ~	regieren	(46, 8)

Angst lässt sich nieder in Y

instalar-se	o ~ (em)	sich niederlassen	(119, 32)

Angst wächst (in Y)

crescer	o ~ (em)	wachsen	

Es gibt Hass

haver	~	haben -> geben	(33, 24)

Hass bricht aus

nascer	o ~	entstehen, entspringen	(34, 8)

6.3. Differenzierung polysemer und synonymer Substantive anhand der Kookkurrenzdaten

Im Gegensatz zu den ausführlichen Angaben aus Kapitel 6.2 in Form von (morpho)syntaktischen Informationen, Beispielsätzen und der gruppierenden Wirkung der Glossen, wird in der folgenden Darstellung der Kollokate polysemer und synonymer Substantive eine reduzierte Form der Kollokationsangaben gewählt. Sie variiert leicht je nach dem untersuchten Substantiv und den erforderlichen Angaben, die eine Differenzierung der Bedeutung des entsprechenden Substantivs gewähren. Präzise Auskünfte über die externen Kollokationsinformationen sind in den Exzerptionsdateien anhand von Corpusbelegen zu finden, die Rankinglisten bieten die numerischen Werte der statistischen Assoziationsmaße.

pena

a) Strafe

cumprir	uma | a | ∅ pena (1028) - penas (206)	eine Strafe verbüßen
condenado	a uma pena (457) - a penas (348)	zu einer Strafe verurteilt
aplicar	a | uma pena (278) - penas (155)	eine Strafe verhängen
incorrer	numa pena (236) - em penas (45)	einer Strafe unterliegen
as penas (185)	ser + ADJ	die Strafen sind + ADJ
dar	pena de morte/de prisão (103)* - penas (5)	Haft/Todesstrafe -Strafen geben
abolir	a pena de morte (64)	die Todesstrafe abschaffen
ter	penas (55)	Strafen haben / zu Strafen führen
enfrentar	a | uma pena (42) - penas (8)	einer Strafe entgegentreten
receber	uma | a pena (37) - penas (21)	eine Strafe erhalten
impor	a pena (33) - penas (9)	eine Strafe auferlegen
sofrer	uma | a pena (32) - penas (15)	eine Strafe erhalten
defender	a pena de morte (31) - penas + ADJ (7)	die Todesstrafe/ harte Strafen verteidigen
apanhar	uma | a pena (25) - penas (5)	eine Strafe bekommen
aumentar	a pena (13) - as penas (25)	die Strafen erhöhen
agravar	a pena (1) - as penas (21)	die Strafen verschärfen
exigir	a | uma pena (19) - penas (5)	eine Strafe fordern
acarretar	uma | a pena (10) - penas (4)	eine Strafe mit sich bringen
diminuir	a pena (11) - as penas (10)	die Strafe verringern

b) Kummer, Leid, Mitleid, Qual

vale	a pena (5553) - as penas (1)	das lohnt sich, ist die Mühe wert
é	pena (2340)	es ist schade
ter	pena de/que (841)	jmd/es tut einem Leid
faz	pena (282)*	es tut einem Leid
dá	pena + INF (103)*	es tut einem Leid + INF
merecer	a pena (60)	das lohnt sich, ist die Mühe wert
sentir	pena de (55)	jmd tut einem Leid
ficar	com pena (de) (47)	jmd/es tut einem Leid
meter	pena (10)	traurig stimmen, Mitleid hervorrufen

c) Feder

keine Vorkommen mit den lemmatisierten Verben

Eine klare Bedeutungsunterscheidung anhand der verbalen Kollokate ist bei dem Substantiv *pena* gegeben. Von den drei möglichen Bedeutungen sind zwei etymologisch verwandt: a) *Strafe* und b) *Kummer, Leid, Mitleid, Qual* stammen von dem lateinischen Wort *poena* ab, das wiederum auf das griechische Lexem *poiné* zurückgeht. In seiner dritten Bedeutung *Schreib-*, bzw. *Vogelfeder*, die vom lateinischen Wort *penna* abstammt, kommt das Substantiv *pena* im Corpus *Cetempúblico* zusammen mit den lemmatisierten Verben nicht vor. Die etymologisch verwandten Bedeutungen sind in fast allen Fällen über die Verwendung der verbalen Kollokate zu unterscheiden, die Verben, die mit *pena* in der einen oder anderen Bedeutung kookkurrieren entsprechen (nahezu) komplementären Mengen.
In die Auswertung fließen nur diejenigen Verben mit ein, deren Kookkurrenzfrequenzen mit *pena* oder *penas* einen t-score größer als 3 erzielen. In gewisser Weise kontaminiert werden die Extraktionsergebnisse durch das extrem häufige Vorkommen des Phrasems *vale a pena* ('es lohnt sich'), dem häufig ein Verb im Infinitiv folgt, in diesem Fall wirkt sich die globale Ausdehnung des Suchraums auf ein Wort, das rechts vom Substantiv steht, sehr negativ auf die Precision-Ergebnisse aus. Beispielsweise sind die gesamten 23 Kookkurrenzen von *pena* + *chorar*, die einen t-score von 4.7 ergeben, negative Fundstellen, die in folgendem Kontext vorkommen: "*« Não vale a <pena> chorar »*" (Cetempúblico, 63932658) ('es lohnt sich nicht zu weinen'). Verben, die in der Rankingliste erscheinen, obwohl sie keine Kollokationsstruktur mit dem Substantiv aufweisen, werden in die Disambiguierung nicht inkludiert. Die Durchsicht der Exzerptionsdateien ist nicht zu vermeiden, um falsche Fundstellen auszusortieren. Nur wenige der aufgeführten Verben, kommen sowohl als Kollokat von *pena* als auch in der Infinitivform nach dem Phrasem *vale a pena* häufig vor. Die entsprechenden Verben *dar* und *fazer* werden durch einen Asterisk gekennzeichnet, der hinter der Frequenzangabe der Kookkurrenz in Klammern verdeutlichen soll, dass nur ein Teil der Kookkurrenzen als Kollokation der Form Verb + *pena* existiert. Nur die eher inhaltsarmen und sehr gebräuchlichen Verben *ser, dar, ter* kommen mit *pena* in beiden Bedeutungen vor, in diesen Fällen wirkt die Numerusform von *pena* oder die Satzstellung bedeutungsunterscheidend. Die Kollokate der lexikalisierten Nomenkomposita, in denen *pena* als Kopflexem fungiert, verhalten sich kongruent zu den Kollokaten von *pena*. Die Nomenkomposita werden in dem Fall als Kollokationspartner genannt, wenn ihre Anzahl die Vorkommen des einfachen Lexems *pena* übersteigt, sie sind jedoch auch mit den Kollokaten üblich, für die nur *pena* als Basis verzeichnet ist.

inclinação

a) Neigung, Gefälle

ter	uma inclinação de/ADJ (20) - inclinações ADJ (5)	haben
dar	a | uma inclinação de/ADJ (8)	geben
provocar	uma inclinação de/Adj (2)	hervorrufen
disfarçada	inclinação (2)	verborgene
manifesta-se	a inclinação (1)	sich zeigen

b) Neigung, Zuneigung, Hang zu, Talent

ter	uma | ∅ inclinação para/por/ADJ (35) - inclinações para/por/ADJ (7)	haben
mostrar	uma inclinação para/por (11) - inclinações ADJ (1)	zeigen
revelar	uma | ADJ inclinação para/por (9) - inclinações (1)	offenbaren
manifestar	uma | ADJ inclinação para/por (5)	zeigen
sentir	uma | ADJ inclinação para/por (5)	fühlen
existir	uma inclinação para (3)	existieren
disfarçar	uma inclinação para (1)	verhehlen
provocar	uma inclinação ADJ (1)	hervorrufen

Da das Nomen *inclinação* im Vergleich zum Nomen *pena* im Corpus relativ selten vorkommt, wird hier ein t-score von 1,5 als Grenze angesetzt, um die Kollokationsdaten zur Differenzierung des polysemen Lexems zu verwenden. Die Verteilung der Kollokate zwischen den beiden Bedeutungen von *inclinação* liefert ein ganz anderes Bild als die komplementäre Verwendung der Kollokate bei *pena*. Die Kollokate, die mit *inclinação* in der Bedeutung von 'Gefälle' kollokieren, kommen auch mit *inclinação* in der Bedeutung 'Zuneigung' vor. Die Bedeutungsunterscheidung des polysemen Substantivs funktioniert nicht über die Semantik der Verben, sondern über modifizierende Adjektive, die von *inclinação* regierten Präpositionen, oder weitere Substantive, die im engen Kontext von *inclinação* stehen. Wird *inclinação* von einer Präposition begleitet, ist die Bedeutung eindeutig auch ohne den weiteren Kontext zu bestimmen. Fehlt der Valenzrahmen, der zur Bedeutungsunterscheidung beiträgt, geben meistens die adjektivische Kollokate Aufschluss über die Semantik von *inclinação*: " *« não é um poeta filósofo nem tem <inclinações> intelectuais »* " (Cetempúblico, 111860255) ('er ist kein philosophischer Dichter, noch hat er intellektuelles Talent'). Mitunter sind es aber auch nur die semantischen Eigenschaften des Subjekts, die über die Bedeutung von *inclinação* entscheiden: "*... se o IP5 tem <inclinações> superiores ao que é normal nas estradas do mundo civilizado , ..."* (Cetempúblico, 77836197) ('wenn die IP5 größere Gefälle hat als die Straßen der zivilisierten Welt normalerweise, ... ').

Ein Vergleich der beiden polysemen Substantive *pena* und *inclinação* bezüglich der Disambiguierungsmöglichkeiten durch die verbalen Kollokate zeigt sehr unterschiedliche Resultate. Während die Bedeutungsdifferenzierung anhand der Verben bei *pena* eindeutig ist, werden die gebräuchlichsten Kollokate von *inclinação* mit den verschiedenen Bedeutungen von *inclinação* verwandt. Die Bedeutungsunterscheidung ist bei inclinação nur über den Subkategorisierungsrahmen, die adjektivischen Kollokate oder den weiteren Kontext zu leisten.

raiva, fúria

Ein interessantes Phänomen stellt die Verwendung synonymer Substantive mit unterschiedlichen verbalen Kollokaten dar. *Raiva* und *fúria* werden im *PONS* und im *Langenscheidt* als Übersetzung von *Wut* aufgeführt. Untersucht man die verbalen Kollokate, deren Kookkurrenzen einen t-score größer als 2 erzielen, zeigt sich eine deutliche Abweichung im Kollokationsverhalten. *Raiva* ist die eigene Wut, die man verspürt, zum Ausdruck bringt, oder gegen jemand richtet, *fúria* hingegen eher die Wut, die durch jemanden oder etwas verursacht wird. Die Ursache der Wut steht bei *fúria* in der Subjektposition und übernimmt die Rolle des aktiven Agens. Konstruktionen wie im Deutschen, in denen der Verursacher der Wut die Rolle des Patiens einnimmt *(sie hatten eine riesengroße Wut auf den Schiedsrichter*), sind im Portugiesischen nur mit *raiva* möglich und nur in der Kombination mit einem Kollokat (*ter*). Soll die Ursache der Wut in Kombination mit dem Wortstamm *fúria* als passiver Auslöser zum Ausdruck kommen, übernehmen Derivate von *fúria* wie reflexive Verben oder Adjektive (*enfurecer-se com, estar furioso com alg.*) die entsprechende Funktion. Die beiden Gefühlssubstantive *raiva* und *fúria* sind in ihrer Bedeutung synonym, doch werden die Substantiv-Verb Kollokationen weitgehend in verschiedenen konzeptuellen Kontexten verwendet.

raiva

sentir	raiva (17)	Wut (ver)spüren
chorar	de raiva (16)	vor/aus Wut weinen
ter	raiva (a/de) (30) - raivas (a) (2)	(auf jdn) Wut haben
ficar	com raiva (6) / ficar ADJ de raiva (6)	wütend sein
exprimir	a | uma raiva (ADJ/contra) (8)	Wut ausdrücken
dar	(ADJ) raiva (10)	wütend machen
manifestar	(a) raiva (5) / a raiva manifesta-se (2)	Wut zeigen
há	raiva (10)	Wut existiert
dirigir	a raiva contra/de/para (5)	die Wut gegen jdn richten
destilar	a raiva (de/contra//que) (4)	die Wut bricht aus

fúria

provocar	a fúria de/em/ADJ (52) - fúrias em (1)	etw. provoziert bei jdn Wut
ter	uma fúria (33) - fúrias (2)	wütend sein
esconder	a (sua) fúria (por/contra/com) (13)	Wut verstecken
enfrentar	a fúria de (12)	der Wut von jdm gegenübertreten
exprimir	a sua fúria (8)	seine Wut ausdrücken
desencadear	a fúria de (8) - fúrias (1)	die Wut von jdn entfesseln
acalmar	a (sua) fúria de/ADJ (7) - as fúrias (1)	die Wut beruhigen
entrar	em fúria (8)	wütend werden
travar	a fúria de/ADJ (6)	die Wut von jdn bremsen
despertar	a fúria de/ADJ (5)	die Wut von jdn wecken
suscitar	a fúria de/ADJ (5)	die Wut von jdn erregen
causar	fúria em/de (2) / a fúria causa (3)	Wut bei jdn verursachen
manifestar	a fúria por (2) / manifestar-se na fúria (1)	Wut zeigen über
aumentar	a fúria (de) (4) / a fúria aumenta (1)	die Wut (von jdn) vergrößern
sentir	fúria (5)	Wut (ver)spüren

6.4. Varietätenspezifische Kollokationen

Aufgrund der stark divergierenden Größe der untersuchten Corpora der europäischen und der brasilianischen Standardsprache des Portugiesischen zeigt der Vergleich zwischen den verbalen Kollokaten der Gefühlssubstantive kein ausgewogenes Bild. Wie sich der geringe Umfang des brasilianischen Corpus (*Cetenfolha*) von nur ca. 13% des europäischen Corpus (*Cetempúblico*) auf das Vorkommen von Kollokationen auswirkt, wurde bereits zu Beginn des Kapitels illustriert. Die folgende Darstellung der Verwendung der Kollokate in den beiden Ländern zeigt daher häufig Kollokate der europäischen Variante des Portugiesischen, für die brasilianische Äquivalente fehlen, was jedoch nicht auf eine geringere Anzahl von verfügbaren Kollokaten im Brasilianischen schließen lässt, sondern vielmehr verdeutlicht, dass eine Corpusgröße von 24 Millionen Wörtern für die lexikalische Akquisition von Kollokationen keine geeignete Grundlage bietet. Eine weitere Einschränkung (für beide Corpora) ergibt sich durch die geringe Anzahl von 226 lemmatisierten Verben.
In keinem der untersuchten (deutsch/)portugiesischen Wörterbüchern wird auf kollokationale Divergenzen zwischen den beiden Varietäten des Portugiesischen hingewiesen. Insofern zeigt der folgende Überblick über die abweichende Verwendung verbaler Kollokate in Portugal und Brasilien erstmalig Ergebnisse, die aus dem Vergleich der maschinellen lexikalischen Akquisition von Kollokationen aus Corpora aus beiden Ländern stammen. Neben kollokationalen Divergenzen werden auch unterschiedliche Präferenzen im Numerusgebrauch des Substantivs für die gleiche Kollokation evident. Eventuelle weitere morphosyntaktische Eigenheiten bei der Verwendung der Kollokationen in beiden Varietäten sind in den durch PECCI generierten Ausgabedateien nur durch den manuellen Vergleich der Exzerptionsdateien mit den Umgebungsdaten der einzelnen Kollokationen zu belegen, worauf im Rahmen der vorliegenden Arbeit verzichtet wird. Es ist davon auszugehen, dass die varietätenspezifischen Divergenzen nicht nur verbale Kollokate betreffen, sondern auch in anderen Kollokationsstrukturen anzutreffen sind, und dass bei einer Extraktion aus einem Corpus das POS-Informationen enthält und das somit alle Kollokate zur Verfügung stellt, sehr viel mehr kollokationale Divergenzen festzustellen sind. Diese These ist in der Kollokationsforschung durch weitere corpusbasierte Untersuchungen zu belegen.

Portugal	Brasilien
-	ter asco (4)
ter ciúme (5) ciúmes (55)	ter ciúme (11) ciúmes (12)
provocar ciúme (5) ciúmes (17)	gerar ciúmes (3)
acalentar esperança (70) esperanças (119)	-
-	produzir excitação (3)
roer-se de inveja (37)	-
suscitar ira (10) iras (3)	-
atiçar ódio (12) ódios (8)	-
-	acarretar ódio (2)
lançar pânico (157)	-
apanhar susto (131)	tomar susto (14) levar susto (59)
pregar susto (65)	dar susto (10)

7. Clustering der portugiesischen Gefühlssubstantive

Die Gruppierung der Substantive anhand der Frequenz der mit ihnen kookkurrierenden Verben mittels Clusterverfahren wird unabhängig von den semantischen Eigenschaften der Kookkurrenten und deren syntaktischer Beziehung durchgeführt. Grundlegend für die Berechnung ist allein die Frequenz der Kookkurrenz der Substantive mit den Verben. Die Basis der folgenden Überlegungen bildet die Frage, ob sich durch die statistische Auswertung der Kookkurrenzdaten Cluster von Substantiven ergeben, die bestimmte semantische Eigenschaften teilen. Mel'čuk und Wanner (1994) stellen in dem Artikel "Lexical Co-occurrence and Lexical Inheritance" den Zusammenhang zwischen bestimmten semantischen Eigenschaften von Gefühlssubstantiven und dem kongruenten verbalen Kookkurrenzverhalten dar. Beispielsweise kookkurrieren Gefühlssubstantive mit dem semantischen Merkmal "Excited-state" mit dem Verb *sich legen*, direktionale Gefühlssubstantive kookkurrieren mit Verben wie *sich richten (gegen)*. Auch für die Verben wird verzeichnet, welche Eigenschaften die Lexeme aufweisen, mit denen sie kombinieren. *Ausbrechen* kommt mit Lexemen vor, die gleichzeitig die Eigenschaften "Intense" und "Manifested" besitzen, wie beispielsweise *Begeisterung* und *Panik*.
Ziel der Arbeit von Mel'čuk und Wanner ist die Generierung eines generischen Lexikoneintrags für Gefühlssubstantive, in dem die Werte der Lexikalischen Funktionen stehen (in diesem Fall ausschließlich Verben). Die sprachliche Realisierung der Werte der lexikalischen Funktionen ist an bestimmte semantische Bedingungen gebunden, welche die Argumente (in diesem Fall spezifische Gefühlssubstantive) erfüllen müssen, die mit dem übergeordneten generischen Lexikoneintrag *Gefühl* verbunden sind. Unter den Lexikoneinträgen der einzelnen Gefühlssubstantive erscheinen nur wenige Ergänzungen und Abweichungen der kookkurrierenden Verben, die sich von den Angaben unterscheiden, die im generischen Lexikoneintrag gespeichert sind (vgl. Kapitel 2.4.1). Die Untersuchung von Mel'čuk und Wanner beschränkt sich auf die Beschreibung von 20 Verben als mögliche Kollokate der untersuchten 40 Gefühlssubstantive und beruht auf einer manuellen Klassifikation der Kookkurrenzdaten.

7.1. Das Clusterverfahren K-Means

Um zu untersuchen, ob sich durch die statistische Auswertung des gesamten verbalen Kookkurrenzbereichs Gruppen von Gefühlssubstantiven ergeben, die bestimmte semantische Eigenschaften teilen, wurde das Clusterverfahren K-Means als Modul von PECCI implementiert (vgl. Kapitel 5.2), das u.a. in Manning/Schütze (1999, Kapitel 14: Clustering) neben weiteren Clusterverfahren ausführlich beschrieben ist. Im Gegensatz zu überwachten Klassifikationsverfahren, für die zunächst ein Trainings-Set mit gelabelten Instanzen bereitgestellt wird, sind Clusteranalysen unüberwachte Verfahren, sie brauchen keinen "Lehrer", der ihnen anhand von Trainingsdaten mit korrekten Klassenlabeln zeigt, wie die Klassifikation durchzuführen ist. Das Ziel von Clusterverfahren ist die Zusammenfassung von Objekten zu Gruppen (Cluster), wobei es eine möglichst große Homogenität innerhalb eines Clusters und eine möglichst große Heterogenität zwischen den Cluster zu erreichen gilt.
Clusterverfahren können sich u.a. hinsichtlich ihrer Zuordnungsprinzipien (exakt, probabilistisch, possibilistisch), ihrer Vorgehensweise (hierarchisch, partitionierend, heuristisch, begrifflich) und der benutzten Informationen (partiell, global) unterscheiden.

K-Means gehört zu den partitionierenden, globalen Verfahren mit exakter Zuordnung, das die gewählte Anfangspartition der zu clusternden Objekte durch Neuzuordnung schrittweise verbessert. Die Frequenzen der Kookkurrenzen von jedem Substantiv mit den in PECCI lemmatisierten Verben werden zunächst als Vektor des Substantivs gespeichert, wobei x_i die *i*-te Komponente des Vektors $\vec{x}$ bezeichnet, seinen Wert in der Dimension *i*.

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_i \end{pmatrix}$$

$$\vec{alegria} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ \vdots \\ 27 \end{pmatrix} \qquad \vec{amor} = \begin{pmatrix} 7 \\ 0 \\ 0 \\ \vdots \\ 38 \end{pmatrix}$$

Dimension 1 = *abandonar*
Dimension 2 = *abater*
Dimension 3 = *abolir*

Dimension 226 = *viver*

Die Clusterzahl K muss zu Beginn des Algorithmus festgelegt werden. Als jeweiliges Clusterzentrum wird ein Substantiv gewählt. Mittels der Berechnung der Euklidischen Distanz der Vektoren der verbleibenden Substantive zu den Vektoren der Clusterzentren werden diese demjenigen Clusterzentrum zugeordnet, zu dem die Euklidische Distanz am geringsten ausfällt. Wurden die Substantive auf die verschiedenen Clusterzentren verteilt, erfolgt im nächsten Schritt eine Neuberechnung des Vektors der Clusterzentren, der sich aus dem Mittelwert der im Cluster befindlichen Vektoren der Substantive ergibt. Das Clusterzentrum eines Clusters (der Zentroid) ist nun nicht mehr identisch mit einem der Substantive (außer im Falle eines alleinstehenden Substantivs). In der folgenden Iteration werden die Substantive wieder dem Clusterzentrum zugeordnet, zu dem die Euklidische Distanz am geringsten ausfällt. Der Algorithmus wird beendet, wenn die Cluster stabil sind, d.h., dass keine Vertauschungen mehr vorgenommen werden, und sich jedes Substantiv in dem Cluster befindet, mit dem es die maximale Ähnlichkeit in den Kookkurrenzdaten aufweist.
Eine formale Darstellung des Algorithmus geben Manning/Schütze (1999: 516):

1 Given: a set $X = \{\vec{x}_1, \ldots, \vec{x}_n\} \subseteq \mathbb{R}^m$
2 a distance measure $d : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$
3 a function for computing the mean $\mu : \mathsf{P}(\mathbb{R}) \rightarrow \mathbb{R}^m$
4 Select k initial centers $\vec{f}_1, \ldots, \vec{f}_k$
5 **while** stopping for criterion is not true **do**
6 **for** all clusters c_j **do**
7 $c_j = \{\vec{x}_i \mid \forall \vec{f}_l d(\vec{x}_i, \vec{f}_j) \leq d(\vec{x}_i, \vec{f}_l)\}$
8 **end**
9 **for** all means $\vec{f}_j$ **do**
10 $\vec{f}_j = \mu(c_j)$
11 **end**
12 **end**

Als Abstandsmaß wurde in der vorliegenden Implementierung von K-Means die Euklidische Distanz gewählt, sie misst wie weit entfernt voneinander zwei Vektoren im Vektorraum liegen (Manning/Schütze 1999: 301):

$$|\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Die Bestimmung der Ausgangsclusterzentren ist in PECCI über den Benutzerdialog auf drei Arten möglich. Die einfachste und wohl auch gebräuchlichste Art stellt die zufallsgenerierte Auswahl der Substantive dar, die mit ihren Vektoren die anfänglichen Clusterzentren bilden. Anzugeben ist hier lediglich die Anzahl der zu erzeugenden Cluster. Bei der zweiten Auswahl werden bestimmte Substantive als Clusterzentren festgelegt, was sinnvoll ist, wenn die Clusterbildung um spezifische Substantive von Interesse erscheint. Die dritte Selektion initiiert die Clusterzentren ausgehend von einem Substantiv und einer festzulegenden Schrittgröße, mit der die folgenden Substantive gewählt werden. Dieses Verfahren bietet sich an, um das Verhalten der Cluster zu untersuchen, wenn die gewählten Substantive stabil bleiben, aber die kookkurrierenden Verben und damit die Angaben in den Vektoren differieren.
Um die teilweise erheblichen Unterschiede der Frequenz der untersuchten Gefühlssubstantive im Corpus und damit auch der Kookkurrenzdaten auszugleichen, wurden die Vektoren normalisiert. Die Länge eines normalisierten Vektors ist gleich 1 $(|\vec{x}| = 1)$. Dadurch soll vermieden werden, dass Substantive, die mit den gleichen Verben in einem proportionalen Verhältnis, aber mit ganz unterschiedlichen Frequenzen kookkurrieren, aufgrund ihrer unterschiedlichen Kookkurrenzfrequenzen verschiedenen Cluster zugewiesen werden. Im Gegensatz zu der separaten Verarbeitung der Singular- und Pluralform der Nomina für die Berechnung der statistischen Assoziationsmaße und der auf den Rankinglisten aufbauenden lexikografischen Angaben, werden die Kookkurrenzfrequenzen für die Anwendung von K-Means für beide Numerusformen eines Substantivs gemeinsam unter der Singularform gespeichert.

7.2. Exemplarische Anwendungen von Clusterverfahren in der Computerlinguistik

Angewandt werden Clusterverfahren in der Computerlinguistik in vielfältigen Bereichen, in denen das Bestreben besteht, ähnliche Objekte in der gleichen Gruppe zu platzieren, und unähnliche Objekte in unterschiedlichen Gruppen zu gruppieren. Als Grundlage der Klassifikation dienen die Kookkurrenzdaten, wobei je nach Bedarf und Verfahren genau zu bestimmen ist, welche Wörter der sprachlichen Nachbarschaft in welchem Abstand für die Berechnung relevant sind. An dieser Stelle soll noch einmal das Zitat aus Manning/Schütze aufgeführt werden, das die Clusterverfahren mit der Idee des Britischen Kontextualismus vergleicht, Sprache durch lexikalische Sets zu strukturieren (vgl. Kapitel 1): "First, the left and right neighbors of tokens of each word in the Brown corpus were tallied. These distributions give a fairly true implementation of Firth's idea that one can categorize a word by the words that occur around it. But now, rather than looking for distinctive collocations, as in chapter 5, we are capturing and using the whole distributional pattern of the word. Word similarity was then measured as the degree of overlap in the distributions of these neighbors for the two words in question" (Manning/Schütze 2002: 495-496).

Manning/Schütze (1999: 498) geben als Beispiel für die Verwendung von Clusterdaten in einem Lernverfahren das Clustering der Substantive in einem bestimmten Corpus an. Durch die ihnen gemeinsame sprachliche Umgebung wie *until, last, morning* werden die Wochentage in einem Cluster zusammengefasst. Kommt in dem Corpus für den Wochentag *Friday* keine Fundstelle mit einer geeigneten Präposition für die Übersetzung von *am Freitag* vor, kann unter der Annahme, dass die sprachliche Umgebung, die für einen Teilnehmer des Clusters vorliegt, auch für die anderen Teilnehmer gilt, über das Vorkommen von *on Monday* geschlossen werden, dass auch für *Friday* die korrekte Übersetzung *on Friday* ist. Ein maschinelles Übersetzungssystem, das auf der Analyse von Corpora basiert, kann durch die Information aus den Clusterverfahren fehlende Übersetzungsäquivalente durch Inferenz herleiten. Ein weiteres Beispiel, das auf der Anwendung des Clusterverfahrens K-Means basiert, präsentiert 20 Substantive aus dem *New York Times* Corpus, die fünf Cluster zugeordnet werden. Dabei handelt es sich um Substantive, die entweder Eigennamen sind, oder aus den Themenbereichen 'Regierung', 'Finanzen', 'Sport' oder 'Forschung' stammen. Durch das Clustering der Substantive mit K-Means anhand ihrer Kookkurrenzdaten ergeben sich Cluster, die jeweils die den Themenbereichen zugehörigen Substantive oder die Eigennamen enthalten (Manning/Schütze 1999: 518).
Substantive können durch die Darstellung der Kookkurrenzfrequenzen im multidimensionalen Vektorraum mit adäquaten Clusterverfahren zu Cluster gruppiert werden, die semantische Ähnlichkeiten aufweisen. Wird die Berechnung der Euklidischen Distanz auf normalisierte Vektoren außerhalb von Clusterverfahren zur Ermittlung der Ähnlichkeit von jeweils zwei Vektoren angewandt, sind die Ranking-Ergebnisse identisch mit den Ergebnissen, die durch die Werte der Kosinus-Abweichung für dieselben Vektorräume erzielt werden. Der Kosinus gibt den Winkel zwischen zwei Vektoren an, bei totaler Übereinstimmung ist der Wert 1, zeigen die Vektoren in gegensätzliche Richtungen liegt der Kosinus-Wert bei -1. Die Berechnung des Kosinus zum Vergleich von Vektorräumen ist ein häufig angewandtes Verfahren, das die semantische Ähnlichkeit zwischen Wörtern anhand ihrer Kookkurrenzdaten bestimmt. Hier werden zu ausgewählten Wörtern diejenigen Wörter im Corpus ermittelt, die die größten Ähnlichkeiten im Kookkurrenzverhalten im Corpus aufweisen und deren Vektoren dadurch eine maximale Ähnlichkeit im Kosinus-Wert zeigen. Die auf diese Weise ermittelten nächsten Nachbarn im *New York Times* Corpus von *garlic* sind *sauce, pepper, salt* und *cup*, die nächsten Nachbarn von *fallen* sind *fell, decline, rise* und *drop* (vgl. Manning/Schütze 1999, Kapitel 8.5: Semantic Similarity).
Das Vektorraummodell und die damit verbundene Metapher von räumlicher und semantischer Nähe wird auch im Information Retrieval zur Bestimmung der Ähnlichkeit zwischen einer Anfrage und den verschiedenen Dokumenten verwandt. Die Vektoren der Dokumente, die die Vorkommen der Wörter beinhalten, werden normalisiert, damit längere Dokumente gegenüber den kürzeren Dokumenten nicht bevorzugt sind, und anhand der Kosinus-Abweichung zur Anfrage gruppiert (Manning/Schütze 1999, Kapitel 13: Topics in Information Retrieval, 541). Dieses Verfahren wird auch von Dörre/Gerstl/Seiffert (2001) in ihrem Artikel "Volltextsuche und Text Mining" näher erläutert, wo außerdem die Anwendung von Clusterverfahren zur Analyse von Textkollektionen vorgestellt wird. Die Aufgabe des Clustering besteht bei der Gruppierung von Dokumenten darin, eine Menge von Texten so zu strukturieren, dass inhaltlich ähnliche Texte im selben Cluster auftreten. Die Distanz zweier Texte wird auch hier anhand der im Vektor repräsentierten Frequenzen der im Dokument enthaltenen Wörter ermittelt.

Im Bereich der Informationstheorie ist es üblicherweise die Gesamtheit der in den Dokumenten existenten Wörtern, die in die Berechnung der Ähnlichkeit der Texte mit einfließt. Der aus den Texten generierte multidimensionale Merkmalsvektor wird lediglich um Stoppwörter wie Artikel oder Präpositionen reduziert, die über die Identität eines Textes keine Auskunft geben, da sie unabhängig von den behandelten Themengebieten Verwendung finden. Auch in den aus Manning/Schütze (1999) zitierten Beispielen sind sämtliche kookkurrierenden Wörter (inklusive der Stoppwörter) präsent. Die Anzahl der Dimensionen der Vektoren wird in den informationstheoretischen Verfahren und bei Manning/Schütze zusätzlich durch die Sammlung der Frequenzen für einzelne Wortformen erhöht, denn die Texte oder Corpora, aus denen die Belegung der Merkmalsvektoren stammt, werden ohne linguistische Aufbereitung verwendet, POS- oder Lemmainformationen sind daher nicht relevant.
Das Clustering englischer Nomina mit einem grammatikalisch eingeschränkten Merkmalsraum stellen Pereira/Tishby/Lee 1993 in dem Artikel "Distributional Clustering of English Words" vor. Sie wenden ein hierarchisches Clusterverfahren auf die Kookkurrenzfrequenzen von Nomina an, die sich in der Position des direkten Objekts von Vollverben befinden. Grundlage sind getaggte Corpora, in denen sich die syntaktische Beziehung der Nomina zu den Verben aufgrund der rigiden SVO-Anordnung im Englischen über die Wortstellung identifizieren lässt. In den Vektoren werden nur die Kookkurrenzfrequenzen der Nomina mit den Verben verzeichnet. Das Ergebnis sind in vielen Fällen "semantisch signifikante" Cluster, wie "structure, relationship, aspect, system" oder "conductor, vicepresident, director, chairman". Die Resultate des Clustering können für die Gruppierung lexikalischer Assoziationen in lexikalisierten Grammatiksystemen benutzt werden, beispielsweise in einer stochastisch lexikalisierten Tree-Adjoining Grammatik (Pereira/ Tishby/Lee 1993: 189).

7.3. Die Ergebnisse des Clustering und die Klassifikation der Gefühlssubstantive anhand semantischer Eigenschaften bei Mel'čuk und Wanner (1994)

Die Methodik der Clusterverfahren unterscheidet sich prinzipiell von der Vorgehensweise bei Mel'čuk und Wanner (1994). Während durch das Clustering der Gefühlssubstantive anhand der Kookkurrenzdaten Klassen von Substantiven entstehen, deren semantischen Gemeinsamkeiten nach der Entstehung der Cluster zu bewerten sind, gehen Mel'čuk und Wanner den umgekehrten Weg und bilden Gruppen von Substantiven aufgrund der semantischen Eigenschaften der Substantive. Ein Substantiv kann dabei mehreren Klassen angehören, jede Klasse wird durch eine der elf semantischen Dimensionen gekennzeichnet (vgl. Kapitel 2.4.1). Für die Verben als Werte der lexikalischen Funktionen im generischen Eintrag der Gefühlssubstantive werden die semantischen Merkmale angegeben, die das spezifische Substantiv in der Argumentposition erfüllen muss, um als Kollokationspartner zu agieren, und für jede semantische Dimension werden spezifische Verben aufgezählt, mit denen die Mitglieder der Klasse kollokieren, oder mit denen sie aufgrund der semantischen Unvereinbarkeit des klassendefinierenden Merkmals der semantischen Dimension nicht auftreten können.
Bei der Untersuchung von Mel'čuk und Wanner handelt es sich um Angaben einzelner Verben, die als Kollokate bestimmter semantischer Dimensionen in Frage kommen oder

ausgeschlossen sind. Die Feststellung, dass Gefühlssubstantive, die das Merkmal "Active" tragen mit dem Verb *überwinden* kollokieren, kann eher in der Ausgabedatei von PECCI verifiziert werden, die alle Substantive für das portugiesische Verb *superar* auflistet. Bei einem stichprobenartigen Vergleich fällt auf, dass die postulierten Vereinbarkeiten semantischer Merkmale mit bestimmten Verben mit den Ergebnissen aus der Corpusuntersuchung nur teilweise übereinstimmen. Das Verb *superar* kommt im Portugiesischen mit Gefühlssubstantiven vor, für deren deutsche Übersetzungen auch bei Mel'čuk und Wanner das semantische Merkmal "Active" nicht verzeichnet ist, wie beispielsweise *desilusão* ('Enttäuschung').
Die Ergebnisse des Clusterverfahrens K-Means können mit der Klassifikation der Gefühlssubstantive bei Mel'čuk und Wanner verglichen werden, indem man nach semantischen Merkmalen sucht, die die Substantive teilen, die einem Cluster angehören. Unter den 226 lemmatisierten Verben, mit denen PECCI arbeitet, findet sich für jedes der von Mel'čuk und Wanner untersuchten 20 deutschen Verben mindestens ein portugiesisches Übersetzungsäquivalent. Das folgende Beispiel zeigt die Ausgabedateien nach der Berechnung von K-Means für 10 Cluster, die Clusterzentren wurden durch einen Zufallsgenerator ausge-wählt:

Ergebnis K-Means (Nomina)
Der Algorithmus wird 2 mal durchlaufen bis die Cluster stabil sind.

Das 1. Clusterzentrum wurde initiiert von admiração.
Zu Cluster 1 gehören: admiração ciúme compaixão paixão

Das 2. Clusterzentrum wurde initiiert von alegria.
Zu Cluster 2 gehören: alegria dor

Das 3. Clusterzentrum wurde initiiert von amor.
Zu Cluster 3 gehören: amor

Das 4. Clusterzentrum wurde initiiert von esperança.
Zu Cluster 4 gehören: encanto esperança estimação inclinação respeito

Das 5. Clusterzentrum wurde initiiert von fúria.
Zu Cluster 5 gehören: agitação alvoroço cólera comoção excitação fúria indignação ira

Das 6. Clusterzentrum wurde initiiert von inveja.
Zu Cluster 6 gehören: furor inveja

Das 7. Clusterzentrum wurde initiiert von medo.
Zu Cluster 7 gehören: medo vergonha

Das 8. Clusterzentrum wurde initiiert von pânico.
Zu Cluster 8 gehören: pânico

Das 9. Clusterzentrum wurde initiiert von susto.
Zu Cluster 9 gehören: susto

Das 10. Clusterzentrum wurde initiiert von tristeza.
Zu Cluster 10 gehören: aflição apreensão asco decepção desesperança desespero desilusão enfado entusiasmo luto ódio pena raiva tristeza

(Pecci: SentimentoCetemp/ClusterNomen/Ergebniskmeans.txt)

In einer weiteren Datei stehen die Vektoren der Clusterzentren, die Werte der Clusterzentren sind hier numerisch sortiert. Man kann zwischen zwei Ausgabevarianten wählen, die eine enthält alle Werte der im Vektor befindlichen Verben, die folgende nur die ersten zehn numerisch relevanten Verben:

Vektor von Clusterzentrum 1 sortiert:

0.2315045355814550	ter
0.0497258967052719	provocar
0.0485866976146525	haver
0.0427428911385748	suscitar
0.0345020522451824	sentir
0.0298066090531568	esconder
0.0284677746297968	estar
0.0271235774526219	despertar
0.0268429216700811	fazer
0.0231034469211426	mostrar

Vektor von Clusterzentrum 2 sortiert:

0.1262077914240020	ter
0.1164711175242570	dar
0.1077324437836220	sentir
0.0471992626101078	provocar
0.0436470075083190	fazer
0.0403098244584107	haver
0.0327183845455368	causar
0.0267457458803003	trazer
0.0256876681964947	sofrer
0.0224322344724288	estar

Vektor von Clusterzentrum 3 sortiert:

0.2973267866884890	fazer
0.0949263502454992	ter
0.0758319694489907	morrer
0.0490998363338789	perder
0.0425531914893617	haver
0.0212765957446809	viver
0.0201854882705947	dar
0.0201854882705947	estar
0.0169121658483361	sentir
0.0141843971631206	existir

Vektor von Clusterzentrum 4 sortiert:

0.2864556789129840	ter
0.0490223123676012	haver
0.0408614484066810	estar
0.0385797879630238	fazer
0.0341511215631274	perder
0.0279990547055012	dar
0.0275886772265252	manifestar
0.0224352628489156	mostrar
0.0221350762527233	merecer
0.0216364788354824	manter

Vektor von Clusterzentrum 5 sortiert:

0.2351537417943670	provocar
0.0621621511163400	causar
0.0435711850151899	ter
0.0384569134835875	haver
0.0328827737927859	estar
0.0290414977749258	suscitar
0.0257335232000931	manifestar
0.0235437736592715	criar
0.0218086899319511	sentir
0.0188376641917328	esconder

Vektor von Clusterzentrum 6 sortiert:

0.6177811550151980	fazer
0.0761822294479395	causar
0.0528557291298508	ter
0.0305718526896162	provocar
0.0281155015197568	roer
0.0121580547112462	corar
0.0113981762917933	olhar
0.0109210433307415	suscitar
0.0098784194528875	haver
0.0091185410334346	ficar

Vektor von Clusterzentrum 7 sortiert:

0.5838802550925640	ter
0.0471159492891893	sentir
0.0400782013685239	corar
0.0278055303827702	perder
0.0258317338451695	meter
0.0235926849449232	estar
0.0210787679907239	passar
0.0182435462294295	haver
0.0168040068820206	fazer
0.0096063101449765	considerar

Vektor von Clusterzentrum 8 sortiert:

0.2056338028169010	entrar
0.1183098591549300	provocar
0.1105633802816900	lançar
0.0598591549295775	causar
0.0549295774647887	semear
0.0485915492957746	estar
0.0457746478873239	gerar
0.0450704225352113	haver
0.0366197183098592	instalar
0.0338028169014084	viver

Vektor von Clusterzentrum 9 sortiert:

0.2658450704225350	apanhar
0.1971830985915490	ganhar
0.1338028169014080	pregar
0.0545774647887324	provocar
0.0457746478873239	passar
0.0352112676056338	ter
0.0334507042253521	sofrer
0.0281690140845070	haver
0.0211267605633803	recuperar
0.0176056338028169	morrer

Vektor von Clusterzentrum 10 sortiert:

0.0653677258848829	ter
0.0492846788104044	estar
0.0444072910798692	haver
0.0422179929613630	manifestar
0.0416482661119087	valer
0.0403772172098047	esconder
0.0385477847758705	sentir
0.0350354238731408	fazer
0.0333255535150067	provocar
0.0281371090626876	ficar

(Pecci: SentimentoCetemp/ClusterNomen/VektorenClusterzentrenSortShort.txt)

Startet man K-Means ein weiteres mal erhält man ein Ergebnis, das sich erheblich von dem ersten Ergebnis unterscheidet:

Ergebnis K-Means (Nomina)
Der Algorithmus wird 2 mal durchlaufen bis die Cluster stabil sind.

Das 1. Clusterzentrum wurde initiiert von alvoroço.
Zu Cluster 1 gehören: aflição alvoroço luto pânico

Das 2. Clusterzentrum wurde initiiert von amor.
Zu Cluster 2 gehören: amor furor inveja

Das 3. Clusterzentrum wurde initiiert von medo.
Zu Cluster 3 gehören: medo vergonha

Das 4. Clusterzentrum wurde initiiert von inclinação.
Zu Cluster 4 gehören: ciúme encanto inclinação

Das 5. Clusterzentrum wurde initiiert von decepção.
Zu Cluster 5 gehören: asco decepção desilusão tristeza

Das 6. Clusterzentrum wurde initiiert von fúria.
Zu Cluster 6 gehören: alegria apreensão desesperança desespero dor enfado entusiasmo fúria ódio pena raiva

Das 7. Clusterzentrum wurde initiiert von admiração.
Zu Cluster 7 gehören: admiração compaixão esperança estimação paixão respeito

Das 8. Clusterzentrum wurde initiiert von comoção.
Zu Cluster 8 gehören: agitação cólera comoção excitação indignação

Das 9. Clusterzentrum wurde initiiert von susto.
Zu Cluster 9 gehören: susto

Das 10. Clusterzentrum wurde initiiert von ira.
Zu Cluster 10 gehören: ira

(Pecci: SentimentoCetemp/ClusterNomen/Ergebniskmeans.txt)

Entsprechend variieren die Werte der Vektoren der Clusterzentren:

Vektor von Clusterzentrum 1 sortiert:

0.1224061727570280	estar
0.0802096797281423	provocar
0.0674661975585289	entrar
0.0561899871111902	causar
0.0523455150186064	haver
0.0491954180468943	fazer
0.0441893021434854	viver
0.0424028268551237	vestir
0.0352088176824959	ter
0.0303004195385076	lançar

Vektor von Clusterzentrum 2 sortiert:

0.5109630322396280	fazer
0.0668792695017336	ter
0.0507881529652930	causar
0.0298365936663809	morrer
0.0209267888634539	provocar
0.0207700101317123	haver
0.0187436676798379	roer
0.0163666121112930	perder
0.0120770976782852	estar
0.0101966591328293	sentir

Vektor von Clusterzentrum 3 sortiert:

0.5838802550925640	ter
0.0471159492891893	sentir
0.0400782013685239	corar
0.0278055303827702	perder
0.0258317338451695	meter
0.0235926849449232	estar
0.0210787679907239	passar
0.0182435462294295	haver
0.0168040068820206	fazer
0.0096063101449765	considerar

Vektor von Clusterzentrum 4 sortiert:

0.3693330262035700	ter
0.0482199308020573	provocar
0.0412782624669350	fazer
0.0358706805133977	haver
0.0338094960683596	mostrar
0.0323624595469256	render
0.0290168432551202	estar
0.0257900835031364	perder
0.0234215460903341	dar
0.0227336289903712	revelar

Vektor von Clusterzentrum 5 sortiert:		Vektor von Clusterzentrum 6 sortiert:	
0.0898919828167177	manifestar	0.0846711353018469	ter
0.0894141688800474	esconder	0.0505933145210620	valer
0.0765764105872276	sentir	0.0488106905419004	provocar
0.0668943115214725	ter	0.0419128333049501	haver
0.0665692568927138	ficar	0.0403652335142530	sentir
0.0521802325581395	invadir	0.0352932976456901	fazer
0.0507267441860465	olhar	0.0336744844775459	dar
0.0396639625581275	haver	0.0333086485287064	aumentar
0.0373651699768271	provocar	0.0299982160304998	estar
0.0352032646314376	sofrer	0.0272659971684292	causar

Vektor von Clusterzentrum 7 sortiert:		Vektor von Clusterzentrum 8 sortiert:	
0.2083829097133390	ter	0.2078780368541010	provocar
0.0553077184594038	haver	0.0746008636030977	causar
0.0385213017978719	estar	0.0444992311110584	haver
0.0325691597815301	suscitar	0.0369969085200824	ter
0.0306106478351238	merecer	0.0354196688661808	manifestar
0.0294059731824398	fazer	0.0344624281859131	suscitar
0.0280297022441318	manifestar	0.0313252884568398	estar
0.0228981094703381	sentir	0.0312870591314301	criar
0.0193750895222389	perder	0.0284011649515743	sentir
0.0187575113920574	esconder	0.0224021674686773	esconder

Vektor von Clusterzentrum 9 sortiert:		Vektor von Clusterzentrum 10 sortiert:	
0.2658450704225350	apanhar	0.5138888888888890	provocar
0.1971830985915490	ganhar	0.0451388888888889	suscitar
0.1338028169014080	pregar	0.0381944444444444	desencadear
0.0545774647887324	provocar	0.0381944444444444	enfrentar
0.0457746478873239	passar	0.0347222222222222	despertar
0.0352112676056338	ter	0.0312500000000000	acalmar
0.0334507042253521	sofrer	0.0277777777777778	motivar
0.0281690140845070	haver	0.0243055555555556	causar
0.0211267605633803	recuperar	0.0173611111111111	lançar
0.0176056338028169	morrer	0.0173611111111111	sofrer

(Pecci: SentimentoCetemp/ClusterNomen/VektorenClusterzentrenSortShort.txt)

Wie aus den beiden Beispielen hervorgeht, ist die Zusammensetzung der in einem bestimmten Cluster befindlichen Substantive in starkem Maße abhängig von der Wahl der Ausgangsclusterzentren. Neben den Abweichungen ist bei einer zweimaligen Initiierung des Algorithmus mit zufallsgenerierten Clusterzentren aber auch eine gewisse Stabilität zu beobachten, *medo* und *vergonha* befinden sich beispielsweise bei beiden Ergebnissen in einem Cluster, ebenso *agitação, cólera, comoção, excitação* und *indignação,* oder *decepção, desilusão* und *tristeza*.

Um eine relativ stabile Gruppierung der Substantive zu erzielen, wird im Benutzerdialog von PECCI zusätzlich die Möglichkeit angeboten, die gesammelten Ergebnisse von K-Means nach 100 Initiierungen des Algorithmus zu untersuchen (vgl. Kapitel 5.2, Anhang B1). Bei dieser Variante wird in der Ausgabedatei für jedes Substantiv aufgeführt, wie oft es mit welchen weiteren Substantiven in einem Cluster steht (Anhang C7 enthält die komplette Ausgabedatei):

> Nachdem K-Means 100 mal durchlaufen wurde, kommen die Nomina mit folgenden weiteren Nomina X-mal im gleichen Cluster vor:
>
> admiração:
> 100-admiração 63-paixão 58-compaixão 57-respeito 57-esperança 54-estimação 43-desilusão 36-raiva 33-dor 31-alegria 30-decepção 29-ódio 27-tristeza 19-ciúme 19-desesperança 14-

entusiasmo 11-encanto 11-pena 10-desespero 10-fúria 9-susto 9-asco 9-aflição 9-apreensão 8-luto 8-enfado 7-inclinação 6-medo 6-vergonha 3-indignação 2-amor 2-pânico 1-agitação 1-cólera 1-excitação 1-comoção

aflição:
100-aflição 82-desespero 59-luto 57-entusiasmo 55-susto 52-ódio 43-pena 42-tristeza 42-apreensão 41-desesperança 40-raiva 40-alegria 33-dor 27-asco 26-pânico 25-desilusão 22-decepção 17-enfado 15-alvoroço 14-compaixão 13-estimação 12-paixão 12-fúria 10-amor 9-respeito 9-esperança 9-indignação 9-admiração 8-agitação 8-cólera 7-excitação 6-furor 5-inveja 2-comoção 1-ira

agitação:
100-agitação 89-excitação 83-comoção 78-cólera 76-alvoroço 62-fúria 58-ira 47-pânico 46-indignação 34-enfado 10-susto 8-aflição 7-desespero 7-decepção 6-apreensão 5-luto 4-ódio 4-desilusão 4-entusiasmo 3-dor 2-tristeza 2-desesperança 2-pena 1-raiva 1-asco 1-paixão 1-alegria 1-compaixão 1-admiração

alegria:
100-alegria 86-raiva 65-ódio 64-dor 50-tristeza 47-desespero 45-entusiasmo 44-compaixão 41-paixão 40-susto 40-aflição 38-desilusão 37-pena 35-estimação 34-respeito 34-esperança 31-desesperança 31-admiração 29-asco 28-luto 27-apreensão 24-decepção 13-amor 9-enfado 7-fúria 4-furor 4-inveja 3-ciúme 3-pânico 2-encanto 1-alvoroço 1-agitação 1-medo 1-cólera 1-vergonha 1-excitação 1-comoção 1-indignação 1-inclinação

alvoroço:
100-alvoroço 76-agitação 72-excitação 71-cólera 66-comoção 60-pânico 60-fúria 48-ira 41-indignação 30-enfado 17-susto 15-desespero 15-aflição 9-luto 9-entusiasmo 9-apreensão 8-decepção 7-tristeza 7-ódio 6-dor 5-desesperança 5-desilusão 4-asco 4-pena 1-furor 1-raiva 1-alegria

amor:
100-amor 78-furor 76-inveja 13-luto 13-alegria 13-pena 10-raiva 10-asco 10-aflição 9-desesperança 8-desespero 7-ódio 6-tristeza 6-dor 6-paixão 6-compaixão 6-apreensão 5-respeito 5-susto 4-entusiasmo 4-estimação 3-desilusão 2-esperança 2-decepção 2-admiração 1-enfado 1-indignação

...

(Pecci: SentimentoCetemp/ClusterNomen/Ergebniskmeans100.txt)

Der Vektor für jedes Substantiv wird aus dem Mittelwert der 100 Vektoren der Zentroide der Cluster ermittelt, in denen das Substantiv nach jeder Berechnung von K-Means steht. In einem weiteren Verzeichnis sind auch die Cluster für jede der 100 Berechnungen von K-Means sowie deren Zentroidvektoren abgelegt. Diese Daten sind auf den beiliegenden CDs oder nach einer Installation von PECCI nach jedem Programmaufruf mit der Wahl dieses Moduls über den Benutzerdialog auf dem eigenen Rechner zu finden. Die Häufigkeit mit der ein spezifisches Substantiv mit den weiteren untersuchten Substantiven aus dem Sample in einem Cluster steht, variiert mit jeder Initiierung des Moduls. Die numerische Reihenfolge der im selben Cluster vorkommenden Substantive bleibt für die signifikanten Clusternachbarn konstant. Durch die Akkumulation der Ergebnisse in dieser Form nach einer gewissen Anzahl von Initiierungen des K-Means Algorithmus, führt die Berechnung von der Clusterbildung wieder zu den spezifischen Substantiven. Für jedes untersuchte Substantiv sind diejenigen Substantive, die ihm aufgrund der Ähnlichkeit in den Kookkurrenzdaten nahe stehen, in absteigender Reihenfolge verzeichnet.

Durch die Anwendung statistischer Verfahren auf die gesamten verbalen Kookkurrenzdaten der Substantive, werden Ergebnisse erzielt, in denen bedeutungsverwandte Substantive sehr häufig in demselben Cluster stehen. *Aflição* ('Trauer'), *apreensão* ('Sorge'), *desespero* ('Verzweiflung'), *luto* ('Trauer') und *tristeza* ('Traurigkeit') befinden sich aufgrund ihrer Kookkurrenzdaten oft in demselben Cluster, eine weitere Gruppe semantisch ähnlicher Substantive bilden *agitação* ('Aufregung'), *excitação* ('Erregung'), *cólera* ('Zorn'), *ira* ('Zorn')

und *fúria* ('Wut'). Daneben ist aber auch die Bündelung von Substantiven zu beobachten, die semantisch gesehen nicht sehr viele Gemeinsamkeiten aufweisen, wie beispielsweise *amor, inveja* und *furor*. Ihr häufiges miteinander Vorkommen in demselben Cluster ist auf das sehr frequente Kookkurrieren mit dem Verb *fazer* zurückzuführen, dessen Übersetzung ('machen') als Kollokat auch im Deutschen mit den genannten Substantiven (oder deren Derivaten) gebräuchlich ist (*Liebe machen, neidisch machen, Furore machen*[100]). In diesem Fall haben die Substantive nur eine geringe semantische Ähnlichkeit, wenn die Synonymie der ausschlaggebende Faktor für die Definition von semantischer Verwand-schaft ist.
Es gibt auf der einen Seite Substantive wie *medo* ('Angst') und *vergonha* ('Scham', 'Schande'), die sehr häufig in einem Cluster stehen - *medo* und *vergonha* kommen nach 100 Initiierungen von K-Means mit 10 Clustern mindestens 90 mal in demselben Cluster. Auf der anderen Seite existieren Substantive, die sich aufgrund ihrer Kookkurrenzdaten, mit keinem der weiteren untersuchten Substantive kongruent verhalten. Beispielsweise kommt *susto* ('Schreck') am häufigsten zusammen mit *desespero* ('Verzweiflung') vor, die beiden Substantive stehen bei 100 Initiierungen von K-Means aber nur ca. 50 mal in einem Cluster. Der Grund für das häufige Alleinstehen von *susto* in einem Cluster, oder dem weniger stringenten Auftreten mit bestimmten anderen Gefühlssubstantiven in einem Cluster, ist auf den prominentesten Kollokationspartner von *susto,* das Verb *apanhar* (vgl. Kapitel 6.2) zurückzuführen, welches für keines der anderen Gefühlssubstantive sehr gebräuchlich ist. Das Substantiv *pânico* verhält sich in ähnlicher Weise. Durch das hochfrequente Auftreten des Kollokats *entrar (em)* ('in Panik ausbrechen'), welches wie *apanhar* mit keinem der anderen Gefühlssubstantive häufig kookkurriert, steht *pânico* oft allein in einem Cluster oder mit unterschiedlichen Substantiven.
Interessant erscheint, dass sich die in Kapitel 6.3 beschriebenen Unterschiede zwischen *raiva* und *fúria*, die als synonyme portugiesische Übersetzung von *Wut* in den deutsch-portugiesischen Wörterbüchern zu finden sind, auch in der Zuordnung zu unterschiedlichen Clustern manifestieren. Während *fúria* eher die 'Wut' bezeichnet, die durch eine bestimmte Ursache ausgelöst wird, die im Satz die Subjektposition einnimmt, bezeichnet *raiva* bevorzugt die 'Wut', die ein belebtes Wesen in der Subjektposition des Satzes verspürt. Dementsprechend kommt *raiva* oft mit weiteren Substantiven in einem Cluster vor, die meistens "gefühlt" werden, wie *alegria* ('Freude'), *dor* ('Schmerz') oder *ódio* ('Hass'). *Fúria* hingegen steht mit Gefühlssubstantiven in einem Cluster, die bevorzugt "verursacht" werden, wie *cólera* ('Zorn'), *indignação* ('Empörung'), oder *excitação* ('Erregung').
Synonymie ist für die Bestimmung der semantischen Ähnlichkeit der Gefühlssubstantive offenbar nur ein Kriterium, das für die Beschreibung der Bedeutung maßgeblich ist. In dem Fall von *Wut* ist die Abbildung der semantischen Aktanten auf die syntaktische Struktur des Satzes ausschlaggebend für die portugiesische Übersetzung des deutschen Gefühls-substantivs. Erst durch die Kenntnis der semantischen Rolle, die das Nomen in der Subjektposition für das Gefühlssubstantiv spielt, kann entschieden werden, wie das deutsche Substantiv *Wut* ins Portugiesische zu Übersetzen ist.
In der Notation des FrameNet-Projekts wären die Substantive in der Subjektposition einmal "Experiencer" (*raiva*), das andere mal "Stimulus" (*fúria*) der *Wut.* Die Realisierung des Frame-Elements "Experiencer" oder "Stimulus" in der Subjektposition des Satzes führt zur Verwendung bestimmter Verben zur Verknüpfung des Subjekts mit dem Gefühlssubstantiv.

100 *Furor* wird nur in der Kombination mit *fazer* mit 'Furore' übersetzt, in anderen Kontexten bedeutet *furor:* a) Begeisterung b) Wut, Zorn.

Im Rahmen der *Meaning-Text-Theory* äußert sich diese Tatsache im Lexikoneintrag von *raiva* durch das Fehlen etlicher Werte für die lexikalische Funktion 'CausFunc', die für *fúria* verzeichnet sind. In der Lexikalischen Funktion $Oper_1$ hingegen, sind für *raiva* eine größere Anzahl von Verben präsent. Interessant erscheinen auch die Werte der log-likelihood für die jeweilige Kollokation. Die Aufnahme statistischer Werte, die auf die Gebräuchlichkeit einer Wortkombination hinweisen, würde die Wahl des passenden substantivischen Übersetzungsäquivalents erleichtern.

	fúria	log-likelihood	*raiva*	log-likelihood
$Oper_1$	ter	27	chorar	174
	ferver	33	sentir	104
	sentir	16	ter	32
			corar	30
			tremer	27
CausFunc	provocar	401	dar	24
	desencadear	58	meter	24
	despertar	37	causar	13
	suscitar	30	provocar	8
	acalmar	67		
	causar	24		
	aumentar	17		
	alimentar	14		
	pôr	11		

Die semantischen Dimensionen, die Mel'čuk und Wanner (1999) zur Beschreibung der 40 deutschen Gefühlssubstantive anführen, haben weder mit mit dem Konzept der Synonymie noch mit dem der semantischen Rollen Gemeinsamkeiten. Die 11 semantischen Dimensionen, die zur Beschreibung und Klassifizierung der Gefühlssubstantive dienen, stammen aus der Psychologie und werden dort für die Bestimmung von Bedeutung der Gefühle und Gefühlslexeme angewandt. Die Semantik eines Substantivs setzt sich zusammen aus der Bündelung der Werte (positiv, neutral, negativ), die die Relevanz jeder semantischen Dimension für jedes Substantiv festlegen. In diesem Schema unterscheiden sich synonyme Substantive mitunter deutlich in den Werten der semantischen Dimensionen, *Angst* und *Furcht* haben in drei semantischen Dimensionen verschiedene Werte. *Angst* und *Eifersucht* hingegen weichen nur im Wert einer semantischen Dimension voneinander ab. *Angst* und *Eifersucht* haben positive Werte für die Dimensionen "Directionality, Reactivity, Excitation, Self-control, Permanent" und negative Werte für die Dimensionen "Polarity und Manifestability", *Angst* trägt einen zusätzlichen positiven Wert für "Activity".
Mel'čuk und Wanner postulieren zwischen dem Kollokationsverhalten der Substantive und deren semantischen Eigenschaften, die mit den semantischen Dimensionen beschrieben werden, eine direkte Abhängigkeit: "lexemes with common restricted lexical co-occurrence also share semantic features" (1994: 88). Aufgrund der Ähnlichkeit in der semantischen Beschreibung von *Angst* und *Eifersucht* ist zu erwarten, dass die spezifischen Lexikoneinträge von *Angst* und *Eifersucht* nicht stark voneinander abweichen. Die semantische Dimension "Activity" ist als Bedingung für die Kombinierbarkeit der Werte der Lexikalischen Funktionen mit den spezifischen Gefühlssubstantiven im Lexikoneintrag von *Gefühl* nicht vorhanden (der Lexikoneintrag von *Gefühl* ist in Kapitel 2.4.1 abgebildet):

Angst, *fem*
Angst von X vor Y wegen Z, ...

Y = II
1. *vor* N_{dat} 2. *zu* V_{inf} 3. *daß* PROP

$IncepOper_1$	:	bekommen [$\sim_{acc}$]
$Caus_2Oper_1$	:	versetzen [N_{acc} in $\sim_{acc}$]
$CausContFunc_1$	:	schüren [in N_{dat} $\sim_{acc}$]
$\uparrow Caus_2Func_1$	:	einflößen [N_{dat} $\sim_{acc}$] , erregen, wecken [in N_{dat} $\sim_{acc}$]
$Caus_{(2)}Func_1$	:	machen [N_{dat} $\sim_{acc}$]

(Mel'čuk/Wanner 1999: 143)

Eifersucht, *fem*
Eifersucht von X auf Y wegen Z, ...

Y = II
1. *auf* N_{acc}

$Oper_1$	:	¬haben
$\uparrow Caus_2Func_1$	:	wecken [in N_{dat} $\sim_{acc}$] , ¬hervorrufen
$CausContFunc_1$	:	schüren [in N_{dat} $\sim_{acc}$]

(Mel'čuk/Wanner 1999: 144-145)

Der Pfeil-Operator gibt an, dass die aufgeführten Verben der Lexikalischen Funktion, den Verben des generischen Lexikoneintrags von *Gefühl* in der betreffenden Lexikalischen Funktion hinzuzufügen sind, die Negation gibt an, welche Verben des übergeordneten Lexikoneintrags zu tilgen sind.
Die Tatsache, dass die spezifischen Lexikoneinträge von *Angst* und *Eifersucht* nur zwei gemeinsame Kollokate aufweisen, widerspricht der von Mel'čuk und Wanner formulierten Korrelation von semantischen Werten und Kollokationsverhalten der Gefühlssubstantive. Die Abweichung in den Kollokaten ist nicht allein auf den unterschiedlichen Wert für die semantische Dimension "Activity" zurückzuführen. Die aufgeführten Verben kombinieren jeweils mit weiteren Gefühlssubstantiven unabhängig von dem Merkmal "Activity".
Der Lexikoneintrag von *Furcht z*eigt wesentlich mehr Gemeinsamkeiten mit dem Lexikoneintrag von *Angst,* obwohl sich die beiden Gefühlssubstantive in drei semantischen Dimensionen unterscheiden ("Manifestability", "Mentality", "Self-control"). Synonymie scheint auf das Kollokationsverhalten der Substantive einen größeren Einfluss zu haben, als die Beschreibung der Substantive mit semantischen Dimensionen.

Furcht, *fem*
Furcht von X vor Y wegen Z, ...

Y = II
1. *vor* N_{dat} 2. *zu* V_{inf} 3. *daß* PROP

$IncepOper_1$	:	bekommen [$\sim_{acc}$]
$Caus_2Oper_1$	:	versetzen [N_{acc} in $\sim_{acc}$]
$\uparrow Caus_2Func_1$	:	einflößen [N_{dat} $\sim_{acc}$] , wecken [in N_{dat} $\sim_{acc}$]
$CausContFunc_1$	:	schüren [in N_{dat} $\sim_{acc}$]

(Mel'čuk/Wanner 1999: 146-147)

Die Zone der syntaktischen Kombinatorik von *Furcht* und *Angst* ist identisch, und die beiden Gefühlssubstantive haben die gleiche propositionale Form. Die von Mel'čuk und Wanner formulierte These über die Korrelation der semantischen Eigenschaften von Lexemen und deren Kollokationsverhalten mag durchaus ihre Berechtigung haben, doch scheint ein Zusammenhang zwischen dem Kollokationsverhalten der Gefühlssubstantive und deren Bedeutungsbeschreibung durch semantische Dimensionen nicht unbedingt zu bestehen. Relevant für die Korrelation von Bedeutung und Kollokationsverhalten sind semantische Konzepte wie Synonymie und die Definition semantischer Aktanten, sowie die Subkategorisierungseigenschaften der Gefühlssubstantive.
Auch die Substantive *Entsetzen* und *Panik* unterscheiden sich nur in einer semantischen Dimension, *Entsetzen* erhält für "Mentality" einen positiven Wert, *Panik* verhält sich in dieser semantischen Dimension neutral. Die Lexikoneinträge der beiden Substantive weichen wiederum erheblich voneinander ab (im Lexikoneintrag von *Entsetzen* fehlt die syntaktische Zone). Das Merkmal "Mentality" ist als Bedingung für die Verwendung der Werte der lexikalischen Funktionen im Lexikoneintrag von *Gefühl* nicht vorhanden:

Entsetzen, *neut*
Entsetzen von X über Y ...

$Oper_1$	:	¬fühlen
$Magn + IncepOper_1$	:	¬geraten, ¬ausbrechen,
$fast\ FinFunc_0$	:	¬verfliegen
$Caus_2Func_1$	:	¬erregen

(Mel'čuk/Wanner 1999: 145)

Panik, *fem*
Panik von X wegen Y

Y = II
1. *vor* N_{dat}

$Oper_1$	:	¬empfinden
$IncepOper_1$	:	bekommen [$_{acc}$],
$Caus_2Oper_1$	:	versetzen [N_{acc} in ~$_{acc}$] , ¬hervorrufen
$fast\ FinFunc_0$	:	¬verfliegen
$CausContFunc_1$	:	schüren [in N_{dat} ~$_{acc}$]

(Mel'čuk/Wanner 1999: 150)

Das Kollokationsverhalten der Gefühlssubstantive steht daher in Abhängigkeit von einer komplexen Semantik der Substantive, für die - neben einer möglichen Beschreibung der Bedeutung durch semantische Dimensionen - Faktoren der semantischen Relationen, die zwischen den einzelnen Substantiven bestehen (Synonymie, Antonymie, Hyponymie, Hyperonymie), genauso zu berücksichtigen sind wie die Argumentstruktur der Substantive und deren Realisierung im Satz. Eine Korrelation zwischen der Semantik der Gefühlssubstantive und deren verbaler Kollokate würde der Kollokationsdefinition widersprechen. Der idiosynkratische Charakter der Kollokation unterscheidet sie von den freien Wortverbindungen, deren Kombinatorik durch die Semantik der Sprache festgelegt ist. In Kapitel 3.1.5. wurde der Begriff 'konzeptionelle Kollokationen' näher bestimmt und die Möglichkeit aufgezeigt, dass viele der vermeintlich rein idiomatisch geprägten Kollokationen auf semantische Regelmäßigkeiten zurückzuführen sind, die aufgrund ihrer Komplexität erst nach einer genauen Analyse der Sprachdaten festzustellen sind.

Untersucht man die von K-Means berechneten Cluster bezüglich der semantischen Eigenschaften der Substantive, ergibt sich ein heterogenes Bild. Zum einen entstehen Cluster von Substantiven, deren Gemeinsamkeiten in den Kookkurrenzdaten auf den ersten Blick auf semantische Ähnlichkeiten zurückzuführen sind. Die Synonyme *agitação, alvoroço* und *excitação* ('Aufregung') stehen häufig in einem Cluster. Diese Substantive teilen neben ihrer Bedeutungsähnlichkeit auch eine identische Argumentstruktur, wodurch sie sich von den beiden synonymen Übersetzungsäquivalenten für *Wut* (*raiva* und *fúria*) unterscheiden, die vornehmlich in verschiedenen Clustern vorkommen. Im *Langenscheidts Taschenwörter-buch* (2001) wird als weitere Übersetzungsmöglichkeit von *Aufregung* das Substantiv *aflição* aufgeführt, das jedoch sehr selten mit den anderen Übersetzungsäquivalenten von *Aufregung* in einem Cluster steht. Die beiden häufigsten Clusternachbarn von *aflição* sind *desespero* ('Verzweiflung') und *luto* ('Trauer'). Das Clusterverhalten von *aflição* weist darauf hin, dass die anderen Bedeutungen *Kummer* und *Schmerz* des polysemen Substantivs viel gebräuchlicher sind. Die Anwendung von Clusterverfahren auf die Kookkurrenzdaten der Substantive kann dem Lexikografen Bedeutungspräferenzen polysemer Substantive aufzeigen.

Die Polysemie vieler Substantive verhindert eine eindeutige Clusterbildung. Ein Substantiv mit stark voneinander abweichenden Bedeutungen wie *pena*, das sowohl 'Strafe' als auch 'Kummer, Leid' bedeuten kann, zeigt keine eindeutigen Clusterpräferenzen und steht häufig allein in einem Cluster oder mit unterschiedlichen Substantiven. Wie in Kapitel 6.3 demonstriert wurde, ist die Bedeutungsbestimmung polysemer Substantive gerade durch die Kollokate oder die spezifischen Subkategorisierungseigenschaften zu leisten. Die semantische Annotation eines Corpus hätte dem Rechnung zu tragen, indem sie die kookkurrierenden Wörter und die syntaktische Struktur zur Bedeutungsdifferenzierung der polysemen Substantive verwendet. Inwieweit die semantische Ähnlichkeit von Substantiven einen Einfluss auf das Kollokationsverhalten der Substantive hat, ist auf der Grundlage elektronisch verfügbarer Corpora erst zu entscheiden, wenn Corpora in geeigneter Größe vorliegen, die nicht nur syntaktische, sondern auch semantische Informationen zu den einzelnen Wörtern enthalten. Erst dann wären adäquate Anfragen an einen Corpus zu stellen, die der Komplexität und Granularität natürlicher Sprache Rechnung tragen könnten.

Literaturverzeichnis

Abney, Steven (1991): "Parsing by Chunks". In: Berwick, Robert / Abney, Steven / Tenny, Carol (eds.): *Principle-Based Parsing: Computation and Psycholinguistics.* Dordrecht, Kluwer Academic Publishers: 257-278.

Alencar, Leonel F. de (2002): "Der *Constructor* - ein interaktives Werkzeug für Recherchen in portugiesischen Korpora auf dem WWW". In: Pusch, Claus D. / Raible, Wolfgang (eds.): *Romanistische Korpuslinguistik: Korpora und gesprochene Sprache.* Tübingen, Gunter Narr: 147-154.

Alonso Ramos, Margarita (2000): "Critères heuristiques pour l'encodage des collocations au moyen de fonctions lexicales". In: Heid, Ulrich / Evert, Stefan / Lehmann, Egbert / Rohrer, Christian (eds.): *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000*. Stuttgart, IMS: 463-473.

Apresjan, Ju.D. / Boguslavsky, I.M. / Iomdin, L.L. / Tsinman, L.L. (2002): "Lexical Functions in NLP: Possible Uses". In: Klenner, Manfred (ed.): *Computational Linguistics for the New Millennium: Divergence or Synergy? Festschrift in Honour of Peter Hellwig on the Occasion of his 60th Birthday*. Frankfurt, Peter Lang: 55–72.

Athayde, Maria Francisca (2001): *Construções com verbo-suporte (Funktionsverbgefüge) do Português e do Alemão*. Cadernos do CIEG No. 1.

Athayde, Maria Francisca Queiroz de (2002): "O uso do artigo em construções com verbo-suporte do português". In: *Lusorama* 51/52: 58-84

Bahns, Jens (1996): *Kollokationen als lexikographisches Problem: Eine Analyse allgemeiner und spezieller Lernerwörterbücher des Englischen.* Tübingen, Max Niemeyer.

Bally, Charles (1951[2]): *Traité de stylistique français*. Band 1,2. Genf, Librairie Georg&Cie. (1909)

Bartsch, Sabine (2004): *Structural and Functional Properties of Collocations in English. A Corpus Study of Lexical and Pragmatic Constraints on Lexical Co-occurrence.* Tübingen, Gunter Narr.

Benson, Morton (1985): "Lexical combinability". *Papers in Linguistics* 18: 3-15.

Benson, Morton / Benson, Evelyn / Ilson, Robert (1986): *Lexicographic Description of English.* Amsterdam: Benjamins.

Biber, Douglas / Conrad, Susan / Reppen, Randi (1998): *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: CUP.

Böhmer, Heiner (1994): *Komplexe Prädikatsausdrücke im Deutschen und Französischen. Theoretische Aspekte, kontrastive Aspekte, Aspekte der Anwendung*. Frankfurt, Peter Lang.

Braasch, Anna / Olsen, Sussi (2000): "Formalised Representation of collocations in a Danish Computational Lexicon." In: Heid, Ulrich / Evert, Stefan / Lehmann, Egbert / Rohrer, Christian (eds.): *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000*. Stuttgart, IMS: 475-487.

Burger, Harald (2003[2]): *Phraseologie. Eine Einführung am Beispiel des Deutschen*. Berlin, Erich Schmidt. (1998)

Bußmann, Hadumod (1990): *Lexikon der Sprachwissenschaft*. Stuttgart, Kröner.

Butina-Koller, Ekaterina (2005): *Kollokationen im zweisprachigen Wörterbuch. Zur Behandlung lexikalischer Kollokationen in allgemeinsprachlichen Wörterbüchern des Sprachenpaares Französisch/Russisch*. Tübingen, Max Niemeyer.

Caro Cedillo, Ana (2004): *Fachsprachliche Kollokationen. Ein übersetzungsorientiertes Datenbankmodell Deutsch-Spanisch*. Tübingen, Gunter Narr.

Christiansen, Tom / Torkington, Nathan (1999): *Perl Kochbuch*. Köln, O'Reilly.

Church, Kenneth et al. (1991): "Using Statistics in Lexical Analysis". In: Zernik, Uri (ed.): *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. Hillsdale, NJ, Lawrence Erlbaum: 115-164.

Claveau, Vincent / L'Homme, Marie-Claude (2004). "Discovering Specific Semantic Relationships between Nouns and Verbs in a Specialized French Corpus". *Computerm 2004 , dans le cadre de Coling 2004, Université de Genève (Suisse).*

Côco, Ute Fragoso (2001): "Funktionsverbgefüge in Zeitungstexten: ein portugiesisch-spanischer Vergleich". In: Schönberger, Axel / Thielemann, Werner (eds.): *Neuere Studien zur lusitanistischen Sprachwissenschaft*. Frankfurt, Domus Editora: 11-100.

Coseriu, Eugenio 1967: "Lexikalische Solidaritäten". *Poetica* 1,3: 293-303.

Cowie, Anthony P. (1978): "The Place of Illustrative Material and Collocations in the Design of Learner's Dictionary". In: Strevens, Peter (ed.): *In Honour of A.S. Hornby*. Oxford, University Press*:* 127-139.

Detges, Ulrich (1996): *Nominalprädikate. Eine valenztheoretische Untersuchung der französischen Funktionsverbgefüge des Paradigmas "être Präposition Nomen" und verwandter Konstruktionen.* Tübingen, Max Niemeyer.

Dias, Gaël / Nunes, Sérgio (2001): "Combining Evolutionary Computing and Similarity Measures to Extract Collocations from Unrestricted Texts". In: *Proceedings of RANLP - 2001 (Recent Advances in NLP)* (Tzigov Chark, Bulgaria, 5-7 September 2001).

Döll, Cornelia / Hundt, Christine (2002): "Funktionsverbgefüge im Portugiesischen: komplexe sprachliche Einheiten zwischen Syntax und Lexikon". In: Große, Sybille / Schönberger, Axel (eds.): *Ex oriente lux. Festschrift für Eberhard Gärtner zu seinem 60. Geburtstag*. Frankfurt, Valentia: 145-170.

Dorr, Bonnie J. / Pamela W. Jordan / and John W. Benoit (1999): "A Survey of Current Research in Machine Translation". *Advances in Computers* 49: 1--68.

Dörre, Jochen / Gerstl, Peter / Seiffert, Roland (2001): "Volltextsuche und Text Mining". In: Carstensen, Kai-Uwe et al.: *Computerlinguistik und Sprachtechnologie*. Heidelberg, Spektrum Akademischer Verlag: 425-441.

Dras, Mark / Johnson, Mike (1996): "Death and Lightness: Using a Demographic Model to Find Support Verbs". In: *Proceedings of the Fifth International Conference on the Cognitive Science of Natural Language Processing*. Dublin, Irland.

Dunning, Ted (1993): "Accurate Methods for the Statistics of Surprise and Coincidence". *Computational Linguistics* 19(1): 61-74.

Evert, Stefan (2005a): *The Statistics of Word Cooccurrences: Word Pairs and Collocations.* PhD. Dissertation, Universität Stuttgart, IMS.

Evert, Stefan (2005b): *The CQP Query Language Tutorial*. http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/UsersCorner.html

Evert, Stefan (2005c last update): *Computational Approaches to Collocations.* http://www.collocations.de/

Evert, Stefan / Kermes, Hannah (2003): "Experiments on Candidate Data for Collocation Extraction". In: *Companion Volume to the Proceedings of the 10th Conference of The European Chapter of the Association for Computational Linguistics.* Budapest, Ungarn: 83-86.

Evert, Stefan / Krenn, Brigitte (2001): "Methods for the qualitative evaluation of lexical association measures". In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France: 188-195.

Fiker, Marcia Epstein / Foley, Stela (2004): "Os Corpora como Ferramentas para Solução de Problemas de Tradução de Colocações Verbais". *CROP - Revista da Área de Língua e Literatura Inglesa e Norte-Americana* 10. Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo: 65-112.

Firth, John Rupert [1935] (1964): "The Technique of Semantics". In: Firth, John Rupert: *Papers in Linguistics 1934-1951.* Oxford, University Press: 7-33. (Reprint) .

Firth, John Rupert (1957): "A Synopsis of Linguistic Theory 1930-1955". In: Firth, John Rupert: *Studies in Linguistic Analysis*. Oxford, Blackwell: 1-32.

Fontenelle, Thierry (1997): *Turning a Bilingual Dictionary into a Lexical-Semantic Database*. Tübingen, Max Niemeyer.

Freire, N.A. (1985): *Verbformen Portugiesisch zum Nachschlagen.* München, Max Hueber.

Grefenstette, Gregory / Teufel, Simone (1995): "Corpus-based Method for Automatic Identification of Support Verbs for Nominalizations". In: *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*. Dublin, Irland: 98-103.

Grossmann, Francis / Tutin, Agnès (2003): "Quelques pistes pour le traitement des collocations". In: Grossmann, Francis / Tutin, Agnès (eds.): *Les Collocations - analyse et traitement*. Amsterdam, Editions 'De Werelt': 5-21.

Halliday, M.A.K. (1961): "Categories of the Theory of Grammar". *Word* 17: 241-292.

Halliday, M.A.K. (1966): "Lexis as a Linguistic Level". In: Bazell, Charles Ernest / Catford, John / Halliday, M.A.K., Robins, R.H. (eds.): *In Memory of J.R. Firth.* London, Longman: 148-162.

Halliday, M.A.K. / McIntosh, Angus / Strevens, Peter (1964): *The Linguistic Sciences and Language Teaching*. London: Longman.

Hausmann, Franz Josef (1979): "Un dictionnaire des collocations est-il possible?" In: *Travaux de Linguistique et de Littérature* 17,1: 187-195.

Hausmann, Franz Josef (1984): "Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen". *Praxis des neusprachlichen Unterrichts* 31: 395-406.

Hausmann, Franz Josef (1985): "Kollokationen im Deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels". In: Bergenholtz, Henning / Mugdan, Joachim (eds.): *Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch 1984.* Tübingen, Max Niemeyer: 118-129.

Hausmann, Franz Josef (1988): "Grundprobleme des zweisprachigen Wörterbuchs". In: Hyldgaard-Jensen, K. / Zettersten, A. (eds.): *Symposium on Lexicography* III. Tübingen, Max Niemeyer: 137-154.

Hausmann, Franz Josef (1989): "Le dictionnaire de collocations". In: Hausmann, Franz Josef / Reichmann, Otto / Wiegand, Herbert / Zgusta, Ladislav (eds.): *Wörterbücher - Dictionaries - Dictionnaires. Ein internationales Handbuch zur Lexikographie* 1. Berlin, Walter de Gruyter: 1010-1019.

Hausmann, Franz Josef (1997): "Semiotaxis und Wörterbuch". In: Konerding, Klaus-Peter / Lehr, Andrea (eds.): *Linguistische Theorie und lexikographische Praxis*. Tübingen, Max Niemeyer: 171-179.

Hausmann, Franz Josef (2000): "Vorwort". In: *Dicionário Contextual Básico da Língua Portuguesa.* Pöll, Bernhard. Wien, Edition Praesens: I-IV.

Hausmann, Franz Josef (2004): "Was sind eigentlich Kollokationen?". In: Steyer, Kathrin (ed.): *Wortverbindungen - mehr oder weniger fest. (Institut für Deutsche Sprache, Jahrbuch 2003)* Berlin, Walter de Gruyter: 309-334.

Hausmann, Franz Josef (2005): "Lexicographie française et phraséologie". Manuskript, Beitrag zu: *Collocations, corpus, dictionnaires. Colloque international de Cologne sur les collocations (1.-2.7.2005).*

Heid, Ulrich (1994): "On Ways Words Work Together - Topics in Lexical Combinatories". In: Willy, Martin et al. (ed.): *Proceedings of the VIth EURALEX International Congress*. Amsterdam, Euralex: 226-257.

Heid, Ulrich (1996): „Using Lexical Functions for the Extraction of Collocations from Dictionaries and Corpora“. In: Wanner, Leo (ed.): *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam, John Benjamins: 115-146.

Heid, Ulrich (1997): *Zur Strukturierung von einsprachigen und kontrastiven elektronischen Wörterbüchern.* Tübingen, Max Niemeyer.

Heid, Ulrich (1998): "Towards a corpus-based dictionary of german noun-verb collocations". In: Fontenelle, Thierry et al. (eds.): *EURALEX 98 Proceedings. Actes from the 8th EURALEX International Congress.* Belgium, Université de Liège: 301-312.

Heid, Ulrich (2001): "Computergestützte Lexikographie". In: Carstensen, Kai-Uwe et al.: *Computerlinguistik und Sprachtechnologie*. Heidelberg, Spektrum Akademischer Verlag: 418-424.

Heid, Ulrich (2004): "On the presentation of collocations in monolingual dictionaries". In: Williams, G. / Vesser, S (eds.): *Proceedings of the 11th EURALEX International Congress*. Lorient, Frankreich: 729-738.

Heid, Ulrich (2005): "Corpusbasierte Gewinnung von Daten zur Interaktion von Lexik und Grammatik: Kollokation - Distribution - Valenz". In: Lenz, Friedrich / Schierholz, Stefan (2005): *Corpuslinguistik in Lexik und Grammatik*. Tübingen, Stauffenburg Verlag: 97-122.

Heid, Ulrich / Raab, Sybille (1989): "Collocations in Multilingual Generation". In: *Proceedings of the 4th Conference of the European Chapter of the Association for Computational Linguistics*. Manchester, Association for Computational Linguistics: 130-136.

Heid, Ulrich / Freibott, Gerhard (1990): "Zur Darstellung von Äquivalenten in einer terminologisch-lexikalischen Datenbank für Übersetzer und technische Autoren". In: Schaeder, Burkhard/ Rieger, Burghard (eds.): *Lexikon und Lexikographie*. Hildesheim, Georg Olms Verlag: 244-254.

Heid, Ulrich et al. (2000): "Software Demonstration: Computational linguistic tools for semi-automatic corpus-based updating of dictionaries". In: Heid, Ulrich / Evert, Stefan / Lehmann, Egbert / Rohrer, Christian (eds.): *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000*. Stuttgart, IMS: 183-195.

Heid, Ulrich / Säuberlich, Bettina / Debus-Gregor, Esther / Scholze-Stubenrecht, Werner (2004): "Tools for upgrading printed dictionaries by means of corpus-based lexical acquisition". In: *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*. Lissabon, Portugal: 419-422.

Helbig, Gerhard / Schenkel, Wolfgang (1983[7]): *Wörterbuch zur Valenz und Distribution deutscher Verben*. Tübingen, Max Niemeyer. (1969)

Herbst, Thomas / Klotz, Michael (2003): *Lexikografie*. Paderborn, Ferdinand Schöningh.

Heylen, Dirk / Maxwell, Kerry / Verhagen, Marc (1994): "Lexical Functions and Machine Translation". In: *Proceedings of the 13th International Conference on Computational Linguistics (COLING)*. Kyoto, Japan: 1240-1244.

Hill, Jimmie / Lewis, Michael (eds.) (2000): *Handbuch der wichtigsten englischen Kollokationen*. Stuttgart, Ernst Klett.

Hollós, Zita (2004): *Lernerlexikographie: syntagmatisch. Konzeption für ein deutsch-ungarisches Lernerwörterbuch.* Tübingen, Max Niemeyer.

Hundertmark-Santos Martins, Maria Teresa (1982): *Portugiesische Grammatik*. Tübingen, Max Niemeyer.

Hundt, Christine (1994a): *Untersuchung zur portugiesischen Phraseologie*. Wilhelmsfeld, Gottfried Egert.

Hundt, Christine (1994b): "Construções de verbo + substantivo: estrutura, semântica e posição dentro da fraseologia". In: Endruschat, Annett / Vilela, Mário / Wotjak, Gerd (eds.): *Verbo e estruturas frásicas. Actas do IV Colóquio Internacional de Linguística Hispânica*. Porto, Faculdade de Letras do Porto: 267- 275.

Jones, S. / Sinclair, John (1973): "English Lexical Collocations: A Study in Computational Linguistics". *Cahiers de Lexicologie* 24: 15-61.

Jurafsky, Daniel / Martin, James (2000): *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. New Jersey, Prentice Hall.

Kermes, Hannah / Heid, Ulrich (2003): "Using chunked corpora for the acquisition of collocations and idiomatic expressions". In: Kiefer, Ferenc / Pajzs, Júlia (eds.): *Proceedings of Complex 2003. 7th Conference on Computational Lexicography and Text Research*. Budapest, Linguistics Institute - Hungarian Academy of Sciences.

Klotz, Michael (2000): *Grammatik und Lexik: Studien zur Syntagmatik englischer Verben.* Tübingen, Stauffenburg.

Köster, Lutz / Neubauer, Fritz (2002): "Kollokationen und Kompetenzbeispiele im DE GRUYTER WÖRTERBÜCH DEUTSCHH ALS FREMMDSPRACHE". In: Wiegand, Herbert Ernst (ed.): *Perspektiven der pädagogischen Lexikographie des Deutschen* II. Tübingen, Max Niemeyer: 283-310.

Krenn, Brigitte (2000): *The Usual Suspects: Data-Oriented Models for the Identification and Representation of Lexical Collocations*. PhD Dissertation, Universität des Saarlandes, Saarbrücken.

Krenn, Brigitte (2004): "Manual zur Identifikation von Funktionsverbgefügen und figurativen Ausdrücken in PP-Verb Listen". http://www.ofai.at/publications.html

Krenn, Brigitte / Evert, Stefan (2001): "Can we do better than frequency? A case study on extracting PP-verb collocations". In: *Proceedings of the ACL Workshop on Collocations*, Toulouse, France: 39-46.

Krenn, Brigitte / Evert, Stefan / Zinsmeister, Heike (2004): "Determining Intercoder Agreement for a Collocation Identification Task". In: Buchberger, E. (ed.): *Proceedings of Konvens 2004.* Wien, OEGAI: 89-96.

Kunze, Claudia (2001): "Lexikalisch-semantische Wortnetze". In: Carstensen, Kai-Uwe et al.: *Computerlinguistik und Sprachtechnologie*. Heidelberg, Spektrum Akademischer Verlag: 386-393.

Leffa, Vilson J. (1997): "Solving lexical ambiguity through collocation". In: *1st Annual Convention and Exposition of TESOL* (Orlando, EUA, 11-15 de Março de 1997): pp. 139.

Lehr, Andrea (1996): *Kollokationen und maschinenlesbare Korpora. Ein operationales Analysemodell zum Aufbau lexikalischer Netze.* Tübingen, Max Niemeyer.

Lehr, Andrea (1998): "Kollokationen in LANGENSCHEIDTS GROSSWÖRTERBUCH DEUTSCH ALS FREMDSPRACHE". In: Wiegand, Herbert Ernst (ed.): *Perspektiven der pädagogischen Lexikographie des Deutschen* I. Tübingen, Max Niemeyer: 257-281.

Lemnitzer, Lothar (1997): *Akquisition komplexer Lexeme aus Textkorpora.* Tübingen: Niemeyer.

L'Homme, Marie-Claude/ Bertrand, Claudine (2000): "Specialized Lexical Combinations: Should they be described as Collocations or in Terms of Selectional Restrictions?". In: Heid, Ulrich / Evert, Stefan / Lehmann, Egbert / Rohrer, Christian (eds.): *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000*. Stuttgart, IMS: 497-506.

Louro, Inês da Conceição dos Anjos (2001): *'Enxergando' as colocações: para ajudar a vencer o medo de um texto autêntico*. São Paulo, Universidade de São Paulo, Faculdade de Filosofia, Letras e Ciências Humanas (FFLCH). http://www.teses.usp.br/teses/disponiveis/8/8147/tde-22112001-000335/

Manning, Christopher / Schütze, Hinrich (2002[5]): *Foundations of Statistical Natural Language Processing*. Cambridge, MIT Press. (1999)

Martin, Willy (1992): "Remarks on Collocations in Sublanguage". In: *Terminologie et Traduction* 2-3: 157-164.

McKeown, Kathleen / Radev, Dragomir (2000): "Collocations". In: Dale, Robert / Moisl, Hermann / Somers, Harold (eds.): *Handbook of Natural Language Processing*. New York, Marcel Dekker.

Mel'čuk, Igor (1974): *Opyt teorii lingvističeskix modelej "Smysl <=> Text"*. Moskau, Nauka.

Mel'čuk, Igor (1996): "Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon". In: Wanner, Leo (ed.): *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam, John Benjamins: 37-102.

Mel'čuk, Igor (1997): *Vers une linguistique Sens-Texte. Leçon inaugurale.* Paris, Collège de France. http://www.olst.umontreal.ca/melcuk/#anchorPublications

Mel'čuk, Igor (1998): "Collocations and Lexical Functions". In: Cowie, Anthony P. (ed.): *Phraseology. Theory, Analysis, and Applications*. Oxford, University Press: 23-52.

Mel'čuk, Igor / Wanner, Leo (1994): „Lexical Co-occurrence and Lexical Inheritance. Emotion Lexemes in German: A Lexicographic Case Study". *Lexikos* 4: 86-161.

Meyer, Ingrid / Mackintosh, Kristen (1994): "Phraseme Analysis and Concept Analysis: Exploring a Symbiotic Relationship in the Specialized Lexicon". In: Willy, Martin et al. (ed.): *Proceedings of the VIth EURALEX International Congress*. Amsterdam, Euralex: 339-348.

Noll, Volker (1999): *Das brasilianische Portugiesisch. Herausbildung und Kontraste*. Heidelberg, C. Winter.

Oksefjell, Signe / Santos, Diana (1998): "Breve panorâmica dos recursos de português mencionados na Web". In: *Anais do Terceiro Encontro de Processamento da Língua Portuguesa (Escrita e falada), PROPOR'98* (Porto Alegre, 3-4 November 1998): 38-47.

Orenha, Adriane (2004a): "Aplicações léxico-terminográficas da lingüística de corpus: relato da elaboração de um glossário bilíngüe de colocações na área de negócios". *Intercâmbio* 13: 1-8.

Orenha, Adriane (2004b): *A compilação de um glossário bilíngüe de colocações, na área de negócios, baseado em corpus comparável*. São Paulo, Universidade de São Paulo, FFLCH, Dissertação de Mestrado.

Pereira, Fernando / Tishby, Naftali / Lee, Lillian (1993): "Distributional Clustering of English Words". In: *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics:* 183-190.

Pöll, Bernhard (1996): *Portugiesische Kollokationen im Wörterbuch: ein Beitrag zur Lexikographie und Metalexikographie.* Bonn, Romanistischer Verlag.

Polguère, Alain (2000): "Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French". In: Heid, Ulrich / Evert, Stefan / Lehmann, Egbert / Rohrer, Christian (eds.): *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000*. Stuttgart, IMS: 517-527.

Pusch, Claus D. / Raible, Wolfgang (eds.) (2002): *Romanistische Korpuslinguistik: Korpora und gesprochene Sprache.* Tübingen, Gunter Narr.

Pusch, Claus D. / Kabatek, Johannes / Raible, Wolfgang (eds.) (2002): *Romanistische Korpuslinguistik: Korpora und diachrone Sprachwissenschaft* II. Tübingen, Gunter Narr.

Quasthoff, Uwe (1998): "Projekt Der Deutsche Wortschatz." In: Heyer, Gerhard / Wolff, Christian (eds.). *Linguistik und neue Medien*. Wiesbaden: 93-99.

Rothkegel, Annely (1969): "Funktionsverbgefüge als Gegenstand maschineller Sprachanalysen". *Beiträge zur Linguistik und Informationsverarbeitung* 17: 7-26.

Rothkegel, Annely (1973): *Feste Syntagmen. Grundlagen, Strukturbeschreibung und automatische Analyse*. Tübingen, Max Niemeyer.

Sardinha, Tony Berber (1999): "Padrões lexicais e colocações do português". (Trabalho apresentado no Simpósio Processamento Computacional do Português, 9). São Paulo, PUCSP. http://www2.lael.pucsp.br/~tony/cursos/grandeimprensa/bibl/imprensa/ementa.pdf

Sardinha, Tony Berber (2004): *Lingüística de Corpus*. São Paulo, Editora Manole.

Sardinha, Tony Berber (2005): *A Língua Portuguesa no Computador*. Campinas, Mercado de Letras.

Schemann, Hans (1981): *Das idiomatische Sprachzeichen. Untersuchung der Idiomatizitätsfaktoren anhand der Analyse portugiesischer Idioms und ihrer deutschen Entsprechungen.* Tübingen, Max Niemeyer.

Scherfer, Peter (2001): "Zu einigen wesentlichen Merkmalen lexikalischer Kollokationen". In: Lorenz-Bourjot, Martine / Lüger, Heinz-Helmut (eds.): *Phraseologie und Phraseodidaktik*. Wien, Edition Praesens: 3-19.

Schickinger, Thomas / Steger, Angelika (2001): *Diskrete Strukturen 2: Wahrscheinlichkeitstheorie und Statistik*. Berlin, Springer-Verlag.

Schmid, Helmut (2005): *Statistische Methoden in der Maschinellen Sprachverarbeitung.* http://www.ims.uni-stuttgart.de/~schmid/ParsingII/

Schmidt-Radefeldt, Jürgen (2000): "Zweisprachige Lexikographie Portugiesisch-Deutsch/Deutsch-Portugiesisch". In: Wiegand, Herbert Ernst (ed.): *Studien zur zweisprachigen Lexikographie mit Deutsch V*. Hildesheim, Georg Olms: 215-226.

Schulte im Walde, Sabine et al. (2001): "Statistical Grammar Models and Lexicon Acquisition". In: Rohrer, Christian / Rossdeutscher, Antje / Kamp, Hans (eds.): *Linguistic Form and its Computation.* Stanford, CSLI Publications: 387-440.

Seretan, Violeta / Nerima, Luka / Wehrli, Eric (2004): "A Tool for Muli-Word Collocation Extraction and Visualization in Multilingual Corpora". In: *Proceedings of the Eleventh EURALEX International Congress (EURALEX 2004)*. Lorient, Frankreich: 755-766.

Seretan, Violeta / Nerima, Luka / Wehrli, Eric (2003): "Extraction of Multi-Word Collocations Using Syntactic Bigram Composition". In: *Proceedings of the Fourth International Conference on Recent Advances in NLP (RANLP-2003)*. Borovets, Bulgarien: 424-431.

Siepmann, Dirk (2005): "Collocation, Colligation and Encoding Dictionaries. Part I: Lexicological Aspects". *International Journal of Lexicography* 18(4): 409-443.

Silva, Jaime F. da (1994): "Zum Stand der zweisprachigen Lexikographie Deutsch-Portugiesisch / Portugiesisch-Deutsch: allgemeinsprachliche Äquivalenzwörterbücher". In: Figge, Udo L. (ed.): *Portugiesische und portugiesisch-deutsche Lexikographie.* Tübingen, Max Niemeyer: 67-85.

Sinclair, John (1966): "Beginning the Study of Lexis". In: Bazell, Charles / Catford, John / Halliday, M.A.K., Robins, R.H. (eds.): *In Memory of J.R. Firth.* London, Longman: 410-430.

Smadja, Frank (1993): "Retrieving Collocations from Text: Xtract". *Computational Linguistics* 19(1): 143-177.

Sommerfeldt, Karl-Ernst / Schreiber, Herbert (1983[3]): *Wörterbuch zur Valenz und Distribution deutscher Adjektive*. Tübingen, Max Niemeyer. (1974)

Sommerfeldt, Karl-Ernst / Schreiber, Herbert (1983[3]): *Wörterbuch zur Valenz und Distribution deutscher Substantive*. Tübingen, Max Niemeyer. (1977)

Stadler, Heike (1996): *Das doppelte Partizip im Portugiesischen. Der Wortartwechsel als transkategorialer Prozeß.* Magisterarbeit, Freie Universität Berlin, Institut für Romanische Philologie.

Steyer, Kathrin (2000): "Usuelle Wortverbindungen des Deutschen. Linguistisches Konzept und lexikografische Möglichkeiten". *Deutsche Sprache* 28: 101-125.

Steyer, Kathrin (2004): "Kookkurrenz, Korpusmethodik, linguistisches Modell, lexikografische Perspektiven". In: Steyer, Kathrin (ed.): *Wortverbindungen - mehr oder weniger fest. (Institut für Deutsche Sprache, Jahrbuch 2003)* Berlin, Walter de Gruyter: 87-116.

Tagnin, Stella E.O. (1989): *Expressões Idiomáticas e Convencionais.* São Paulo, Urtica.

Tagnin, Stella E.O. (1999a): *Convencionalidade e Produção de Texto: um Dicionário de Colocações Verbais Inglês/Português, Português/Inglês.* Tese de Livre-Docência, Universidade de São Paulo.

Tagnin, Stella E.O. (1999b): "Collecting data for a bilingual dictionary of verbal collocations: from scraps of paper to corpora research". In: Lewandowska-Tomaszczyk, Barbara (ed.): *PALC - Practical Applications in Language Corpora.* Lodz, University Press: 399-407.

Tagnin, Stella E.O. (2002): "The Brazilian Lexicographic Road to Bilingual Verbal Collocations". In: Braasch, Anna (ed.): *Proceedings of the Tenth EURALEX International Congress*. Copenhagen, Center for Sprogteknologi: 735-740.

Wall, Larry / Christiansen, Tom / Orwant, Jon (2001): *Programmieren mit Perl*. Köln, O'Reilly.

Wanner, Leo (1996): "Introduction". In: Wanner, Leo (ed.): *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam, John Benjamin: 1-36.

Wanner, Leo (2004): "Towards automatic fine-grained semantic classification of verb-noun collocations". *Natural Language Engineering* 10(2): 95-143.

Wanner, Leo / Bohnet, Bernd / Alonso, Margarita / Vázques, Nancy (2005a): "The *True, Deep Happiness:* Towards the Automatic Semantic Classification of Adjective-Noun Collocations". Kiefer, Ferenc / Kiss, Gábor / Pajzs, Júlia (eds.): *Papers in Computational Lexicography. COMPLEX 2005.* Budapest, Linguistics Institute - Hungarian Academy of Sciences: 255- 265

Wanner, Leo / Bohnet, Bernd / Giereth, Mark / Vidal, Vanesa (2005b): "The first steps towards the automatic compilation of specialized collocation dictionaries". In: *Terminology* 11(1): 143-180.

Welker, Herbert A. (2002): "Die Behandlung von Phraseologismen in einem deutsch-portugiesischen Wörterbuch der deutschen Verben (und in einigen anderen zweisprachigen Wörterbüchern)". *Zeitschrift für romanische Philologie* 118: 392-429. = Welker, Herbert (2002a): "A apresentação de fraseologismos num dicionário alemão-português de verbos (e em seis outros dicionários)". http://www.unb.br/il/let/welker/fraseo.doc

Woll, Dieter (1990): "Portugiesische Lexikographie". In: Hausmann, Franz Josef / Reichmann, Otto / Wiegand, Herbert / Zgusta, Ladislav (eds.): *Wörterbücher - Dictionaries - Dictionnaires. Ein internationales Handbuch zur Lexikographie* 2. Berlin, Walter de Gruyter: 1723 – 1736.

Yuan, Jie (1986): *Funktionsverbgefüge im heutigen Deutsch: Eine Analyse und Kontrastierung mit ihren chinesischen Entsprechungen*. Heidelberg, Julius Groos.

Zinsmeister, Heike / Heid, Ulrich (2002): "Collocations of Complex Words: Implications for the Acquisition with a Stochastic Grammar". In: *Proceedings of the International Workshop on Computational Approaches to Collocations.* Wien.

Zinsmeister, Heike / Heid, Ulrich (2003): "Significant Triples: Adjective+Noun+Verb Combinations". In: Kiefer, Ferenc / Pajzs, Júlia (eds.): *Proceedings of Complex 2003. 7th Conference on Computational Lexicography and Text Research*. Budapest, Linguistics Institute - Hungarian Academy of Sciences: 92-101.

Zinsmeister, Heike / Heid, Ulrich (2004): "Collocations of Complex Nouns: Evidence for Lexicalisation". In: Williams, G. / Vesser, S (eds.): *Proceedings of the 11th EURALEX International Congress*. Lorient, Frankreich.

Print-Wörterbücher

The BBI Combinatory Dictionary of English: A Guide to Word Combinations. (1986) Benson, Morton / Benson, Evelyn / Ilson, Robert (eds.). Amsterdam, Benjamins. *BBI*

Collins COBUILD English Language Dictionary. (1987) Sinclair, John (ed.). London, Collins.

Collins COBUILD English Dictionary. (1995) Sinclair, John (ed.). London, Collins.

Dicionário Contextual Básico da Língua Portuguesa. Portugiesisches Kontext-wörterbuch. (2000) Pöll, Bernhard. Wien, Edition Praesens.

Dicionário de Alemão-Português. (1989) Schau, Udo. Porto, Porto Editora.

A Dictionary of English Collocations: Based on the Brown Corpus. I-III. (1994) Kjellmer, Göran (ed.). Oxford, Clarendon Press.

Dictionnaire combinatoire du français - Expressions, locutions et constructions. (2003) Zinglé, Henri / Brobeck-Zinglé, Marie-Louise. Paris: La Maison di Dictionnaire.

Dictionnaire explicatif et combinatoire du français contemporain: recherches lexico-sémantiques. I-IV. (1984, 1988, 1992, 1999) Mel'čuk, Igor et al. Montréal, Presses de l'Université de Montréal. *DEC*

Duden. Das Fremdwörterbuch. Duden Band 5. (2005). Mannheim, Dudenverlag.

Idiomatik Deutsch-Portugiesisch. Dicionário Idiomático Alemão-Português. Pons. (2002) Schemann, Hans et al. Stuttgart, Klett.

Langenscheidts Großwörterbuch Deutsch als Fremdsprache. (2003) Langenscheidt-Redaktion. Berlin, Langenscheidt.

Langenscheidts Taschenwörterbuch Portugiesisch. Portugiesisch-Deutsch/Deutsch-Portugiesisch. (2001) Langenscheidt-Redaktion. Berlin, Langenscheidt.

Longman Dictionary of Contemporary English. (2003) Summers, Della (ed.). Harlow: Pearson.

Novo Dicionário Aurélio da Língua Portuguesa. (1986) Ferreira, Aurélio Buarque de Holanda. Rio de Janeiro, Editora Nova Fronteira. *Aurélio*

Oxford Collocations Dictionary for Students of English. (2002). Oxford, University Press. *OCD*

Oxford Dictionary of Current Idiomatic English 1: Verbs with Prepositions & Particles.(1975) Cowie, A.P. / Mackin, R.. Oxford, University Press. *ODCIE1*

Oxford Dictionary of Current Idiomatic English 2: Phrase, Clause & Sentence Idioms. (1983) Cowie, A.P. / Mackin, R. / McCaig, I.R.. Oxford, University Press. *ODCIE2*

PONS Standardwörterbuch. Portugiesisch-Deutsch/Deutsch-Portugiesisch. (2002)Bearbeitet von: Mafalda, Joana / Seixas, Pimentel / Weber, Antje. Barcelona, Ernst Klett Sprachen.

Elektronische Wörterbücher

Diccionario de Colocaciones del Español *DiCE*
http://www.dicesp.com
Dictionnaire des Collocations
http://www.tonitraduction.net/
Dictionnaire en Ligne de Combinatoire du Français *DiCouèbe*
http://olst.ling.umontreal.ca/dicouebe/
Digitales Wörterbuch der deutschen Sprache des 20. Jahrhunderts
http://www.dwds.de/
eldit - Elektronisches Wörterbuch Deutsch-Italienisch
http://dev.eurac.edu:8081/MakeEldit1/Eldit.html
elexiko
http://www.elexiko.de/
LEO Wörterbücher
http://dict.leo.org/
Le Trésor de la Langue Française informatisé
http://atilf.atilf.fr/tlf.htm
Wortschatz-Portal der Universität Leipzig
http://wortschatz.uni-leipzig.de/

URLs zu Corpora, Tools und Projekten

Corpora

British National Corpus http://www.natcorp.ox.ac.uk/
Brown Corpus http://clwww.essex.ac.uk/w3c/corpus_ling/content/corpora/list/private/brown/brown.html
Cetempúblico http://www.linguateca.pt/CETEMPublico
Cetenfolha http://www.linguateca.pt/CETENFolha
COSMAS I Corpus http://www.ids-mannheim.de/kt/projekte/cosmas_I/
IMS Corpus Workbench http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

Tools für die linguistische Corpusannotation

Deutscher Chunker http://www.ims.uni-stuttgart.de/tcl/SOFTWARE/German-Chunker.html
LoPar Parser http://www.ims.uni-stuttgart.de/tcl/SOFTWARE/LoPar.html
Stuttgart-Tübinger Tagset http://www.sfs.uni-tuebingen.de/Elwis/stts/stts.html
Tree-Tagger http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
VISL http://visl.sdu.dk/visl/pt

Baumbanken, semantisch annotierte Corpora, Wortnetze

EuroWordNet http://www.illc.uva.nl/EuroWordNet/
FrameNet framenet.icsi.berkeley.edu/
Penn Treebank http://www.cis.upenn.edu/~treebank/
SALSA http://www.coli.uni-saarland.de/projects/salsa/
TIGER http://www.ims.uni-stuttgart.de/projekte/TIGER/
WordNet http://wordnet.princeton.edu/

Projekte

COBUILD http://www.collins.co.uk/books.aspx?group=140
Linguateca http://www.linguateca.pt
Observatoire de linguistique Sens-Texte http://www.olst.umontreal.ca/dicoeng.html
Transferbereich 32 IMS http://www.ims.uni-stuttgart.de/projekte/TFB/projekt.shtml

```
###############
# FILE: Cetemp
# LANGUAGE: Perl
# AUTHOR: Heike Stadler
# PURPOSE: Corpusaufbereitung CetemPublico1.7 für Pecci

#!/usr/bin/perl
use locale;

#### CetemPublico1.7 wird in Zeilen geschrieben

open (IN, "+< Corpora/Cetempublico/CetemPublico1.7");
open (OUT, "> Corpora/Cetempublico/ablage.txt");

while (<IN>) {

   $a = substr ($_,0,-1);
   if ($a =~ /<ext/) {print OUT "$a";}
   elsif ($a =~ /<p>/<\/p>/) {print OUT "\n$a";}
   elsif ($a =~ "</ext>") {print OUT "\n$a\n";}
   elsif ($a =~ "<s>|<a>|<t>|<li>") {print OUT "\n$a ";}
   else {print OUT "$a ";}
}

close (IN); close (OUT);

#### Die Wörter und bestimmte SGML-Tags werden gezählt und die aktuelle Wortanzahl
#### an den Satzanfang gestellt

open (IN, "+< Corpora/Cetempublico/ablage.txt");
open (OUT, "> Corpora/Cetempublico/Cetempublico");

while (<IN>) {

   @zeile = split;

   if ($zeile[$#zeile] !~ "</s>|</a>|</t>|</li>") {print OUT "@zeile\n";}

   else {

      if ($zeile[$#zeile] =~ "</s>") {$allesaetze++;}
      if ($zeile[$#zeile] =~ "</a>") {$alleautoren++;}
      if ($zeile[$#zeile] =~ "</t>") {$alletitel++;}
      if ($zeile[$#zeile] =~ "</li>") {$alleaufz++;}

      for $j (0..$#zeile) {
         if (($zeile[$j] !~ "<*>") && ($zeile[$j] =~ /[A-Za-zÀ-ü0-9]/))
            {$alleworte++;}
      }

      for $j (0..$#zeile) {
         if (($zeile[$j] =~ "<*>") && ($zeile[$j] !~ /\//)) {
            print OUT $zeile[$j]; print OUT " $alleworte:";}
         else {print OUT " $zeile[$j]";}
      }

      print OUT "\n";
   }
}
```

```
close (IN); close (OUT);

open (OUTW, "> Corpora/Cetempublico/wortanzahl.txt");
print OUTW "$alleworte Wörter $allesaetze Sätze $alletitel Titel $alleautoren
Autoren $alleaufz Aufzählungen\n";
close (OUTW);

system  "rm Corpora/Cetempublico/ablage.txt";
```

```
###############
# FILE: Cetenf
# LANGUAGE: Perl
# AUTHOR: Heike Stadler
# PURPOSE: Corpusaufbereitung CETENFolha-1.0 für Pecci

#!/usr/bin/perl
use locale;

#### Analog zum Punkt werden die weiteren Satzzeichen tokenisiert

@arraq = (',', ':','!', '»', '\?', '\)', '\.\.\.');
@ersatz = (',', ':','!', '»', '?', ')', '...');

system "cp Corpora/Cetenfolha/CETENFolha-1.0 Corpora/Cetenfolha/ablage.txt";

foreach $satzzeichen(@arraq) {

   open (IN, "+< Corpora/Cetenfolha/ablage.txt");
   open (OUT, "> Corpora/Cetenfolha/ablage1.txt");

   while (<IN>) {
     @zeile = split;
     if ($zeile[0] =~ /<s/) {
        @zeile1 = "";
        for ($j=0, $k=0; $j<=$#zeile; $j++, $k++) {
           $zeile1[$k]= $zeile[$j];
           if ($zeile1[$k] =~ /$satzzeichen/) && ($zeile1[$k] !~ /[0-9](:|,)[0-
9]/)){
              if ($satzzeichen =~ "...")
                 {$zeile1[$k] = substr ($zeile1[$k],0,-3);}
              else {$zeile1[$k] = substr ($zeile1[$k],0,-1);}
              for $i(($k+1)..$#zeile1) {
                 $zeile2[$i] = $zeile1[$i];
              }
              $zeile1[$k+1] = $ersatz[$i]; $k++;
              push @zeile1,@zeile2;
              @zeile2 = "";
           }
        }
        for $j (0..$#zeile1) {
           print OUT "$zeile1[$j] ";
        }
        print OUT "\n";
     }
     else {print OUT "@zeile\n";}
   }

   close (IN);  close (OUT);
   system "mv Corpora/Cetenfolha/ablage1.txt Corpora/Cetenfolha/ablage.txt";
   $i++;
}

#### Tokenisierung von Satzzeichen, die vor dem Wort stehen

@arraq = ('«', '\(');
@ersatz = ('«', '(');
$i=0;
```

```
foreach $satzzeichen(@arraq) {

   open (IN, "+< Corpora/Cetenfolha/ablage.txt");
   open (OUT, "> Corpora/Cetenfolha/ablage1.txt");

   while (<IN>) {
      @zeile = split;
      if ($zeile[0] =~ /<s/) {
         @zeile1 = "";
         for ($j=0, $k=0; $j<=$#zeile; $j++, $k++) {
            $zeile1[$k]= $zeile[$j];
            if ($zeile1[$k] =~ /$satzzeichen/) {
               $zeile1[$k+1] = substr ($zeile1[$k],1);
               $zeile1[$k] = $ersatz[$i];
               for $i(($k+2)..$#zeile1) {
                  $zeile2[$i] = $zeile1[$i];
               }
               $k++;
               push @zeile1,@zeile2;
               @zeile2 = "";
            }
         }
         for $j (0..$#zeile1) {
            print OUT "$zeile1[$j] ";
         }
         print OUT "\n";
      }
      else {print OUT "@zeile\n";}
   }

   close (IN);  close (OUT);
   system "mv Corpora/Cetenfolha/ablage1.txt Corpora/Cetenfolha/ablage.txt";
   $i++;
}

#### Die Wörter und bestimmte SGML-Tags werden gezählt und die aktuelle Wortanzahl
#### an den Satzanfang gestellt

open (IN, "+< Corpora/Cetenfolha/ablage.txt");
open (OUT, "> Corpora/Cetenfolha/ablage1.txt");

while (<IN>) {
   @zeile = split;
   if ($zeile[0] !~ /<s/) {print OUT "@zeile\n";}
   else {
      if ($zeile[1] !~ /</) {$allesaetze++;}
      if ($zeile[1] =~ /<a|<sit/) {$alleautoren++;}
      if ($zeile[1] =~ /<t|<cai/) {$alletitel++;}
      if ($zeile[1] =~ /<li/) {$alleaufz++;}
      for $j (0..$#zeile) {
         if (($zeile[$j] !~ /<|>/) && ($zeile[$j] =~ /[A-Za-zÀ-ü0-9]/))
            {$alleworte++;}
      }
      for $j (0..$#zeile) {
         if ($zeile[$j] =~ /<|>/ && $zeile[$j] !~ /\//) {
            print OUT "$zeile[$j] ";
            if ($zeile[$j+1] !~ /<|>/) {printf OUT "$alleworte: ";}
            }
         else {print OUT "$zeile[$j] ";}
      }
```

```
      print OUT "\n";
   }
}

close (IN); close (OUT);
system  "rm Corpora/Cetenfolha/ablage.txt";

open (OUTW, "> Corpora/Cetenfolha/wortanzahl.txt");
print OUTW "$alleworte Wörter $allesaetze Sätze $alletitel Titel $alleautoren
Autoren $alleaufz Aufzählungen\n";
close (OUTW);

#### Die SGML-Tags <s> und </s> werden nur noch um Sätze gesetzt
#### Aus dem Tag <s frag> wird das Leerzeichen getilgt <sfrag>

open (IN, "+< Corpora/Cetenfolha/ablage1.txt");
open (OUT, "> Corpora/Cetenfolha/Cetenfolha");

while (<IN>) {
   @zeile = split;
   if (($zeile[0] =~ /<s/)  && ($zeile[1] =~ /frag>/)) {
      print OUT "<sfrag>";
      for $j(2..$#zeile)
         {print OUT " $zeile[$j]";}
      print OUT "\n";
      }
   elsif (($zeile[0] =~ /<s/) && ($zeile[1] !~ />/)) {
      for $i (0..$#zeile) { print OUT "$zeile[$i] ";}
      print OUT "\n";
      }
   elsif (($zeile[0] =~ /<s/) && ($zeile[1] =~ />/)) {
      for $i (1..$#zeile-1) { print OUT "$zeile[$i] ";}
      print OUT "\n";
      }
   else {print OUT "@zeile\n";}
}

close (IN); close (OUT);
system  "rm Corpora/Cetenfolha/ablage1.txt";
```

Mit PECCI können einzelne Extraktions- und Exzerptionsschritte separat gesteuert werden. Nach der Eingabe von `Pecci` in die Kommandozeile meldet sich das Programm mit verschiedenen Optionen:

```
heike@linux:~/D> ./Pecci
Möchten Sie ein bereits existierendes Sample bearbeiten?
1 = Substantive der Gefühle = Sample 'Sentimento'
X = Neues Sample
1
Mit welchem Corpus möchten Sie arbeiten?
1 = Cetempublico
2 = Cetenfolha
3 = Testcorpus 2 Millionen
1
Folgende Möglichkeiten stehen zur Auswahl
0 = Fundstellen neuer Verben
1 = Anzahl der Fundstellen der Verben
2 = Extraktion (zusätzlicher) Substantive
3 = Extraktion der Substantiv-Verb Kollokationen
4 = Berechnung des Kollokationspotenzials
5 = Clusterverfahren K-Means Nomina
6 = Clusterverfahren K-Means Verben
Es kann nur ein Modul oder mehrere Module ausgewählt werden.
(Eingabe getrennt durch Leerzeichen)
1 2 3 4 5 6
Die Extraktion der Verben ist abgeschlossen.
Bisher wurden 39 Substantive extrahiert (Singular- und Pluralform
separat): admiração admirações aflição aflições agitação agitações alegria
alegrias alvoroço alvoroços amor amores apreensão apreensões asco ascos
ciúme ciúmes cólera cóleras comoção comoções compaixão compaixões decepção
decepções desesperança desesperanças desespero desesperos desilusão
desilusões dor dores encanto encantos enfado enfados entusiasmo
entusiasmos  estimação estimações excitação excitações fúria fúrias furor
furores inclinação inclinações indignação indignações inveja invejas ira
iras luto lutos medo medos ódio ódios paixão paixões pânico pânicos pena
penas raiva raivas respeito respeitos susto sustos tristeza tristezas
vergonha vergonhas
Geben Sie nun die (zusätzlich) zu extrahierenden Substantive in Singular-
und Pluralform durch Leerzeichen getrennt ein.
esperança esperanças
Wie möchten Sie das Kollokationspotenzial berechnen?
Es können eins oder zwei der Assoziationsmaße gleichzeitig ausgewählt
werden. Default-Einstellung sind t-score und Mutual Information.
1 = t-score
2 = log-likelihood
3 = chi-square
4 = Mutual Information
1 2
Wo soll die Signifikanzgrenze liegen?
1
Wie sollen die Clusterzentren von K-Means für Nomina bestimmt werden?
1 = Zufallsgenerierte Clusterzentren
2 = Selektion des ersten Clusterzentrums und der Schrittgröße
3 = Einzelne Nomina bestimmen die Ausgangsclusterzentren
4 = Zufallsgenerierte Clusterzentren, K-Means wird 100 mal gestartet und
die Ergebnisse gesammelt
1
Wie viele Cluster sollen gebildet werden?
10
Wie sollen die Clusterzentren von K-Means für Verben bestimmt werden?
1 = Zufallsgenerierte Clusterzentren
```

```
2 = Selektion des ersten Clusterzentrums und der Schrittgröße
3 = Einzelne Verben bestimmen die Ausgangsclusterzentren
3
Die nummerierten Verben finden Sie in SentimentoCetemp/ClusterVerb/
zinfoverb.
Eingabe der einzelnen Clusterzentren gefolgt von einem Leerzeichen.
1 4 33 54 67 101 118 204
heike@linux:~/D>
```

Zunächst wählt man das betreffende Sample und das Corpus aus. Das in dieser Arbeit näher betrachtete Wortfeld besteht aus Gefühlssubstantiven, die erstellten Verzeichnisse beginnen mit dem Wort "Sentimento". Man kann beliebige weitere Felder bilden. Pro Sample wird für jedes Corpus ein eigenes Verzeichnis angelegt indem an den Namen des Samples die ersten sechs Buchstaben des Corpusnamen angefügt werden (z.B. `SentimentoCetemp`).
Danach folgt die Wahl der Module.[101] **Modul 1** ermittelt die Frequenz der lemmatisierten Verben im Gesamtcorpus. Es beinhaltet zwei Subroutinen. Mit der Option 1 im Benutzerdialog wird die Zählung der bereits lemmatisierten Verben initiiert. Sie wird nur einmal pro Corpus ausgeführt. Dieser Teil der Corpusstatistik wird von den verschiedenen Samples der Nomina benötigt zur Berechnung der Assoziationsmaße und der Clusterverfahren. Die Frequenzlisten werden im Verzeichnis des Corpus abgelegt (`Corpora/Cetempublico/verbfrequenzen`).
Die lemmatisierten Verben stehen in `Programme/verben`. Dort findet man die Lemmaform (Infinitiv) und, im Falle der unregelmäßigen Verben und der Verben, denen eine homographe Konjugationsform fehlt, den Suchausdruck, im Falle der regelmäßigen Verben den Aufruf der betreffenden Subroutine des Konjugationsmuster mit der Stammform des Verbs. Die Subroutinen sind in `Lemma.pm` enthalten, dort werden die Konjugationsmuster der Verben auch mit dem Suchausdruck verbunden, der die verschiedenen Personalpronomina in Akkusativ und Dativ enthält, die im Portugiesischen durch einen Bindestrich an das Verb angehängt sind. Die Hilfsverben (*ficar, ser, estar, ter, haver*) werden separat verarbeitet und gespeichert (`Programme/verbenhilf`, `Corpora/Cetempublico/verbhilffrequenzen`).

```
abandonar     -3 1  abandon(ar|as|a|amo(s?)|ais|am|ada|adas|ado|ados|ava|avas|
ávamo(s?)|áveis|avam|ei|aste|ou|ámo(s?)|astes|aram|ara|aras|áramo(s?)|áreis|
arei|arás|ará|aremo(s?)|areis|arão|e|es|emo(s?)|eis|em|asse|asses|ássemo(s?)|
ásseis|assem|ares|armo(s?)|ardes|arem|aria|arias|aríamo(s?)|aríeis|ariam|
ando|á)
abater        -3 1  abat(o|er|emo(s?)|eis|em|ida|idas|ido|idos|ia|ias|
íamo(s?)|íeis|iam|i|este|eu|emo(s?)|estes|eram|era|eras|êramo(s?)|êreis|erei|
erás|erá|eremo(s?)|ereis|erão|a|as|amo(s?)|ais|am|esse|esses|êssemo(s?)|
êsseis|essem|eres|ermo(s?)|erdes|erem|eria|erias|eríamo(s?)|eríeis|eriam|
endo|ê)
abolir        -3 1  lemmair    abol
abrandar      -3 1  lemmaar    abrand
acalentar     -3 1  lemmaar    acalent
acalmar       -3 1  lemmaar    acalm                                  ...
(Programme/verben)
```

Die Zahlen nach dem Lemma in der Datei `verben` geben an, wie viele Wörter das Verb links oder rechts vom Nomen stehen kann, um als Kookkurrenzpartner des Nomens zu gelten. In `verben` kann die Größe des Fensters verbindividuell eingestellt werden (vgl. Kapitel 5.3.2). In der ersten Spalte wird die Anzahl der Wörter, die das Verb links vom Substantiv vorkommen kann, als negative Zahl eingetragen, in die zweite Spalte kommt die Anzahl der Wörter rechts vom Substantiv als positive Zahl. Die Default-Einstellung für die Kollokationsspanne ist: drei Wörter links vom Substantiv und ein Wort rechts davon.

101 Die folgende Beschreibung der Module gibt nur einen groben Überblick, genauere Erklärungen sind in den Programmen in Form von Kommentaren vorhanden.

Die Option 0 bietet die Möglichkeit der automatischen Integration neuer lemmatisierter Verben und deren Frequenz. Der Suchausdruck wird in `extraktionverbneu` eingetragen und getestet, die Richtigkeit des neuen Suchmusters kann in `xxlemmakontrolle` überprüft werden. Das Modul 1, das die Fundstellen der Verben zählt, ist das zeitintensivste, da es für jedes Wort im Corpus einen Vergleich mit einem komplexen regulären Ausdruck vornimmt.

Die Ergebnisse aller weiteren Module werden in dem Verzeichnis abgelegt, welches sich aus der Wahl eines Samples mit einem bestimmten Corpus ergibt. **Modul 2** extrahiert im Gegensatz zu Modul 1 nicht die Anzahl der Fundstellen (der Verben), sondern die Konkordanzen der Substantive. Für jedes Substantiv wird eine eigenes Verzeichnis generiert (`SentimentoCetemp/Nomina/esperança/esperança`). Die Konkordanzen stehen in der Datei mit dem Namen des Nomens. Auf diese Weise wird für jedes untersuchte Nomen ein Subcorpus gebildet, das als Grundlage für den eigentlichen Schritt der Kollokationsextraktion dient.

In **Modul 3** wird für jedes dem Sample zugehörige Substantiv überprüft, ob sich eines der lemmatisierten Verben im gleichen Satz befindet. Im Verzeichnis des Substantivs wird ein weiteres Verzeichnis erstellt, in dem die Kookkurrenzkonkordanzen in Dateien unter dem Namen der Verben gespeichert sind (`SentimentoCetemp/Nomina/esperança/Verben/`). Befindet sich das Verb zwar im gleichen Satz, aber außerhalb der Kollokationsspanne, kommen die Konkordanzen in Dateien, deren verbale Bezeichnungen die Ergänzung 'raus' erhält.

```
                              esperança + acalentar
  2612707: Mas a SIC queria um bom espectáculo televisivo , repleto de frases
mortais , acalentado certamente a <esperança> de poder exibir dirigentes
políticos em explícitas cenas de pugilato .
   4710954: Desta feita a vontade dos « encarnados » é quebrar a malapata e
regressar a Lisboa , senão com uma vitória , pelo menos com um resultado que
lhes permita acalentar a <esperança> de seguir em frente .       ...
(SentimentoCetemp/Nomina/esperança/Verben/acalentar)

                        esperança + acalentar  außerhalb des Suchraums
  3136958: Afinal , saíram-lhes galos do campo , rijos , vorazes e , ainda
por cima , assassinos , pois que com a vitória (1-0) conseguida na Luz ,
mataram por completo alguma réstia de <esperança> que os mais optimistas
pudessem acalentar em relação à revalidação do título .
  16271713: Porém , Fábio acalenta há muito a <esperança> de entrar na
selecção nacional brasileira que irá disputar no próximo ano os Jogos
Olímpicos em Atlanta .                                            ...
(SentimentoCetemp/Nomina/esperança/Verben/acalentarraus)
```

Eine dritte Datei wird generiert, die die Kookkurrenz des Nomens mit einem Verb innerhalb der Kollokationsspanne nicht mehr mit dem ganzen Satz, sondern in einem stark verkürzten Kontext anzeigt, und die wichtigsten Umgebungsdaten (vgl. Kapitel 5.2) ausgibt:

```
                        esperança + acalentar: Umgebungsdaten
, acalentado certamente a <esperança> de poder exibir dirigentes políticos em
permita acalentar a <esperança> de seguir em frente .
, acalenta a <esperança> de recuperar a presidência da câmara
clientela acalenta a <esperança> de encontrar os livros de borla
Hispano-Americano acalentou a <esperança> de um maior protagonismo .
vai acalentando a <esperança> de ir a tempo para se
ainda acalenta a <esperança> de um encontro exploratório tripartido --
...
acalentar kommt 70 mal mit esperança vor
1 mal direkt aufeinander folgend
56 mal steht der bestimmte Artikel dazwischen
2 mal steht ein unbestimmter Artikel dazwischen
2 mal folgt que
14 mal folgt de que
40 mal folgt de
3 mal folgt em
(SentimentoCetemp/Nomina/esperança/Verben/acalentarumgebung)
```

Für jedes Nomen werden die Kookkurrenzfrequenzen aller lemmatisierten Verben und der erste Satz der Kookkurrenzkonkordanzen in `.zablage` verwaltet und dort mit den Frequenzen der Substantive und Verben und der Suchraumeinstellung zusammen aufgeführt. Gleichzeitig werden Verzeichnisstatistiken der Kookkurrenzen für das gesamte Wortfeld erstellt und im Ordner `Berechnung` abgelegt. Diese Dateien dienen als Grundlage für die unterschiedlich sortierten Ausgabedateien (vgl. folgende Seiten).
Die Dateien `Kollokationsanteil` bzw. `Samplerelevanz` geben Informationen, die das gesamte Wortfeld betreffen. Mit 'Kollokationsanteil' wird der prozentuale Anteil des Vorkommens eines Nomens mit einem Verb im Suchraum am Gesamtvorkommen des Nomens im Corpus bezeichnet. Die 'Samplerelevanz' stellt den prozentualen Anteil der Kookkurrenzen des Verbs mit den untersuchten Nomina am Gesamtvorkommen des Verbs im Corpus dar.

```
Nomen          Vorkommen mit Verb   Vorkommen total   Kollokationsanteil
admiração                 628               1709              36.7 %
admirações                  4                 16              25.0 %
aflição                   105                456              23.0 %
aflições                   41                150              27.3 %
agitação                  759               3046              25.0 %
agitações                  10                 66              15.2 %
alegria                 1313               4775              27.5 %        ...
(SentimentoCetemp/Berechnung/Kollokationsanteil)

Verb                  Corpus               Sample        Samplerelevanz
abandonar              28588                 63              0.2 %
abater                  5155                  9              0.2 %
abolir                  1436                 64              4.5 %
abrandar                1663                  7              0.4 %
acalentar                558                195              34.9 %
acalmar                 2137                 62              2.9 %
acarretar               1916                 21              1.1 %        ...
(SentimentoCetemp/Berechnung/Samplerelevanz)
```

Modul 3 liefert auch Informationen zur Corpusgröße in Wörtern, sowie deren Verteilung auf Sätze, Titel und weitere mit SGML-Tags markierte Corpussequenzen. Die untersuchten Nomina werden genannt sowie die Anzahl der extrahierten Okkurrenzen, die Zahl der lemmatisierten Verben und die Anzahl der Kookkurrenzen der Samplesubstantive mit den Verben innerhalb des Suchraums.

```
Das untersuchte Corpus 'Cetempublico' hat eine Größe von 174184724
Wörtern verteilt auf 7049916 Sätze, 655058 Titel, 247392 Autorennamen
und 80060 Aufzählungen.

Im Sample 'Sentimento' sind folgende Nomina enthalten: admiração
admirações aflição aflições agitação agitações alegria alegrias ...

Das sind 40 Nomina in Singular- und Pluralform mit insgesamt 159701
Fundstellen. Davon enthalten 53362 Fundstellen ein lemmatisiertes Verb
im Suchraum. 226 Verben sind lemmatisiert.

(SentimentoCetemp/Information)
```

Wie in Kapitel 5.2 dargestellt, müssen Verben, die auch Auxiliarfunktionen übernehmen können, aus einem Subcorpus extrahiert werden, der um die Konkordanzen mit den lemmatisierten Verben innerhalb des Suchraums reduziert ist. Der Subcorpus liegt im Ordner des Nomens mit der Erweiterung "test" (`SentimentoCetemp/Nomina/esperança/esperançatest`).
Modul 3 stellt alle Okkurrenz- und Kookkurrenzdaten zur Verfügung, die für die Berechnung der Assoziationsmaße und Clusterverfahren notwendig sind, sowie die Exzerptionsdateien mit dem sprachlichen Belegmaterial der Kookkurrenzen. Werden dem Sample neue Nomina hinzugefügt, oder weitere Verben lemmatisiert, muss man auch Modul 3 neu kompilieren, um die Module 4, 5 und 6 auszuführen.

Aus der Berechnung des Kollokationspotenzials in **Modul 4** ergeben sich die relevanten Ergebnisse der Kollokationsextraktion. Das Kollokationspotenzial kann nach einem der vier in Kapitel 2.1 vorgestellten Assoziationsmaße ermittelt werden: t-score, log-likelihood, χ^2 und Mutual Information stehen zur Verfügung. Zum Vergleich kann ein zweites Assoziationsmaß gewählt werden. Festgelegt wird auch die Signifikanzgrenze, die angibt wie oft ein Verb mit einem Nomen innerhalb des Suchraums mindestens vorkommen muss, um in den Ausgabedateien zu erscheinen. Die Datei `.zablage` wird mit den neu gewonnenen Daten angereichert zu `outalphabetisch`, wo die Informationen im Ordner des Nomens alphabetisch sortiert nach Verben erscheinen. Die Datei `outrankinglong` enthält die gleichen Daten, jedoch absteigend nach den Werten des gewählten Assoziationsmaßes geordnet.

```
esperança: 11113 Fundstellen im Corpus sortiert nach Kollokationspotenzial

ter          t-score: 26.98  MI: 2.46   esperança: 870  Sample: 8933    alle:
1163390    -3/1   318503: Temos <esperança> de mais História e menos mito ,
mais exactidão e menos romance , no segundo episódio de Histórias que o Tempo
Apagou ( hoje , na TV2 , às 20h45) .

manifestar   t-score: 20.18  MI: 5.39   esperança: 411  Sample: 943     alle:
29248      -3/1    2162344: No final da audiência , Deus Pinheiro manifestou
a <esperança> de que aquele país venha a beneficiar de um aumento substancial
nas dotações que lhe são atribuídas no quadro da Convenção de Lomé , assunto
que está sob a responsabilidade do comissário português .

ser          t-score: 17.41  MI: 1.16   esperança: 640  Sample: 9341    alle:
3129383    -3/1    207284: Poderá não ser muito , mas é uma <esperança> .

(SentimentoCetemp/Nomina/esperança/outrankinglong)
```

Aus den Dateien `outrankingshort` (die keinen Beispielsatz mehr enthalten) der einzelnen Nomina setzt sich `AusgabeNomina` zusammen.

```
Corpus: Cetempublico, 174184724 Wörter
Sample: Sentimento, 40 Nomina (Singular- und Pluralform separat)
Lemmatisierte Verben: 226
Gesamtzahl der Fundstellen der untersuchten Nomina: 159701
Fundstellen mit einem Verb im Suchraum: 53362
Signifikanzgrenze = 1              (CF = Corpusfrequenz, SF = Samplefrequenz)

admiração (1712)   t-score      MI   Kookkurrenz Suchraum       CF       SF
ter                   9.91      2.35     120       -3/1       1163390    8933
esconder              7.66      6.04      59       -3/1         14390     452
manifestar            5.52      4.68      31       -3/1         29248     943
suscitar              5.37      5.80      29       -3/1          8920     407
causar                5.36      5.20      29       -3/1         16253     650
nutrir                5.29      8.55      28       -3/1           555      76
confessar             5.08      5.43      26       -3/1         11639     148
merecer               4.76      4.93      23       -3/1         16877     373
provocar              4.71      4.03      23       -3/1         41836    1308
expressar             4.46      5.72      20       -3/1          6673     132
exprimir              4.11      5.80      17       -3/1          5231     220
mostrar               3.61      3.32      14       -3/1         51696     344
sentir                3.49      3.48      13       -3/1         40915     860
ganhar                3.44      3.11      13       -3/1         59127     250
ser                   2.95      0.53      52       -3/1       3129383    9341
testemunhar           2.64      5.63       7       -3/1          2560      30
conquistar            2.59      3.86       7       -3/1         15082      63
despertar             2.43      4.91       6       -3/1          4512     248
olhar                 2.41      4.16       6       -3/1          9574      62
demonstrar            2.38      3.63       6       -3/1         16293     130
inspirar              1.97      4.09       4       -3/1          6798      63
gerar                 1.94      3.49       4       -3/1         12477     242
haver                 1.87      0.89      10       -3/1        417781    1415
ficar                 1.80      1.32       6       -3/1        162652     391
...
(SentimentoCetemp/AusgabeNomina)
```

Die Datei `AusgabeNomina` kann man in Ausschnitten für *Cetempúblico* in Anhang C1 und C2 finden. Alternativ steht `AusgabeNomina4Scores` zur Verfügung, wo die Werte und das Ranking der vier Assoziationsmaße beim jeweiligen Nomen zu finden sind (vgl. Kapitel 2.1, Anhang C3). In `AusgabeVerben` wird die Ausgabe nach Verben sortiert (Anhang C5):

```
Kookkurrenzen sortiert nach Verben und Kollokationspotenzial
...
acalentar      t-score      MI      Kookkurrenz      Nomen      acalentar     558
esperanças       10.91      8.93          119        4900
esperança         8.36      7.58           70       11113
ódios             1.00      6.16            1         659
paixões           1.00      5.32            1        1520
raiva             1.00      5.50            1        1275
paixão            0.98      4.10            1        5170
amor              0.96      3.34            1       11054
medo              0.96      3.23            1       12288
                                            195                      ...
nutrir         t-score      MI      Kookkurrenz      Nomen      nutrir        555
admiração         5.29      8.55           28        1709
ódio              3.60      7.46           13        2347
paixão            3.46      6.59           12        5170
esperanças        2.82      6.24            8        4900
respeito          2.42      4.37            6       23710
amor              1.98      4.73            4       11054
esperança         1.71      4.44            3       11113
entusiasmo        0.98      4.18            1        4786
medo              0.96      3.24            1       12288
                                             76                      ...
(SentimentoCetemp/AusgabeVerben)
```

Zusätzlich wird eine Rankingliste erstellt, die alle Substantiv-Verb Kookkurrenzen des Wortfeldes enthält, die hier rein numerisch sortiert sind.

```
Wortfeld: Sentimento    Corpus: Cetempublico
                       t-score      MI       Kookkurrenz
valer pena               74.45     6.94          5553
ter medo                 62.39     3.90          4055
ser pena                 40.16     1.77          2340
cumprir pena             31.93     5.47          1028
ter esperança            26.98     2.46           870
ter pena                 23.91     1.74           841
ter vergonha             22.09     3.25           528
fazer amor               21.88     2.85           539
ser vergonha             21.30     2.32           558
condenar pena            21.25     5.12           457
manifestar esperança     20.18     5.39           411
ter esperanças           19.13     2.57           429
condenar penas           18.63     6.57           348
ser amor                 18.56     1.24           684
fazer inveja             18.34     4.64           343
depositar esperanças     18.27     7.76           334
meter medo               17.93     6.11           323
ser esperança            17.41     1.16           640
entrar pânico            17.02     5.58           292
haver esperança          16.64     2.51           328
aplicar pena             16.52     4.68           278
ser paixão               16.08     1.52           424
merecer respeito         15.79     4.71           254
ser desilusão            15.65     2.22           308
incorrer pena            15.35     7.34           236
ter respeito             14.71     1.11           481
perder esperança         14.62     3.88           223
ser alegria              14.61     1.45           365
fazer furor              14.58     5.19           215
cumprir penas            14.30     5.58           206
...
(SentimentoCetemp/AusgabeWortfeld)
```

Die jeweils erste Seite von `AusgabeWortfeld`, berechnet nach t-score, log-likelihood, χ^2 oder MI, befindet sich in Anhang C4.

Das Clusterverfahren K-Means (**Modul 5** und **6**) und seine Ergebnisse werden in Kapitel 7 erläutert. Die Resultate sind, je nach dem ob sie für Verben oder Nomina ermittelt werden, im Ordner `ClusterNomen` bzw. `ClusterVerb` des betreffenden Verzeichnisses enthalten.

Die Verzeichnisstruktur auf der folgenden Seite entsteht nach der Auswahl aller Module von PECCI. Dargestellt sind jeweils nur die Dateien von einem Corpus, einem Sample, einem Nomen und einem Verb.

<- 1: Datei wird generiert von Modul 1 -> 3 4: Datei wird weiterverarbeitet von Modul 3 und 4

```
D/./Pecci
   ./Cetenf
   ./Cetemp
D/Programme/ ExtraktionVerb.pm (extraktionverb und extraktionverbneu)   (Option 0 1)
             ExtraktionNomen.pm                                        (Option 2)
             ExtraktionKollok.pm                                       (Option 3)
             Kollokationspotenzial.pm                                  (Option 4)
             KmeansNomen.pm                                            (Option 5)
             KmeansVerb.pm                                          (Option 6)
             Lemma.pm                                                          -> 0 1 3
             Modulextraktion.pm                                                -> 3
             verben                                                            -> 1 3
             verbenhilf                                                        -> 1 3
             xxverbneu                                                         -> 0
             xxverbneufund                                          <- 0
             xxlemmakontrolle                                       <- 0
D/Corpora/Cetempublico/ Cetempublico                                           -> 1 2
          (---ZZZ)      verbfrequenzen                            <- 01     -> 3
                        verbhilffrequenzen                        <- 01     -> 3
                        verbfrequenzenalpha                       <- 3
                        wortanzahl                                          -> 3 4
D/SentimentoCetemp/ AusgabeNomina                                 <- 4
   (---ZZZ)         AusgabeNomina4Scores                          <- 4
                    AusgabeVerben                                 <- 4
                    AusgabeWortfeld                               <- 4
                    Information                                   <- 3
                    Kollokationsanteil                            <- 3
                    Samplerelevanz                                <- 3
                    .nomen                                        <- 2      -> 3456
                ~/Nomina/admiração/ admiração                     <- 2      -> 3
                          (---ZZZ)  admiraçãotest                 <- 3
                                    outalphabetisch               <- 4
                                    outrankinglong                <- 4
                                    outrankingshort               <- 4
                                    .zablage                      <- 3       -> 4
                                  ~/Verben/ abandonar             <- 3
                                            abandonarraus         <- 3
                                            abandonarumgebung     <- 3
                                            (---zzz)
                ~/ClusterNomen/ Ergebniskmeans                    <- 5
                                VektorenClusterzentrenSort        <- 5
                                VektorenClusterzentrenSortShort   <- 5
                                zinfonomen                        <- 5
                              ~/Matrix/ matrix                    <- 5
                                        matrixnormal              <- 5
                                      ~/Clusterablage/ clusterwert0 (..n)  <- 5
                                                       euklidistanz0 (..n) <- 5
                                                       zuordnung0 (..n)    <- 5
                ~/ClusterVerb/ Ergebniskmeans                     <- 6
                               VektorenClusterzentrenSort         <- 6
                               VektorenClusterzentrenSortShort    <- 6
                               zinfoverb                          <- 6
                             ~/Matrix/ matrix                     <- 6
                                       matrixnormal               <- 6
                                     ~/Clusterablage/ clusterwert0 (..n)  <- 6
                                                      euklidistanz0 (..n) <- 6
                                                      zuordnung0 (..n)    <- 6
                ~/Berechnung/ 2sortnachnomen                      <- 3      -> 4 6
                              2sortnachverben                     <- 3      -> 4 5
                              2sortnachverbenkp                   <- 4      -> 4
```

Die CD1 enthält alle Programme und extrahierten Daten nach der Ausführung der Programme für das Wortfeld der Gefühlssubstantive. Auf der CD2 befinden sich die Corpora in kompromierter Form und die Dateien mit den Frequenzen der Verben im Verzeichnis der jeweiligen Corpora. Möchte man nur mit den Modulen 4, 5 und 6 von PECCI arbeiten, das heißt keine neuen Substantive extrahieren oder lemmatisierte Verben hinzufügen, reicht es ein Verzeichnis anzulegen, in das man die ausführbare Datei 'Pecci', den Ordner 'Programme' und das zu bearbeitende Sample mit der entsprechenden Corpuserweiterung von CD1 kopiert. Die Dateien mit den Verbfrequenzen und der Gesamtwortzahl des Corpus ist in der vorgegebenen Verzeichnisstruktur von CD2 zu kopieren. Möchte man darüber hinaus zusätzliche Substantive extrahieren, ein neues Sample anlegen oder weitere Verben lemmatisieren, müssen auch die Corpora kopiert werden. Dies beansprucht ca. 700 MB Speicherplatz. Danach werden die Corpora in ihrem jeweiligen Verzeichnis entpackt und zur Corpusaufbereitung die ausführbaren Dateien 'Cetenf' bzw. 'Cetemp' gestartet, die wie PECCI im Hauptverzeichnis liegen. Für diesen einmaligen Vorgang sollten 4 GB freier Speicherplatz auf der Festplatte verfügbar sein. Ist die Corpus-aufbereitung abgeschlossen, kann man die komprimierten Corpusdateien und die entpackten Originalcorpora wieder löschen, wodurch sich der benötigte Speicherplatz auf 1,6 GB reduziert. Wahlweise kann natürlich auch nur mit dem kleineren Corpus *Cetenfolha* oder dem Testcorpus von 2 Millionen Wörtern gearbeitet werden, den man auch bequem mit dem Texteditor öffnen kann. Der Testcorpus enthält die ersten 2 Millionen Wörter aus *Cetempúblico.*

Corpus: Cetempublico, 174184724 Wörter
Sample: Sentimento, 40 Nomina (Singular- und Pluralform separat)
Lemmatisierte Verben: 226
Gesamtzahl der Fundstellen der untersuchten Nomina: 159701
Fundstellen mit einem Verb im Suchraum: 53362
Signifikanzgrenze = 2 (CF = Corpusfrequenz, SF = Samplefrequenz)

...

alegria (4775)	t-score	MI	Kookkurrenz	Suchraum	CF	SF
ser	14.61	1.45	365	-3/1	3129383	9341
dar	11.13	3.36	133	-3/1	168517	813
ter	6.68	1.12	98	-3/1	1163390	8933
sentir	6.54	3.69	45	-3/1	40915	860
haver	6.42	1.69	62	-3/1	417781	1415
fazer	5.77	1.45	57	-3/1	489880	2190
esconder	5.50	4.36	31	-3/1	14390	452
manifestar	5.42	3.65	31	-3/1	29248	943
encher	5.25	4.95	28	-3/1	7260	77
chorar	5.08	5.61	26	-3/1	3459	105
saltar	5.07	5.17	26	-3/1	5389	34
viver	4.88	2.80	27	-3/1	59896	419
trazer	4.54	3.43	22	-3/1	26021	218
perder	4.15	2.36	21	-3/1	72509	826
exprimir	3.57	4.51	13	-3/1	5231	220
mostrar	3.51	2.36	15	-3/1	51696	344
exultar	3.46	7.21	12	-3/1	325	12
receber	3.18	1.89	14	-3/1	77287	247
disfarçar	3.14	4.81	10	-3/1	2984	87
expressar	3.10	4.00	10	-3/1	6673	132
transmitir	3.03	3.17	10	-3/1	15300	58
provocar	2.62	2.06	9	-3/1	41836	1308
recuperar	2.62	2.62	8	-3/1	21293	78
demonstrar	2.48	2.75	7	-3/1	16293	130
faltar	2.45	2.61	7	-3/1	18699	83
levar	2.34	1.35	10	-3/1	94774	242
tirar	2.29	2.71	6	-3/1	14518	55
espalhar	2.16	3.32	5	-3/1	6589	61
causar	2.04	2.42	5	-3/1	16253	650
ganhar	2.03	1.46	7	-3/1	59127	250
rebentar	1.96	3.97	4	-3/1	2754	12
testemunhar	1.96	4.04	4	-3/1	2560	30
oferecer	1.92	1.94	5	-3/1	26143	38
deixar	1.90	1.01	9	-3/1	119915	336
festejar	1.69	3.62	3	-3/1	2944	11
conservar	1.68	3.43	3	-3/1	3543	24
cair	1.61	1.63	4	-3/1	28455	51
confessar	1.55	2.24	3	-3/1	11639	148
ficar	1.51	0.70	9	-3/1	162652	391
corar	1.40	4.35	2	-3/1	944	109
excitar	1.40	4.64	2	-3/1	708	8
comover	1.39	4.21	2	-3/1	1080	10
cultivar	1.36	3.30	2	-3/1	2689	25
produzir	1.35	1.50	3	-3/1	24378	46
experimentar	1.34	2.95	2	-3/1	3817	37
motivar	1.26	2.23	2	-3/1	7880	96
morrer	1.23	1.24	3	-3/1	31733	273
cumprir	1.19	1.17	3	-3/1	34024	1271
correr	1.17	1.13	3	-3/1	35345	109
responder	1.13	1.06	3	-3/1	37844	61
permanecer	1.12	1.57	2	-3/1	15207	30
vencer	1.11	1.03	3	-3/1	39000	98
desaparecer	1.10	1.50	2	-3/1	16268	57
dominar	1.10	1.51	2	-3/1	16088	79
prosseguir	1.06	1.37	2	-3/1	18460	29
andar	1.01	1.26	2	-3/1	20757	56
representar	1.01	0.88	3	-3/1	45373	97
existir	0.99	0.68	4	-3/1	74025	255
surgir	0.97	0.82	3	-3/1	48224	126
nascer	0.92	1.04	2	-3/1	25744	99
revelar	0.90	0.74	3	-3/1	52331	198
manter	0.85	0.68	3	-3/1	55665	292
criar	0.78	0.59	3	-3/1	60466	260
aumentar	0.71	0.70	2	-3/1	36220	213
chegar	0.70	0.37	5	-3/1	125518	111
entrar	0.69	0.50	3	-3/1	66123	377
dirigir	0.65	0.61	2	-3/1	39667	29
pôr	0.61	0.57	2	-3/1	41440	98
valer	0.59	0.55	2	-3/1	42302	5638

chamar	0.50	0.44	2	-3/1	47211	73
começar	0.46	0.26	4	-3/1	112636	186
estar	0.39	0.09	21	-3/1	700933	916
passar	0.22	0.12	4	-3/1	129693	240
tornar	0.10	0.07	2	-3/1	67746	140
seguir	-1.01	-0.54	2	-3/1	125184	195

alegrias (513)	t-score	MI	Kookkurrenz	Suchraum	CF	SF
dar	5.57	4.17	32	-3/1	168517	813
ter	4.54	2.06	27	-3/1	1163390	8933
trazer	4.10	5.40	17	-3/1	26021	218
fazer	3.50	2.34	15	-3/1	489880	2190
haver	1.95	1.58	6	-3/1	417781	1415
viver	1.63	2.83	3	-3/1	59896	419
ser	1.49	0.49	15	-3/1	3129383	9341
iniciar	1.33	2.88	2	-3/1	38282	18
continuar	1.17	1.75	2	-3/1	117486	300
passar	1.14	1.66	2	-3/1	129693	240

...

ciúme (308)	t-score	MI	Kookkurrenz	Suchraum	CF	SF
ser	2.62	1.06	16	-3/1	3129383	9341
provocar	2.20	4.21	5	-3/1	41836	1308
haver	2.15	2.09	6	-3/1	417781	1415
sentir	1.69	3.72	3	-3/1	40915	860
ter	1.32	0.89	5	-3/1	1163390	8933
estar	1.02	0.88	3	-3/1	700933	916

ciúmes (437)	t-score	MI	Kookkurrenz	Suchraum	CF	SF
ter	7.02	2.94	55	-3/1	1163390	8933
provocar	4.10	5.09	17	-3/1	41836	1308
motivar	2.64	5.87	7	-3/1	7880	96
causar	2.22	4.81	5	-3/1	16253	650
ser	2.04	0.71	16	-3/1	3129383	9341
haver	2.02	1.74	6	-3/1	417781	1415
fazer	1.95	1.59	6	-3/1	489880	2190
matar	1.71	4.47	3	-3/1	13678	53
levar	1.59	2.54	3	-3/1	94774	242
roer	1.41	7.39	2	-3/1	494	44
desencadear	1.40	4.64	2	-3/1	7715	94
suscitar	1.40	4.49	2	-3/1	8920	407
esconder	1.39	4.01	2	-3/1	14390	452
andar	1.38	3.65	2	-3/1	20757	56
mostrar	1.32	2.74	2	-3/1	51696	344
criar	1.31	2.58	2	-3/1	60466	260
deixar	1.20	1.89	2	-3/1	119915	336
ficar	1.13	1.59	2	-3/1	162652	391
estar	1.12	0.82	4	-3/1	700933	916

...

esperança (11113)	t-score	MI	Kookkurrenz	Suchraum	CF	SF
ter	26.98	2.46	870	-3/1	1163390	8933
manifestar	20.18	5.39	411	-3/1	29248	943
ser	17.41	1.16	640	-3/1	3129383	9341
haver	16.64	2.51	328	-3/1	417781	1415
perder	14.62	3.88	223	-3/1	72509	826
alimentar	12.95	5.48	169	-3/1	11053	449
restar	12.13	5.01	149	-3/1	15507	182
manter	10.15	3.43	110	-3/1	55665	292
dar	9.67	2.36	114	-3/1	168517	813
acalentar	8.36	7.58	70	-3/1	558	195
depositar	8.21	5.35	68	-3/1	5081	408
existir	7.34	2.59	63	-3/1	74025	255
exprimir	6.73	4.93	46	-3/1	5231	220
renascer	6.39	6.11	41	-3/1	1425	66
trazer	6.22	3.23	42	-3/1	26021	218
representar	6.12	2.70	43	-3/1	45373	97
ficar	5.60	1.57	50	-3/1	162652	391
viver	5.36	2.24	36	-3/1	59896	419
deixar	5.02	1.63	39	-3/1	119915	336
aumentar	4.65	2.42	26	-3/1	36220	213
expressar	4.14	3.74	18	-3/1	6673	132
continuar	4.11	1.39	30	-3/1	117486	300
nascer	3.98	2.45	19	-3/1	25744	99

restaurar	3.83	4.60	15	-3/1	2359	25
surgir	3.65	1.82	19	-3/1	48224	126
esconder	3.20	2.57	12	-3/1	14390	452
fazer	3.20	0.57	55	-3/1	489880	2190
crescer	3.13	2.34	12	-3/1	18109	102
conservar	3.09	3.79	10	-3/1	3543	24
morrer	3.04	1.86	13	-3/1	31733	273
oferecer	2.98	1.97	12	-3/1	26143	38
tornar	2.92	1.31	16	-3/1	67746	140
retirar	2.81	2.18	10	-3/1	17705	38
confessar	2.75	2.49	9	-3/1	11639	148
recuperar	2.73	2.00	10	-3/1	21293	78
criar	2.71	1.29	14	-3/1	60466	260
guardar	2.67	2.89	8	-3/1	6944	39
transmitir	2.67	2.22	9	-3/1	15300	58
abandonar	2.59	1.70	10	-3/1	28588	63
levar	2.49	0.97	16	-3/1	94774	242
mostrar	2.32	1.20	11	-3/1	51696	344
revelar	2.31	1.19	11	-3/1	52331	198
encher	2.26	2.56	6	-3/1	7260	77
estar	2.19	0.33	62	-3/1	700933	916
diminuir	2.15	2.10	6	-3/1	11467	89
proclamar	2.13	3.08	5	-3/1	3611	23
gerar	2.12	2.02	6	-3/1	12477	242
pôr	2.12	1.22	9	-3/1	41440	98
despertar	2.11	2.85	5	-3/1	4512	248
permanecer	2.05	1.82	6	-3/1	15207	30
suscitar	1.98	2.17	5	-3/1	8920	407
matar	1.85	1.75	5	-3/1	13678	53
tirar	1.82	1.69	5	-3/1	14518	55
motivar	1.75	2.07	4	-3/1	7880	96
nutrir	1.71	4.44	3	-3/1	555	76
olhar	1.69	1.88	4	-3/1	9574	62
semear	1.68	3.57	3	-3/1	1330	104
apelar	1.63	1.69	4	-3/1	11571	69
valer	1.63	0.95	7	-3/1	42302	5638
desfazer	1.61	2.67	3	-3/1	3272	34
demonstrar	1.48	1.35	4	-3/1	16293	130
sobreviver	1.46	1.87	3	-3/1	7268	31
lançar	1.43	0.78	7	-3/1	50287	228
cair	1.42	1.01	5	-3/1	28455	51
infundir	1.41	6.50	2	-3/1	47	15
começar	1.39	0.51	12	-3/1	112636	186
seguir	1.39	0.49	13	-3/1	125184	195
vestir	1.34	1.48	3	-3/1	10660	59
condenar	1.32	1.07	4	-3/1	21468	830
cultivar	1.29	2.46	2	-3/1	2689	25
afastar	1.28	1.02	4	-3/1	22651	40
destruir	1.25	1.28	3	-3/1	13094	39
conceder	1.22	1.22	3	-3/1	13910	12
enterrar	1.21	1.92	2	-3/1	4605	14
responder	1.16	0.73	5	-3/1	37844	61
desaparecer	1.13	1.06	3	-3/1	16268	57
espalhar	1.12	1.56	2	-3/1	6589	61
enganar	1.11	1.55	2	-3/1	6654	26
sentir	1.07	0.65	5	-3/1	40915	860
prosseguir	1.05	0.94	3	-3/1	18460	29
chamar	0.89	0.51	5	-3/1	47211	73
mandar	0.73	0.72	2	-3/1	15188	14
acreditar	0.71	0.53	3	-3/1	27702	51
declarar	0.67	0.49	3	-3/1	28910	59
faltar	0.57	0.52	2	-3/1	18699	83
passar	0.55	0.19	10	-3/1	129693	240
impor	0.53	0.47	2	-3/1	19601	144
cumprir	0.48	0.32	3	-3/1	34024	1271
considerar	0.26	0.09	9	-3/1	128883	169
colocar	0.24	0.13	4	-3/1	55316	70
entrar	-0.11	-0.05	4	-3/1	66123	377
correr	-0.18	-0.12	2	-3/1	35345	109
vencer	-0.35	-0.22	2	-3/1	39000	98
ganhar	-0.45	-0.23	3	-3/1	59127	250
assumir	-0.46	-0.28	2	-3/1	41594	60
chegar	-0.82	-0.29	6	-3/1	125518	111
receber	-2.07	-0.90	2	-3/1	77287	247

esperanças (4900)	t-score	MI	Kookkurrenz	Suchraum	CF	SF
ter	19.13	2.57	429	-3/1	1163390	8933
depositar	18.27	7.76	334	-3/1	5081	408
alimentar	13.43	6.37	181	-3/1	11053	449

perder	12.08	4.30	150	-3/1	72509	826
acalentar	10.91	8.93	119	-3/1	558	195
dar	7.93	2.72	72	-3/1	168517	813
manter	7.61	3.66	61	-3/1	55665	292
haver	7.37	1.87	76	-3/1	417781	1415
manifestar	6.27	3.91	41	-3/1	29248	943
deixar	5.79	2.47	40	-3/1	119915	336
suscitar	4.85	4.56	24	-3/1	8920	407
aumentar	4.58	3.12	23	-3/1	36220	213
trazer	4.53	3.40	22	-3/1	26021	218
existir	4.47	2.44	24	-3/1	74025	255
criar	4.21	2.51	21	-3/1	60466	260
colocar	4.00	2.50	19	-3/1	55316	70
desfazer	3.98	5.16	16	-3/1	3272	34
renascer	3.86	5.92	15	-3/1	1425	66
pôr	3.71	2.62	16	-3/1	41440	98
matar	3.50	3.52	13	-3/1	13678	53
diminuir	3.22	3.53	11	-3/1	11467	89
destruir	3.21	3.40	11	-3/1	13094	39
restar	3.19	3.23	11	-3/1	15507	182
crescer	3.16	3.07	11	-3/1	18109	102
enterrar	2.96	4.24	9	-3/1	4605	14
fazer	2.96	0.78	30	-3/1	489880	2190
continuar	2.86	1.44	14	-3/1	117486	300
retirar	2.83	2.89	9	-3/1	17705	38
nutrir	2.82	6.24	8	-3/1	555	76
abandonar	2.73	2.42	9	-3/1	28588	63
desvanecer	2.64	6.16	7	-3/1	528	18
guardar	2.57	3.58	7	-3/1	6944	39
representar	2.38	1.84	8	-3/1	45373	97
nascer	2.37	2.27	7	-3/1	25744	99
oferecer	2.37	2.25	7	-3/1	26143	38
dissipar	2.22	5.11	5	-3/1	1072	24
expressar	2.15	3.28	5	-3/1	6673	132
encher	2.14	3.20	5	-3/1	7260	77
confessar	2.09	2.73	5	-3/1	11639	148
morrer	2.09	1.91	6	-3/1	31733	273
estar	2.03	0.45	31	-3/1	700933	916
afastar	1.95	2.06	5	-3/1	22651	40
despertar	1.94	3.45	4	-3/1	4512	248
exprimir	1.93	3.30	4	-3/1	5231	220
ser	1.92	0.20	108	-3/1	3129383	9341
surgir	1.90	1.49	6	-3/1	48224	126
cumprir	1.81	1.65	5	-3/1	34024	1271
tirar	1.80	2.28	4	-3/1	14518	55
vencer	1.75	1.52	5	-3/1	39000	98
conservar	1.67	3.40	3	-3/1	3543	24
ganhar	1.49	1.10	5	-3/1	59127	250
desaparecer	1.47	1.88	3	-3/1	16268	57
ficar	1.47	0.68	9	-3/1	162652	391
começar	1.45	0.79	7	-3/1	112636	186
evaporar	1.41	5.77	2	-3/1	222	7
iludir	1.38	3.70	2	-3/1	1765	22
chegar	1.31	0.68	7	-3/1	125518	111
lançar	1.29	1.04	4	-3/1	50287	228
mentir	1.22	1.96	2	-3/1	9970	10
gerar	1.17	1.74	2	-3/1	12477	242
permanecer	1.11	1.54	2	-3/1	15207	30
demonstrar	1.09	1.47	2	-3/1	16293	130
tornar	1.05	0.74	4	-3/1	67746	140
recuperar	0.99	1.21	2	-3/1	21293	78
revelar	0.88	0.71	3	-3/1	52331	198
viver	0.76	0.58	3	-3/1	59896	419
responder	0.66	0.63	2	-3/1	37844	61
seguir	0.66	0.35	5	-3/1	125184	195
considerar	0.61	0.32	5	-3/1	128883	169
passar	0.60	0.32	5	-3/1	129693	240
mostrar	0.39	0.32	2	-3/1	51696	344
levar	0.19	0.12	3	-3/1	94774	242
entrar	0.10	0.07	2	-3/1	66123	377
receber	-0.12	-0.08	2	-3/1	77287	247

...

fúria (1714)	t-score	MI	Kookkurrenz	Suchraum	CF	SF
provocar	7.15	4.84	52	-3/1	41836	1308
ter	3.75	1.06	33	-3/1	1163390	8933
esconder	3.57	4.52	13	-3/1	14390	452
enfrentar	3.43	4.54	12	-3/1	12989	106

exprimir	2.81	5.05	8	-3/1	5231	220
desencadear	2.80	4.66	8	-3/1	7715	94
ser	2.72	0.48	50	-3/1	3129383	9341
acalmar	2.64	5.81	7	-3/1	2137	62
atacar	2.60	3.96	7	-3/1	13542	24
entrar	2.60	2.51	8	-3/1	66123	377
travar	2.42	4.37	6	-3/1	7750	55
despertar	2.22	4.72	5	-3/1	4512	248
deixar	2.20	1.78	7	-3/1	119915	336
suscitar	2.20	4.04	5	-3/1	8920	407
causar	2.16	3.44	5	-3/1	16253	650
manifestar	2.11	2.85	5	-3/1	29248	943
aumentar	2.08	2.64	5	-3/1	36220	213
sentir	2.06	2.52	5	-3/1	40915	860
correr	1.83	2.44	4	-3/1	35345	109
pôr	1.80	2.28	4	-3/1	41440	98
controlar	1.79	2.24	4	-3/1	43443	56
mostrar	1.75	2.06	4	-3/1	51696	344
destilar	1.73	6.52	3	-3/1	448	23
ferver	1.73	6.62	3	-3/1	406	12
motivar	1.69	3.66	3	-3/1	7880	96
alimentar	1.67	3.32	3	-3/1	11053	449
tornar	1.67	1.79	4	-3/1	67746	140
crescer	1.63	2.82	3	-3/1	18109	102
trazer	1.58	2.46	3	-3/1	26021	218
ficar	1.52	1.14	5	-3/1	162652	391
dirigir	1.51	2.04	3	-3/1	39667	29
lançar	1.45	1.80	3	-3/1	50287	228
experimentar	1.39	3.97	2	-3/1	3817	37
poupar	1.37	3.40	2	-3/1	6752	34
temer	1.35	3.16	2	-3/1	8646	21
destruir	1.32	2.74	2	-3/1	13094	39
dominar	1.30	2.54	2	-3/1	16088	79
esqueçer	1.24	2.09	2	-3/1	25223	107
vencer	1.14	1.65	2	-3/1	39000	98
seguir	1.02	0.89	3	-3/1	125184	195
existir	0.90	1.01	2	-3/1	74025	255
haver	0.77	0.38	6	-3/1	417781	1415
levar	0.75	0.76	2	-3/1	94774	242
passar	0.51	0.45	2	-3/1	129693	240
dar	0.24	0.19	2	-3/1	168517	813
fazer	-0.41	-0.19	4	-3/1	489880	2190
estar	-0.85	-0.32	5	-3/1	700933	916

fúrias (79)	t-score	MI	Kookkurrenz	Suchraum	CF	SF
dar	1.69	3.67	3	-3/1	168517	813
ter	1.43	1.74	3	-3/1	1163390	8933
ser	0.41	0.34	2	-3/1	3129383	9341

...

inclinação (768)	t-score	MI	Kookkurrenz	Suchraum	CF	SF
ter	6.72	2.37	55	-3/1	1163390	8933
ser	4.65	1.18	45	-3/1	3129383	9341
mostrar	3.25	3.88	11	-3/1	51696	344
revelar	2.92	3.66	9	-3/1	52331	198
dar	2.57	2.38	8	-3/1	168517	813
manifestar	2.40	3.84	6	-3/1	29248	943
sentir	2.16	3.32	5	-3/1	40915	860
disfarçar	1.72	5.43	3	-3/1	2984	87
provocar	1.63	2.79	3	-3/1	41836	1308
existir	1.54	2.22	3	-3/1	74025	255
esconder	1.37	3.45	2	-3/1	14390	452
aumentar	1.30	2.53	2	-3/1	36220	213
vencer	1.29	2.45	2	-3/1	39000	98
levar	1.12	1.57	2	-3/1	94774	242
haver	0.67	0.49	3	-3/1	417781	1415

inclinações (182)	t-score	MI	Kookkurrenz	Suchraum	CF	SF
ter	3.11	2.29	12	-3/1	1163390	8933
vencer	1.39	3.89	2	-3/1	39000	98
seguir	1.32	2.73	2	-3/1	125184	195
haver	1.11	1.52	2	-3/1	417781	1415
ser	-0.16	-0.09	3	-3/1	3129383	9341

...

inveja (1179)	t-score	MI	Kookkurrenz	Suchraum	CF	SF
fazer	18.34	4.64	343	-3/1	489880	2190
ter	6.73	2.03	60	-3/1	1163390	8933
roer	6.08	9.31	37	-3/1	494	44
ser	5.27	1.09	63	-3/1	3129383	9341
corar	4.00	7.83	16	-3/1	944	109
causar	3.84	4.92	15	-3/1	16253	650
olhar	3.59	5.30	13	-3/1	9574	62
ficar	3.15	2.39	12	-3/1	162652	391
morrer	2.93	3.74	9	-3/1	31733	273
sentir	2.91	3.48	9	-3/1	40915	860
meter	2.80	4.76	8	-3/1	10125	381
provocar	2.54	3.21	7	-3/1	41836	1308
despertar	2.22	5.10	5	-3/1	4512	248
suscitar	1.97	4.19	4	-3/1	8920	407
confessar	1.69	3.64	3	-3/1	11639	148
gerar	1.68	3.57	3	-3/1	12477	242
sofrer	1.61	2.64	3	-3/1	31630	253
deixar	1.59	1.59	4	-3/1	119915	336
haver	1.58	0.91	7	-3/1	417781	1415
criar	1.50	1.99	3	-3/1	60466	260
disfarçar	1.40	4.60	2	-3/1	2984	87
motivar	1.38	3.62	2	-3/1	7880	96
alimentar	1.36	3.29	2	-3/1	11053	449
passar	1.23	1.23	3	-3/1	129693	240
valer	1.21	1.94	2	-3/1	42302	5638
mostrar	1.17	1.74	2	-3/1	51696	344
estar	-0.37	-0.17	4	-3/1	700933	916

invejas (159)	t-score	MI	Kookkurrenz	Suchraum	CF	SF
provocar	2.63	5.21	7	-3/1	41836	1308
despertar	2.45	7.28	6	-3/1	4512	248
suscitar	2.45	6.60	6	-3/1	8920	407
haver	2.29	2.76	6	-3/1	417781	1415
gerar	2.23	6.08	5	-3/1	12477	242
motivar	1.41	5.63	2	-3/1	7880	96
olhar	1.41	5.43	2	-3/1	9574	62
ganhar	1.38	3.61	2	-3/1	59127	250
existir	1.37	3.39	2	-3/1	74025	255
deixar	1.34	2.91	2	-3/1	119915	336
ter	1.12	1.04	3	-3/1	1163390	8933
ser	0.08	0.05	3	-3/1	3129383	9341

...

pena (22112)	t-score	MI	Kookkurrenz	Suchraum	CF	SF
valer	74.45	6.94	5553	-3/1	42302	5638
ser	40.16	1.77	2340	-3/1	3129383	9341
cumprir	31.93	5.47	1028	-3/1	34024	1271
ter	23.91	1.74	841	-3/1	1163390	8933
condenar	21.25	5.12	457	-3/1	21468	830
aplicar	16.52	4.68	278	-3/1	20265	438
incorrer	15.35	7.34	236	-3/1	1211	287
fazer	13.09	1.51	282	-3/1	489880	2190
continuar	9.44	2.06	117	-3/1	117486	300
dar	8.04	1.57	103	-3/1	168517	813
abolir	7.98	5.86	64	-3/1	1436	64
correr	7.51	2.67	65	-3/1	35345	109
merecer	7.47	3.33	60	-3/1	16877	373
perder	6.92	1.95	65	-3/1	72509	826
sentir	6.72	2.36	55	-3/1	40915	860
enfrentar	6.23	3.24	42	-3/1	12989	106
considerar	5.95	1.36	64	-3/1	128883	169
impor	5.31	2.58	33	-3/1	19601	144
sofrer	4.95	2.08	32	-3/1	31630	253
apanhar	4.70	2.82	25	-3/1	11715	204
chorar	4.70	3.96	23	-3/1	3459	105
pagar	4.69	1.85	31	-3/1	38389	41
receber	4.47	1.33	37	-3/1	77287	247
passar	4.35	1.03	46	-3/1	129693	240
seguir	4.34	1.04	45	-3/1	125184	195
defender	4.11	1.34	31	-3/1	64032	90
manter	3.96	1.38	28	-3/1	55665	292
experimentar	3.88	3.50	16	-3/1	3817	37
ficar	3.84	0.82	47	-3/1	162652	391
acreditar	3.69	1.74	20	-3/1	27702	51
iludir	3.68	4.13	14	-3/1	1765	22

prestar	3.45	1.99	16	-3/1	17242	25
exigir	3.37	1.48	19	-3/1	33968	152
esconder	3.25	2.04	14	-3/1	14390	452
acarretar	3.09	3.72	10	-3/1	1916	21
fingir	2.94	3.84	9	-3/1	1524	21
deitar	2.88	2.42	10	-3/1	6999	16
diminuir	2.88	2.02	11	-3/1	11467	89
livrar	2.76	3.80	8	-3/1	1415	25
meter	2.76	2.05	10	-3/1	10125	381
tirar	2.76	1.79	11	-3/1	14518	55
restaurar	2.72	3.29	8	-3/1	2359	25
criar	2.60	0.91	19	-3/1	60466	260
levar	2.59	0.73	25	-3/1	94774	242
confessar	2.51	1.81	9	-3/1	11639	148
pôr	2.34	0.98	14	-3/1	41440	98
aumentar	2.33	1.04	13	-3/1	36220	213
começar	2.29	0.60	26	-3/1	112636	186
disfarçar	2.29	2.76	6	-3/1	2984	87
travar	2.27	1.96	7	-3/1	7750	55
usar	2.26	1.06	12	-3/1	32885	36
executar	2.23	1.85	7	-3/1	8642	17
interrogar	2.23	1.86	7	-3/1	8621	31
prosseguir	2.22	1.35	9	-3/1	18460	29
enganar	2.10	1.96	6	-3/1	6654	26
existir	2.03	0.65	18	-3/1	74025	255
manifestar	1.99	0.99	10	-3/1	29248	943
alimentar	1.88	1.45	6	-3/1	11053	449
inspirar	1.85	1.76	5	-3/1	6798	63
poupar	1.85	1.76	5	-3/1	6752	34
pegar	1.84	1.73	5	-3/1	7001	15
deixar	1.79	0.46	24	-3/1	119915	336
consagrar	1.77	2.16	4	-3/1	3636	12
chamar	1.73	0.69	12	-3/1	47211	73
entrar	1.71	0.58	15	-3/1	66123	377
exprimir	1.67	1.80	4	-3/1	5231	220
morrer	1.66	0.80	9	-3/1	31733	273
expirar	1.63	2.88	3	-3/1	1328	3
recuperar	1.62	0.95	7	-3/1	21293	78
guardar	1.56	1.51	4	-3/1	6944	39
abandonar	1.55	0.79	8	-3/1	28588	63
declarar	1.53	0.78	8	-3/1	28910	59
festejar	1.52	2.08	3	-3/1	2944	11
conservar	1.47	1.90	3	-3/1	3543	24
esqueçer	1.44	0.78	7	-3/1	25223	107
responder	1.40	0.63	9	-3/1	37844	61
suprimir	1.28	2.39	2	-3/1	1450	7
abrandar	1.26	2.25	2	-3/1	1663	7
retirar	1.23	0.80	5	-3/1	17705	38
aliviar	1.19	1.84	2	-3/1	2506	70
destruir	1.17	0.88	4	-3/1	13094	39
suscitar	1.08	0.97	3	-3/1	8920	407
submeter	1.05	0.94	3	-3/1	9239	9
mentir	1.00	0.86	3	-3/1	9970	10
cair	0.97	0.51	6	-3/1	28455	51
abater	0.95	1.12	2	-3/1	5155	9
tomar	0.93	0.40	8	-3/1	42217	58
utilizar	0.92	0.39	8	-3/1	42568	25
iniciar	0.81	0.36	7	-3/1	38282	18
mostrar	0.81	0.32	9	-3/1	51696	344
somar	0.81	0.86	2	-3/1	6697	17
nascer	0.77	0.43	5	-3/1	25744	99
conceder	0.71	0.53	3	-3/1	13910	12
mandar	0.62	0.44	3	-3/1	15188	14
comer	0.59	0.54	2	-3/1	9156	7
gostar	0.50	0.25	5	-3/1	30592	19
produzir	0.45	0.26	4	-3/1	24378	46
apelar	0.38	0.31	2	-3/1	11571	69
trazer	0.35	0.19	4	-3/1	26021	218
gerar	0.29	0.23	2	-3/1	12477	242
atacar	0.20	0.15	2	-3/1	13542	24
matar	0.19	0.14	2	-3/1	13678	53
tornar	0.13	0.05	9	-3/1	67746	140
afastar	0.07	0.04	3	-3/1	22651	40
transmitir	0.04	0.03	2	-3/1	15300	58
restar	0.02	0.02	2	-3/1	15507	182
tratar	-0.07	-0.03	8	-3/1	64665	108
oferecer	-0.18	-0.10	3	-3/1	26143	38
andar	-0.45	-0.28	2	-3/1	20757	56
surgir	-0.50	-0.20	5	-3/1	48224	126
controlar	-0.76	-0.32	4	-3/1	43443	56

chegar	-0.81	-0.20	13	-3/1	125518	111
viver	-1.16	-0.42	5	-3/1	59896	419
lançar	-1.19	-0.47	4	-3/1	50287	228
assumir	-1.32	-0.57	3	-3/1	41594	60
ganhar	-1.75	-0.63	4	-3/1	59127	250
haver	-2.06	-0.28	40	-3/1	417781	1415
provocar	-2.34	-0.98	2	-3/1	41836	1308
revelar	-3.28	-1.20	2	-3/1	52331	198
colocar	-3.55	-1.26	2	-3/1	55316	70
estar	-10.41	-1.05	31	-3/1	700933	916

penas (3976)	t-score	MI	Kookkurrenz	Suchraum	CF	SF
condenar	18.63	6.57	348	-3/1	21468	830
cumprir	14.30	5.58	206	-3/1	34024	1271
aplicar	12.41	5.81	155	-3/1	20265	438
ser	8.35	0.95	185	-3/1	3129383	9341
incorrer	6.70	7.40	45	-3/1	1211	287
aumentar	4.83	3.41	25	-3/1	36220	213
agrar	4.57	5.91	21	-3/1	2494	25
receber	4.20	2.48	21	-3/1	77287	247
ter	3.84	0.73	55	-3/1	1163390	8933
sofrer	3.69	3.03	15	-3/1	31630	253
diminuir	3.08	3.64	10	-3/1	11467	89
impor	2.85	3.00	9	-3/1	19601	144
enfrentar	2.72	3.30	8	-3/1	12989	106
estar	2.27	0.56	28	-3/1	700933	916
executar	2.15	3.23	5	-3/1	8642	17
arrancar	2.14	3.15	5	-3/1	9408	19
usar	2.14	2.08	6	-3/1	32885	36
apanhar	2.12	2.93	5	-3/1	11715	204
defender	2.09	1.57	7	-3/1	64032	90
considerar	2.02	1.12	9	-3/1	128883	169
acarretar	1.98	4.52	4	-3/1	1916	21
aliviar	1.97	4.25	4	-3/1	2506	70
exigir	1.89	1.86	5	-3/1	33968	152
pagar	1.84	1.74	5	-3/1	38389	41
somar	1.64	2.98	3	-3/1	6697	17
vestir	1.59	2.51	3	-3/1	10660	59
levar	1.57	1.02	6	-3/1	94774	242
apelar	1.23	2.02	2	-3/1	11571	69
tirar	1.18	1.80	2	-3/1	14518	55
existir	1.16	0.86	4	-3/1	74025	255
prosseguir	1.12	1.56	2	-3/1	18460	29
fazer	0.99	0.29	15	-3/1	489880	2190
trazer	0.99	1.21	2	-3/1	26021	218
haver	0.96	0.31	13	-3/1	417781	1415
chegar	0.95	0.56	5	-3/1	125518	111
ficar	0.93	0.48	6	-3/1	162652	391
tomar	0.73	0.73	2	-3/1	42217	58
revelar	0.57	0.52	2	-3/1	52331	198
manter	0.52	0.45	2	-3/1	55665	292
dar	0.08	0.04	4	-3/1	168517	813
continuar	-0.48	-0.29	2	-3/1	117486	300
passar	-0.68	-0.39	2	-3/1	129693	240

raiva (1275)	t-score	MI	Kookkurrenz	Suchraum	CF	SF
sentir	4.05	4.04	17	-3/1	40915	860
chorar	3.99	6.45	16	-3/1	3459	105
ter	3.92	1.26	30	-3/1	1163390	8933
ficar	3.12	2.31	12	-3/1	162652	391
exprimir	2.81	5.34	8	-3/1	5231	220
fazer	2.78	1.36	14	-3/1	489880	2190
dar	2.77	2.09	10	-3/1	168517	813
ser	2.70	0.56	40	-3/1	3129383	9341
manifestar	2.56	3.49	7	-3/1	29248	943
tornar	2.46	2.65	7	-3/1	67746	140
haver	2.20	1.18	10	-3/1	417781	1415
dirigir	2.11	2.85	5	-3/1	39667	29
destilar	2.00	7.11	4	-3/1	448	23
meter	1.96	3.99	4	-3/1	10125	381
morrer	1.88	2.85	4	-3/1	31733	273
corar	1.73	6.07	3	-3/1	944	109
tremer	1.73	5.55	3	-3/1	1601	40
festejar	1.72	4.94	3	-3/1	2944	11
rebentar	1.72	5.00	3	-3/1	2754	12
esconder	1.67	3.35	3	-3/1	14390	452
causar	1.66	3.23	3	-3/1	16253	650
provocar	1.56	2.28	3	-3/1	41836	1308

seguir	1.54	1.47	4	-3/1	125184	195
perder	1.43	1.73	3	-3/1	72509	826
ferver	1.41	6.51	2	-3/1	406	12
reprimir	1.41	5.51	2	-3/1	1110	15
encher	1.38	3.63	2	-3/1	7260	77
expressar	1.38	3.71	2	-3/1	6673	132
motivar	1.37	3.55	2	-3/1	7880	96
olhar	1.36	3.35	2	-3/1	9574	62
apanhar	1.35	3.15	2	-3/1	11715	204
destruir	1.35	3.04	2	-3/1	13094	39
dominar	1.33	2.83	2	-3/1	16088	79
merecer	1.33	2.78	2	-3/1	16877	373
crescer	1.32	2.71	2	-3/1	18109	102
andar	1.31	2.58	2	-3/1	20757	56
perceber	1.30	2.52	2	-3/1	21905	75
nascer	1.28	2.36	2	-3/1	25744	99
lançar	1.15	1.69	2	-3/1	50287	228
mostrar	1.15	1.66	2	-3/1	51696	344
ganhar	1.11	1.53	2	-3/1	59127	250
criar	1.10	1.51	2	-3/1	60466	260
tratar	1.08	1.44	2	-3/1	64665	108
estar	1.01	0.44	8	-3/1	700933	916
deixar	0.79	0.82	2	-3/1	119915	336

raivas (44)	t-score	MI	Kookkurrenz	Suchraum	CF	SF
alimentar	1.41	6.57	2	-3/1	11053	449
ter	1.21	1.92	2	-3/1	1163390	8933

...

susto (1263)	t-score	MI	Kookkurrenz	Suchraum	CF	SF
apanhar	11.44	7.34	131	-3/1	11715	204
ganhar	10.54	5.57	112	-3/1	59127	250
pregar	8.06	8.75	65	-3/1	1416	89
ser	7.97	1.52	104	-3/1	3129383	9341
provocar	5.14	4.49	27	-3/1	41836	1308
passar	4.49	3.15	22	-3/1	129693	240
sofrer	3.94	4.25	16	-3/1	31630	253
recuperar	3.42	4.35	12	-3/1	21293	78
morrer	3.09	3.77	10	-3/1	31733	273
causar	2.79	4.22	8	-3/1	16253	650
esqueçer	2.37	3.49	6	-3/1	25223	107
dar	2.18	1.75	7	-3/1	168517	813
ter	2.08	0.70	17	-3/1	1163390	8933
livrar	1.99	5.97	4	-3/1	1415	25
valer	1.55	2.28	3	-3/1	42302	5638
chamar	1.53	2.17	3	-3/1	47211	73
viver	1.48	1.93	3	-3/1	59896	419
meter	1.36	3.30	2	-3/1	10125	381
correr	1.23	2.05	2	-3/1	35345	109
haver	1.21	0.68	6	-3/1	417781	1415
mostrar	1.15	1.67	2	-3/1	51696	344
ficar	1.05	0.93	3	-3/1	162652	391
começar	0.84	0.90	2	-3/1	112636	186
chegar	0.77	0.79	2	-3/1	125518	111
estar	0.37	0.17	6	-3/1	700933	916
fazer	-1.10	-0.57	2	-3/1	489880	2190

sustos (225)	t-score	MI	Kookkurrenz	Suchraum	CF	SF
apanhar	4.47	7.19	20	-3/1	11715	204
pregar	3.32	8.70	11	-3/1	1416	89
haver	2.99	2.92	10	-3/1	417781	1415
provocar	1.97	4.30	4	-3/1	41836	1308
passar	1.92	3.17	4	-3/1	129693	240
ser	1.88	0.91	10	-3/1	3129383	9341
sofrer	1.71	4.30	3	-3/1	31630	253
viver	1.69	3.66	3	-3/1	59896	419
seguir	1.30	2.52	2	-3/1	125184	195
dar	1.26	2.22	2	-3/1	168517	813
ter	0.86	0.69	3	-3/1	1163390	8933

...

Corpus: Cetempublico, 174184724 Wörter
Sample: Sentimento, 40 Nomina (Singular- und Pluralform separat)
Lemmatisierte Verben: 226
Gesamtzahl der Fundstellen der untersuchten Nomina: 159701
Fundstellen mit einem Verb im Suchraum: 53362
Signifikanzgrenze = 2 (CF = Corpusfrequenz, SF = Samplefrequenz)

...

medo (12288)	log-like	t-score	Kookkurrenz	Suchraum	CF	SF
ter	25159.51	62.39	4055	-3/1	1163390	8933
meter	3324.06	17.93	323	-3/1	10125	381
sentir	431.99	9.07	88	-3/1	40915	860
estar	367.50	12.14	236	-3/1	700933	916
ser	360.39	14.13	553	-3/1	3129383	9341
perder	352.85	9.00	91	-3/1	72509	826
tremer	287.02	5.55	31	-3/1	1601	40
viver	273.22	7.99	72	-3/1	59896	419
dominar	172.45	5.72	35	-3/1	16088	79
provocar	145.12	6.03	42	-3/1	41836	1308
vencer	139.80	5.89	40	-3/1	39000	98
ficar	132.81	6.93	69	-3/1	162652	391
confessar	122.54	4.84	25	-3/1	11639	148
instalar	119.35	5.31	32	-3/1	27515	127
haver	108.66	7.18	102	-3/1	417781	1415
esconder	93.44	4.47	22	-3/1	14390	452
morrer	85.01	4.77	27	-3/1	31733	273
deixar	70.85	5.27	43	-3/1	119915	336
gerar	68.47	3.91	17	-3/1	12477	242
inspirar	60.79	3.47	13	-3/1	6798	63
causar	60.01	3.85	17	-3/1	16253	650
fazer	59.75	5.77	89	-3/1	489880	2190
mostrar	57.48	4.38	26	-3/1	51696	344
revelar	56.94	4.38	26	-3/1	52331	198
borrar	54.13	2.23	5	-3/1	119	6
reinar	46.83	2.77	8	-3/1	2284	36
instilar	45.59	2.00	4	-3/1	72	8
crescer	41.61	3.40	14	-3/1	18109	102
superar	40.40	2.88	9	-3/1	5186	36
enfrentar	39.59	3.20	12	-3/1	12989	106
andar	38.16	3.35	14	-3/1	20757	56
semear	38.11	2.41	6	-3/1	1330	104
apoderar	36.68	2.41	6	-3/1	1501	20
alimentar	32.60	2.92	10	-3/1	11053	449
continuar	31.26	3.85	29	-3/1	117486	300
suscitar	31.16	2.79	9	-3/1	8920	407
desaparecer	30.03	2.97	11	-3/1	16268	57
criar	27.32	3.38	19	-3/1	60466	260
chorar	26.93	2.35	6	-3/1	3459	105
existir	26.92	3.44	21	-3/1	74025	255
aumentar	24.75	3.06	14	-3/1	36220	213
motivar	22.58	2.44	7	-3/1	7880	96
permanecer	22.44	2.64	9	-3/1	15207	30
espalhar	19.63	2.26	6	-3/1	6589	61
começar	18.80	3.14	23	-3/1	112636	186
obcecar	18.33	1.70	3	-3/1	752	8
perceber	16.81	2.48	9	-3/1	21905	75
disfarçar	15.98	1.89	4	-3/1	2984	87
atiçar	15.54	1.40	2	-3/1	217	31
experimentar	14.13	1.87	4	-3/1	3817	37
encher	13.81	2.01	5	-3/1	7260	77
manifestar	12.64	2.31	9	-3/1	29248	943
ganhar	11.91	2.45	13	-3/1	59127	250
exprimir	11.81	1.82	4	-3/1	5231	220
esqueçer	11.61	2.20	8	-3/1	25223	107
excitar	10.87	1.38	2	-3/1	708	8
afastar	9.88	2.04	7	-3/1	22651	40
diminuir	9.84	1.87	5	-3/1	11467	89
mergulhar	8.62	1.57	3	-3/1	4096	30
atacar	8.46	1.81	5	-3/1	13542	24
matar	8.38	1.80	5	-3/1	13678	53
livrar	8.19	1.34	2	-3/1	1415	25
agitar	7.22	1.52	3	-3/1	5321	14
demonstrar	7.00	1.72	5	-3/1	16293	130
aliviar	6.06	1.29	2	-3/1	2506	70
testemunhar	5.98	1.29	2	-3/1	2560	30
impor	5.62	1.62	5	-3/1	19601	144
cuidar	5.45	1.27	2	-3/1	2967	4
exercer	5.38	1.52	4	-3/1	13521	13

emergir	4.87	1.24	2	-3/1	3493	6
arrancar	4.38	1.35	3	-3/1	9408	19
assumir	4.04	1.54	7	-3/1	41594	60
despertar	3.99	1.19	2	-3/1	4512	248
nascer	3.76	1.42	5	-3/1	25744	99
considerar	3.62	-2.55	4	-3/1	128883	169
apelar	3.44	1.26	3	-3/1	11571	69
expressar	2.73	1.08	2	-3/1	6673	132
lançar	2.61	1.30	7	-3/1	50287	228
pegar	2.58	1.06	2	-3/1	7001	15
sofrer	2.53	1.24	5	-3/1	31630	253
tirar	2.50	1.14	3	-3/1	14518	55
sobreviver	2.47	1.05	2	-3/1	7268	31
levar	2.33	1.30	11	-3/1	94774	242
usar	2.32	1.20	5	-3/1	32885	36
travar	2.28	1.03	2	-3/1	7750	55
tornar	2.07	-1.97	2	-3/1	67746	140
dar	1.94	1.24	17	-3/1	168517	813
trazer	1.90	1.08	4	-3/1	26021	218
tratar	1.83	-1.81	2	-3/1	64665	108
submeter	1.79	0.95	2	-3/1	9239	9
defender	1.78	-1.78	2	-3/1	64032	90
cair	1.53	1.00	4	-3/1	28455	51
passar	1.43	1.07	13	-3/1	129693	240
manter	1.16	-1.36	2	-3/1	55665	292
entrar	1.01	0.88	7	-3/1	66123	377
produzir	0.78	0.74	3	-3/1	24378	46
correr	0.77	0.75	4	-3/1	35345	109
surgir	0.68	-0.99	2	-3/1	48224	126
conquistar	0.65	0.66	2	-3/1	15082	63
responder	0.57	0.67	4	-3/1	37844	61
controlar	0.42	-0.75	2	-3/1	43443	56
seguir	0.41	-0.69	7	-3/1	125184	195
tomar	0.36	-0.69	2	-3/1	42217	58
pôr	0.33	-0.65	2	-3/1	41440	98
chamar	0.13	0.33	4	-3/1	47211	73
chegar	0.09	-0.30	8	-3/1	125518	111
utilizar	0.00	-0.00	3	-3/1	42568	25

medos (989)	log-like	t-score	Kookkurrenz	Suchraum	CF	SF
vencer	72.41	3.40	12	-3/1	39000	98
desfazer	57.41	2.44	6	-3/1	3272	34
despertar	53.57	2.44	6	-3/1	4512	248
diminuir	51.67	2.62	7	-3/1	11467	89
afastar	42.26	2.60	7	-3/1	22651	40
acalmar	38.43	1.99	4	-3/1	2137	62
criar	35.12	2.71	8	-3/1	60466	260
superar	31.37	1.99	4	-3/1	5186	36
dissipar	31.23	1.73	3	-3/1	1072	24
provocar	27.26	2.35	6	-3/1	41836	1308
atiçar	25.59	1.41	2	-3/1	217	31
alimentar	25.38	1.97	4	-3/1	11053	449
ser	24.41	3.74	42	-3/1	3129383	9341
surgir	19.62	2.11	5	-3/1	48224	126
suscitar	18.60	1.70	3	-3/1	8920	407
viver	17.58	2.08	5	-3/1	59896	419
enfrentar	16.39	1.69	3	-3/1	12989	106
ter	15.52	2.84	19	-3/1	1163390	8933
haver	13.58	2.41	10	-3/1	417781	1415
agitar	12.84	1.39	2	-3/1	5321	14
rodear	11.74	1.39	2	-3/1	7028	11
confessar	9.78	1.37	2	-3/1	11639	148
revelar	8.47	1.56	3	-3/1	52331	198
deixar	7.54	1.66	4	-3/1	119915	336
sofrer	6.00	1.29	2	-3/1	31630	253
aumentar	5.51	1.27	2	-3/1	36220	213
controlar	4.87	1.24	2	-3/1	43443	56
representar	4.72	1.23	2	-3/1	45373	97
lançar	4.36	1.21	2	-3/1	50287	228
fazer	1.43	0.99	5	-3/1	489880	2190
estar	0.00	0.01	4	-3/1	700933	916

ódio (2347)	log-like	t-score	Kookkurrenz	Suchraum	CF	SF
incitar	250.31	4.58	21	-3/1	1498	29
atiçar	176.42	3.46	12	-3/1	217	31
destilar	174.01	3.60	13	-3/1	448	23
nutrir	168.37	3.60	13	-3/1	555	76
sentir	142.09	4.89	25	-3/1	40915	860

alimentar	108.78	3.83	15	-3/1	11053	449
fomentar	108.13	3.31	11	-3/1	2220	29
instigar	103.86	2.83	8	-3/1	336	12
ser	93.72	6.98	118	-3/1	3129383	9341
semear	69.64	2.64	7	-3/1	1330	104
desencadear	45.17	2.61	7	-3/1	7715	94
motivar	44.88	2.61	7	-3/1	7880	96
esconder	43.94	2.76	8	-3/1	14390	452
espalhar	38.76	2.41	6	-3/1	6589	61
gerar	38.57	2.58	7	-3/1	12477	242
suscitar	35.18	2.40	6	-3/1	8920	407
nascer	34.93	2.71	8	-3/1	25744	99
pregar	34.82	1.99	4	-3/1	1416	89
haver	33.01	3.75	24	-3/1	417781	1415
provocar	27.59	2.63	8	-3/1	41836	1308
ter	26.55	3.85	40	-3/1	1163390	8933
reprimir	25.85	1.72	3	-3/1	1110	15
matar	23.39	2.15	5	-3/1	13678	53
cultivar	20.58	1.71	3	-3/1	2689	25
disfarçar	19.96	1.71	3	-3/1	2984	87
criar	17.76	2.34	7	-3/1	60466	260
despertar	17.52	1.70	3	-3/1	4512	248
exprimir	16.65	1.69	3	-3/1	5231	220
manifestar	16.20	2.06	5	-3/1	29248	943
demonstrar	15.67	1.89	4	-3/1	16293	130
inspirar	15.12	1.68	3	-3/1	6798	63
valer	12.86	1.98	5	-3/1	42302	5638
testemunhar	12.31	1.39	2	-3/1	2560	30
existir	11.54	2.04	6	-3/1	74025	255
destruir	11.36	1.63	3	-3/1	13094	39
mergulhar	10.47	1.38	2	-3/1	4096	30
causar	10.15	1.61	3	-3/1	16253	650
continuar	9.99	2.05	7	-3/1	117486	300
deixar	9.77	2.04	7	-3/1	119915	336
superar	9.56	1.36	2	-3/1	5186	36
crescer	9.55	1.59	3	-3/1	18109	102
andar	8.80	1.57	3	-3/1	20757	56
chamar	7.99	1.68	4	-3/1	47211	73
trazer	7.58	1.53	3	-3/1	26021	218
instalar	7.29	1.52	3	-3/1	27515	127
olhar	7.22	1.32	2	-3/1	9574	62
ganhar	6.51	1.60	4	-3/1	59127	250
usar	6.36	1.48	3	-3/1	32885	36
levar	6.21	1.66	5	-3/1	94774	242
aumentar	5.87	1.45	3	-3/1	36220	213
conquistar	5.55	1.27	2	-3/1	15082	63
desaparecer	5.28	1.26	2	-3/1	16268	57
começar	4.96	1.56	5	-3/1	112636	186
condenar	4.31	1.21	2	-3/1	21468	830
produzir	3.88	1.18	2	-3/1	24378	46
estar	3.78	1.64	16	-3/1	700933	916
fazer	2.45	1.33	11	-3/1	489880	2190
vencer	2.40	1.04	2	-3/1	39000	98
lançar	1.69	0.94	2	-3/1	50287	228
viver	1.24	0.84	2	-3/1	59896	419
passar	0.74	0.72	3	-3/1	129693	240
dar	0.03	-0.19	2	-3/1	168517	813

ódios (659)	log-like	t-score	Kookkurrenz	Suchraum	CF	SF
atiçar	131.35	2.83	8	-3/1	217	31
despertar	120.55	3.31	11	-3/1	4512	248
suscitar	105.57	3.31	11	-3/1	8920	407
alimentar	100.87	3.30	11	-3/1	11053	449
gerar	66.32	2.81	8	-3/1	12477	242
provocar	55.17	2.95	9	-3/1	41836	1308
fomentar	41.38	2.00	4	-3/1	2220	29
semear	32.37	1.73	3	-3/1	1330	104
desencadear	31.45	1.99	4	-3/1	7715	94
concitar	28.47	1.41	2	-3/1	159	11
inflamar	21.97	1.41	2	-3/1	804	24
dissipar	20.82	1.41	2	-3/1	1072	24
haver	20.17	2.66	10	-3/1	417781	1415
esqueçer	14.89	1.68	3	-3/1	25223	107
esconder	10.53	1.38	2	-3/1	14390	452
deixar	10.34	1.77	4	-3/1	119915	336
controlar	6.33	1.30	2	-3/1	43443	56
manter	5.43	1.27	2	-3/1	55665	292
ganhar	5.21	1.26	2	-3/1	59127	250
criar	5.14	1.25	2	-3/1	60466	260

começar	3.04	1.11	2	-3/1	112636	186
fazer	0.60	0.66	3	-3/1	489880	2190
estar	0.59	0.67	4	-3/1	700933	916
ter	0.52	0.65	6	-3/1	1163390	8933
ser	0.38	0.58	14	-3/1	3129383	9341

...

Corpus: Cetempublico, 174184724 Wörter
Sample: Sentimento, 40 Nomina (Singular- und Pluralform separat)
Lemmatisierte Verben: 226
Gesamtzahl der Fundstellen der untersuchten Nomina: 159701
Fundstellen mit einem Verb im Suchraum: 53362
Signifikanzgrenze = 1

admiração	t-socre		log-likelihood			chi-square			MI	
ter	9.91	120	esconder	596.73	59	nutrir	143922.18	28	nutrir	8.55	28
esconder	7.66	59	nutrir	424.44	28	esconder	24539.75	59	concitar	7.56	3
manifestar	5.52	31	ter	354.56	120	suscitar	9552.13	29	esconder	6.04	59
suscitar	5.37	29	suscitar	279.35	29	expressar	6069.87	20	suscitar	5.80	29
causar	5.36	29	causar	244.65	29	confessar	5868.25	26	exprimir	5.80	17
nutrir	5.29	28	confessar	230.93	26	concitar	5763.23	3	expressar	5.72	20
confessar	5.08	26	manifestar	229.47	31	exprimir	5597.22	17	testemunhar	5.63	7
merecer	4.76	23	expressar	189.30	20	causar	5216.58	29	confessar	5.43	26
provocar	4.71	23	merecer	181.62	23	manifestar	3287.72	31	comover	5.24	2
expressar	4.46	20	exprimir	163.62	17	merecer	3149.19	23	causar	5.20	29
exprimir	4.11	17	provocar	140.33	23	testemunhar	1936.92	7	merecer	4.93	23
mostrar	3.61	14	mostrar	66.03	14	provocar	1243.49	23	despertar	4.91	6
sentir	3.49	13	sentir	65.32	13	ter	1039.92	120	manifestar	4.68	31
ganhar	3.44	13	testemunhar	64.92	7	despertar	801.28	6	disfarçar	4.63	3
ser	2.95	52	ganhar	56.10	13	sentir	395.49	13	incitar	4.22	1
testemunhar	2.64	7	despertar	47.03	6	comover	373.51	2	olhar	4.16	6
conquistar	2.59	7	conquistar	40.32	7	olhar	371.36	6	inspirar	4.09	4
despertar	2.43	6	concitar	39.44	3	mostrar	359.04	14	proclamar	4.03	2
olhar	2.41	6	olhar	38.09	6	conquistar	317.31	7	provocar	4.03	23
demonstrar	2.38	6	demonstrar	31.84	6	disfarçar	301.44	3	conquistar	3.86	7
inspirar	1.97	4	inspirar	24.90	4	ganhar	265.99	13	guardar	3.78	3
gerar	1.94	4	disfarçar	21.84	3	inspirar	231.97	4	demonstrar	3.63	6
haver	1.87	10	gerar	20.15	4	demonstrar	213.38	6	gerar	3.49	4
ficar	1.80	6	comover	16.99	2	guardar	126.17	3	sentir	3.48	13
concitar	1.73	3	guardar	16.85	3	gerar	122.83	4	desfazer	3.44	1
disfarçar	1.72	3	alimentar	14.14	3	proclamar	108.94	2	alimentar	3.32	3
guardar	1.69	3	ser	12.47	52	alimentar	77.10	3	mostrar	3.32	14
alimentar	1.67	3	proclamar	12.21	2	incitar	66.05	1	ganhar	3.11	13
tornar	1.67	4	declarar	8.72	3	desfazer	29.18	1	render	3.10	1
declarar	1.57	3	tornar	7.69	4	declarar	26.02	3	poupar	2.71	1
assumir	1.50	3	ficar	7.10	6	render	20.28	1	rodear	2.67	1
revelar	1.44	3	assumir	6.79	3	tornar	16.74	4	sobreviver	2.64	1
manter	1.42	3	incitar	6.47	1	assumir	16.47	3	desencadear	2.58	1
comover	1.41	2	haver	6.06	10	ser	15.04	52	declarar	2.36	3
proclamar	1.39	2	revelar	5.62	3	poupar	13.16	1	ter	2.35	120
receber	1.29	3	manter	5.32	3	rodear	12.57	1	assumir	1.99	3
aumentar	1.16	2	desfazer	4.94	1	ficar	12.17	6	transmitir	1.90	1
valer	1.12	2	render	4.29	1	sobreviver	12.10	1	restar	1.88	1
deixar	1.05	3	receber	3.77	3	revelar	12.05	3	tornar	1.79	4
incitar	0.99	1	aumentar	3.62	2	desencadear	11.29	1	prestar	1.78	1
desfazer	0.97	1	poupar	3.56	1	manter	11.03	3	revelar	1.77	3
...											

Wortfeld: Sentimento Corpus: Cetempublico

		t-score	MI	Kookkurrenz
valer	pena	74.45	6.94	5553
ter	medo	62.39	3.90	4055
ser	pena	40.16	1.77	2340
cumprir	pena	31.93	5.47	1028
ter	esperança	26.98	2.46	870
ter	pena	23.91	1.74	841
ter	vergonha	22.09	3.25	528
fazer	amor	21.88	2.85	539
ser	vergonha	21.30	2.32	558
condenar	pena	21.25	5.12	457
manifestar	esperança	20.18	5.39	411
ter	esperanças	19.13	2.57	429
condenar	penas	18.63	6.57	348
ser	amor	18.56	1.24	684
fazer	inveja	18.34	4.64	343
depositar	esperanças	18.27	7.76	334
meter	medo	17.93	6.11	323
ser	esperança	17.41	1.16	640
entrar	pânico	17.02	5.58	292
haver	esperança	16.64	2.51	328
aplicar	pena	16.52	4.68	278
ser	paixão	16.08	1.52	424
merecer	respeito	15.79	4.71	254
ser	desilusão	15.65	2.22	308
incorrer	pena	15.35	7.34	236
ter	respeito	14.71	1.11	481
perder	esperança	14.62	3.88	223
ser	alegria	14.61	1.45	365
fazer	furor	14.58	5.19	215
cumprir	penas	14.30	5.58	206
ser	medo	14.13	0.92	553
alimentar	esperanças	13.43	6.37	181
fazer	pena	13.09	1.51	282
ser	dor	13.01	1.48	284
alimentar	esperança	12.95	5.48	169
provocar	pânico	12.91	5.49	168
sentir	dores	12.56	5.45	159
lançar	pânico	12.46	5.24	157
aplicar	penas	12.41	5.81	155
estar	medo	12.14	1.56	236
restar	esperança	12.13	5.01	149
provocar	agitação	12.11	5.31	148
perder	esperanças	12.08	4.30	150
provocar	ira	12.07	6.59	146
ser	decepção	12.05	2.26	181
ter	paixão	11.91	1.78	205
ter	dores	11.76	2.20	175
sentir	dor	11.55	5.07	135
provocar	indignação	11.47	5.29	133
apanhar	susto	11.44	7.34	131
dar	alegria	11.13	3.36	133
acalentar	esperanças	10.91	8.93	119
morrer	amores	10.66	6.27	114
manifestar	indignação	10.63	5.49	114
dar	dores	10.56	3.73	117
ganhar	susto	10.54	5.57	112
receber	entusiasmo	10.33	3.96	111
ser	tristeza	10.22	1.61	163
manter	esperança	10.15	3.43	110
exigir	respeito	10.10	3.18	111
provocar	dores	9.93	4.96	100
ter	admiração	9.91	2.35	120
ser	entusiasmo	9.77	1.01	236
dar	esperança	9.67	2.36	114
continuar	pena	9.44	2.06	117
manifestar	apreensão	9.39	5.30	89
causar	pânico	9.19	5.75	85
sentir	medo	9.07	3.42	88
corar	vergonha	9.05	8.51	82
suscitar	entusiasmo	9.03	5.81	82
perder	medo	9.00	2.88	91
impor	respeito	8.99	3.47	86
sentir	vergonha	8.98	4.74	82
causar	agitação	8.86	5.63	79
causar	apreensão	8.86	5.77	79

...

Wortfeld: Sentimento Corpus: Cetempublico

		log-like	t-score	Kookkurrenz
valer	pena	68281.74	74.45	5553
ter	medo	25159.51	62.39	4055
cumprir	pena	9282.95	31.93	1028
ser	pena	4594.03	40.16	2340
depositar	esperanças	4559.35	18.27	334
condenar	penas	3911.51	18.63	348
condenar	pena	3792.14	21.25	457
manifestar	esperança	3637.25	20.18	411
meter	medo	3324.06	17.93	323
incorrer	pena	3042.85	15.35	236
ter	esperança	2750.62	26.98	870
entrar	pânico	2710.92	17.02	292
fazer	inveja	2612.49	18.34	343
ter	vergonha	2511.00	22.09	528
fazer	amor	2083.88	21.88	539
aplicar	pena	2059.94	16.52	278
alimentar	esperanças	1953.08	13.43	181
fazer	furor	1936.45	14.58	215
acalentar	esperanças	1918.49	10.91	119
cumprir	penas	1900.80	14.30	206
merecer	respeito	1893.50	15.79	254
apanhar	susto	1677.03	11.44	131
ser	vergonha	1671.87	21.30	558
provocar	ira	1658.83	12.07	146
ter	pena	1561.73	23.91	841
alimentar	esperança	1520.52	12.95	169
provocar	pânico	1520.26	12.91	168
aplicar	penas	1500.65	12.41	155
ter	esperanças	1448.61	19.13	429
sentir	dores	1425.99	12.56	159
lançar	pânico	1341.28	12.46	157
perder	esperança	1296.69	14.62	223
provocar	agitação	1285.50	12.11	148
corar	vergonha	1240.90	9.05	82
morrer	amores	1214.12	10.66	114
restar	esperança	1201.74	12.13	149
provocar	indignação	1149.26	11.47	133
semear	pânico	1125.41	8.83	78
sentir	dor	1106.52	11.55	135
haver	esperança	1052.39	16.64	328
ganhar	susto	1033.89	10.54	112
manifestar	indignação	1030.63	10.63	114
pregar	susto	1014.37	8.06	65
perder	esperanças	998.24	12.08	150
acalentar	esperança	931.42	8.36	70
ser	desilusão	862.54	15.65	308
causar	pânico	811.54	9.19	85
provocar	dores	797.95	9.93	100
despertar	paixões	792.36	7.87	62
suscitar	entusiasmo	791.95	9.03	82
manifestar	apreensão	770.01	9.39	89
causar	apreensão	757.05	8.86	79
ser	amor	743.09	18.56	684
causar	agitação	734.46	8.86	79
causar	dores	684.86	8.57	74
despertar	entusiasmo	673.57	7.98	64
receber	entusiasmo	663.22	10.33	111
dar	dores	648.63	10.56	117
ser	paixão	647.47	16.08	424
dar	alegria	640.59	11.13	133
ser	esperança	628.42	17.41	640
abolir	pena	625.64	7.98	64
gerar	pânico	619.63	8.04	65
roer	inveja	619.08	6.08	37
sentir	vergonha	616.96	8.98	82
esconder	admiração	596.73	7.66	59
depositar	esperança	593.02	8.21	68
incorrer	penas	577.81	6.70	45
causar	indignação	556.69	7.84	62
sofrer	dores	547.12	8.30	70
manter	esperança	543.62	10.15	110
ser	decepção	521.93	12.05	181
ser	alegria	515.62	14.61	365
vestir	luto	515.18	6.92	48
exigir	respeito	493.65	10.10	111

...

Wortfeld: Sentimento Corpus: Cetempublico

		chi-square	t-score	Kookkurrenz
valer	pena	5733012.09	74.45	5553
acalentar	esperanças	901929.18	10.91	119
depositar	esperanças	779848.04	18.27	334
pregar	susto	411376.50	8.06	65
roer	inveja	409353.34	6.08	37
corar	vergonha	406098.30	9.05	82
incorrer	pena	361870.75	15.35	236
semear	pânico	275558.84	8.83	78
condenar	penas	246472.91	18.63	348
cumprir	pena	242696.31	31.93	1028
apanhar	susto	201778.85	11.44	131
ter	medo	193622.33	62.39	4055
meter	medo	145435.71	17.93	323
nutrir	admiração	143922.18	5.29	28
acalentar	esperança	137507.92	8.36	70
provocar	ira	105513.99	12.07	146
alimentar	esperanças	105011.51	13.43	181
despertar	paixões	97508.81	7.87	62
manifestar	esperança	89724.96	20.18	411
render	encantos	81120.02	5.29	28
atiçar	ódios	77939.55	2.83	8
entrar	pânico	77166.21	17.02	292
condenar	pena	75741.71	21.25	457
incorrer	penas	73168.42	6.70	45
pregar	sustos	66131.64	3.32	11
morrer	amores	60135.47	10.66	114
cumprir	penas	54240.81	14.30	206
aplicar	penas	51635.03	12.41	155
atiçar	ódio	49225.99	3.46	12
provocar	pânico	40336.25	12.91	168
alimentar	esperança	40169.36	12.95	169
corar	inveja	40033.41	4.00	16
fazer	furor	38261.11	14.58	215
causar	furor	38190.20	6.24	39
sentir	dores	36715.32	12.56	159
fazer	inveja	34896.54	18.34	343
despertar	entusiasmo	32912.97	7.98	64
provocar	agitação	29711.57	12.11	148
aplicar	pena	29495.44	16.52	278
lançar	pânico	29238.82	12.46	157
ganhar	susto	29045.31	10.54	112
destilar	ódio	27971.08	3.60	13
merecer	respeito	27584.15	15.79	254
manifestar	indignação	27498.46	10.63	114
suscitar	entusiasmo	27273.08	9.03	82
vestir	luto	27246.14	6.92	48
causar	pânico	26625.89	9.19	85
apanhar	sustos	26394.75	4.47	20
provocar	indignação	26119.70	11.47	133
causar	apreensão	25189.91	8.86	79
esconder	admiração	24539.75	7.66	59
sucumbir	encantos	23912.39	2.64	7
esconder	decepção	23301.22	6.70	45
nutrir	ódio	22573.46	3.60	13
abolir	pena	22344.39	7.98	64
restar	esperança	22146.39	12.13	149
causar	agitação	21846.44	8.86	79
incitar	ódio	21807.11	4.58	21
sentir	dor	21258.62	11.55	135
aliviar	dor	20931.51	5.74	33
gerar	pânico	20281.32	8.04	65
causar	dores	20042.54	8.57	74
suscitar	entusiasmos	19762.37	3.87	15
renascer	esperança	18409.19	6.39	41
manifestar	apreensão	17701.02	9.39	89
suscitar	paixões	17514.66	6.07	37
inflamar	paixões	17224.53	3.31	11
exultar	alegria	16139.25	3.46	12
infundir	respeito	15612.92	3.16	10
aliviar	dores	14894.40	4.99	25
causar	indignação	14633.03	7.84	62
depositar	esperança	14129.83	8.21	68
provocar	dores	14126.69	9.93	100
instigar	ódio	14120.58	2.83	8
suscitar	apreensões	14118.82	4.89	24

...

Wortfeld: Sentimento Corpus: Cetempublico

		MI	chi-square	Kookkurrenz
roer	inveja	9.31	409353.34	37
atiçar	ódios	9.18	77939.55	8
ferver	indignações	9.02	8248.53	1
acalentar	esperanças	8.93	901929.18	119
pregar	susto	8.75	411376.50	65
pregar	sustos	8.70	66131.64	11
acalmar	furores	8.60	5432.00	1
nutrir	admiração	8.55	143922.18	28
corar	vergonha	8.51	406098.30	82
arrebatar	desesperos	8.47	4779.12	1
acometer	invejas	8.44	4620.38	1
atiçar	ódio	8.32	49225.99	12
fingir	cóleras	8.25	3807.85	1
semear	pânico	8.17	275558.84	78
sucumbir	encantos	8.14	23912.39	7
concitar	ódios	8.11	6645.51	2
incorrer	iras	8.09	3267.01	1
afogar	tristezas	8.04	12383.52	4
acalmar	pânicos	8.01	3016.89	1
render	encantos	7.97	81120.02	28
atiçar	paixões	7.88	13192.35	5
corar	inveja	7.83	40033.41	16
livrar	aflições	7.81	7379.99	3
suscitar	admirações	7.80	4878.11	2
infundir	paixões	7.80	2436.22	1
despertar	admirações	7.79	2410.86	1
depositar	esperanças	7.76	779848.04	334
crivar	ciúmes	7.67	2140.98	1
destilar	ódio	7.67	27971.08	13
evaporar	encanto	7.60	3983.91	2
acalentar	esperança	7.58	137507.92	70
concitar	admiração	7.56	5763.23	3
atiçar	furor	7.54	1882.27	1
concitar	amores	7.52	3703.31	2
acalmar	iras	7.52	1850.50	1
acalmar	raivas	7.52	1850.50	1
instigar	ódio	7.48	14120.58	8
acalmar	excitações	7.46	1732.26	1
nutrir	ódio	7.46	22573.46	13
acometer	ciúmes	7.43	1679.83	1
excitar	aflições	7.40	1638.17	1
incorrer	penas	7.40	73168.42	45
roer	ciúmes	7.39	3223.48	2
atiçar	medos	7.39	3242.51	2
encher	furores	7.38	1597.56	1
despertar	paixões	7.36	97508.81	62
inflamar	paixões	7.36	17224.53	11
livrar	fúrias	7.35	1556.23	1
infundir	respeito	7.35	15612.92	10
incorrer	pena	7.34	361870.75	236
apanhar	susto	7.34	201778.85	131
despertar	invejas	7.28	8728.94	6
despertar	entusiasmos	7.24	11113.61	8
acarretar	agitações	7.23	1375.45	1
exultar	alegria	7.21	16139.25	12
repassar	tristeza	7.21	1353.43	1
suscitar	iras	7.19	3988.46	3
apanhar	sustos	7.19	26394.75	20
suscitar	entusiasmos	7.18	19762.37	15
acometer	dores	7.14	6310.69	5
borrar	inveja	7.12	1239.52	1
destilar	raiva	7.11	4871.18	4
insinuar	vergonhas	7.08	1185.56	1
invadir	asco	7.06	1165.91	1
roer	ciúme	7.04	1142.81	1
evaporar	excitação	7.00	1089.27	1
dissipar	aflições	6.99	1081.25	1
sobreviver	lutos	6.99	1087.41	1
incitar	ódio	6.95	21807.11	21
acalmar	fúrias	6.94	1029.77	1
valer	pena	6.94	5733012.09	5553
dissipar	decepções	6.90	988.77	1
causar	furor	6.89	38190.20	39
suscitar	comoções	6.88	974.42	1
germinar	excitação	6.81	901.96	1

...

Kookkurrenzen sortiert nach Verben und Kollokationspotenzial

...

abolir	t-score	MI	Kookkurrenz	Nomen	abolir	1436
pena	7.98	5.86	64	22112		
			64			

abrandar	t-score	MI	Kookkurrenz	Nomen	abrandar	1663
entusiasmo	1.38	3.78	2	4786		
pena	1.26	2.25	2	22112		
ira	0.99	4.83	1	839		
agitação	0.97	3.54	1	3040		
esperanças	0.95	3.06	1	4900		
			7			

acalentar	t-score	MI	Kookkurrenz	Nomen	acalentar	558
esperanças	10.91	8.93	119	4900		
esperança	8.36	7.58	70	11113		
ódios	1.00	6.16	1	659		
paixões	1.00	5.32	1	1520		
raiva	1.00	5.50	1	1275		
paixão	0.98	4.10	1	5170		
amor	0.96	3.34	1	11054		
medo	0.96	3.23	1	12288		
			195			

acalmar	t-score	MI	Kookkurrenz	Nomen	acalmar	2137
dores	2.82	5.41	8	2907		
ira	2.82	6.66	8	839		
fúria	2.64	5.81	7	1714		
agitação	2.43	5.08	6	3040		
medos	1.99	5.80	4	989		
apreensões	1.73	5.73	3	794		
paixões	1.72	5.08	3	1520		
indignação	1.71	4.47	3	2792		
entusiasmo	1.37	3.53	2	4786		
entusiasmos	1.00	5.91	1	222		
excitações	1.00	7.46	1	47		
fúrias	1.00	6.94	1	79		
furores	1.00	8.60	1	15		
iras	1.00	7.52	1	44		
pânicos	1.00	8.01	1	27		
raivas	1.00	7.52	1	44		
ciúmes	0.99	5.23	1	437		
estimação	0.99	5.26	1	422		
excitação	0.99	4.73	1	719		
furor	0.99	5.25	1	426		
ódios	0.99	4.82	1	659		
raiva	0.98	4.16	1	1275		
desespero	0.96	3.28	1	3082		
dor	0.96	3.12	1	3605		
pânico	0.96	3.34	1	2890		
paixão	0.94	2.76	1	5170		
medo	0.85	1.89	1	12288		
			62			

acarretar	t-score	MI	Kookkurrenz	Nomen	abrandar	1916
pena	3.09	3.72	10	22112		
penas	1.98	4.52	4	3976		
agitações	1.00	7.23	1	66		
inveja	0.99	4.35	1	1179		
fúria	0.98	3.97	1	1714		
tristeza	0.98	3.92	1	1810		
agitação	0.97	3.40	1	3040		
apreensão	0.97	3.54	1	2639		
dor	0.96	3.23	1	3605		
			21			

acometer	t-score	MI	Kookkurrenz	Nomen	acometer	237
dores	2.23	7.14	5	2907		
ciúmes	1.00	7.43	1	437		
fúria	1.00	6.06	1	1714		
invejas	1.00	8.44	1	159		
pânico	1.00	5.54	1	2890		
			9			

acreditar	t-score	MI	Kookkurrenz	Nomen	acreditar	27702
pena	3.69	1.74	20	22112		
...						

	Corpus	Sample	Samplerelevanz
acalentar	558	195	34.9 %
infundir	47	15	31.9 %
incorrer	1211	287	23.7 %
atiçar	217	31	14.3 %
nutrir	555	76	13.7 %
valer	42302	5638	13.3 %
corar	944	109	11.5 %
instilar	72	8	11.1 %
roer	494	44	8.9 %
depositar	5081	408	8.0 %
semear	1330	104	7.8 %
concitar	159	11	6.9 %
pregar	1416	89	6.3 %
despertar	4512	248	5.5 %
destilar	448	23	5.1 %
borrar	119	6	5.0 %
renascer	1425	66	4.6 %
suscitar	8920	407	4.6 %
abolir	1436	64	4.5 %
exprimir	5231	220	4.2 %
alimentar	11053	449	4.1 %
causar	16253	650	4.0 %
condenar	21468	830	3.9 %
acometer	237	9	3.8 %
meter	10125	381	3.8 %
cumprir	34024	1271	3.7 %
exultar	325	12	3.7 %
instigar	336	12	3.6 %
desvanecer	528	18	3.4 %
evaporar	222	7	3.2 %
manifestar	29248	943	3.2 %
esconder	14390	452	3.1 %
provocar	41836	1308	3.1 %
chorar	3459	105	3.0 %
ferver	406	12	3.0 %
inflamar	804	24	3.0 %
acalmar	2137	62	2.9 %
disfarçar	2984	87	2.9 %
aliviar	2506	70	2.8 %
repassar	71	2	2.8 %
tremer	1601	40	2.5 %
aplicar	20265	438	2.2 %
arrebatar	828	18	2.2 %
dissipar	1072	24	2.2 %
merecer	16877	373	2.2 %
sentir	40915	860	2.1 %
expressar	6673	132	2.0 %
gerar	12477	242	1.9 %
incitar	1498	29	1.9 %
livrar	1415	25	1.8 %
apanhar	11715	204	1.7 %
reinar	2284	36	1.6 %
fingir	1524	21	1.4 %
reprimir	1110	15	1.4 %
sucumbir	972	14	1.4 %
apoderar	1501	20	1.3 %
confessar	11639	148	1.3 %
fomentar	2220	29	1.3 %
saciar	225	3	1.3 %
desencadear	7715	94	1.2 %
iludir	1765	22	1.2 %
motivar	7880	96	1.2 %
restar	15507	182	1.2 %
testemunhar	2560	30	1.2 %
acarretar	1916	21	1.1 %
afogar	1190	13	1.1 %
crivar	186	2	1.1 %
encher	7260	77	1.1 %
excitar	708	8	1.1 %
obcecar	752	8	1.1 %
perder	72509	826	1.1 %
restaurar	2359	25	1.1 %
agrar	2494	25	1.0 %
desfazer	3272	34	1.0 %
experimentar	3817	37	1.0 %
jurar	1261	12	1.0 %
comover	1080	10	0.9 %
cultivar	2689	25	0.9 %

...

Ergebnis K-Means (Nomina)

Nachdem K-Means 100 mal durchlaufen wurde, kommen die Nomina mit folgenden weiteren Nomina X-mal im gleichen Cluster vor:

admiração:
100-admiração 63-paixão 58-compaixão 57-respeito 57-esperança 54-estimação 43-desilusão 36-raiva 33-dor 31-alegria 30-decepção 29-ódio 27-tristeza 19-ciúme 19-desesperança 14-entusiasmo 11-encanto 11-pena 10-desespero 10-fúria 9-susto 9-asco 9-aflição 9-apreensão 8-luto 8-enfado 7-inclinação 6-medo 6-vergonha 3-indignação 2-amor 2-pânico 1-agitação 1-cólera 1-excitação 1-comoção

aflição:
100-aflição 82-desespero 59-luto 57-entusiasmo 55-susto 52-ódio 43-pena 42-tristeza 42-apreensão 41-desesperança 40-raiva 40-alegria 33-dor 27-asco 26-pânico 25-desilusão 22-decepção 17-enfado 15-alvoroço 14-compaixão 13-estimação 12-paixão 12-fúria 10-amor 9-respeito 9-esperança 9-indignação 9-admiração 8-agitação 8-cólera 7-excitação 6-furor 5-inveja 2-comoção 1-ira

agitação:
100-agitação 89-excitação 83-comoção 78-cólera 76-alvoroço 62-fúria 58-ira 47-pânico 46-indignação 34-enfado 10-susto 8-aflição 7-desespero 7-decepção 6-apreensão 5-luto 4-ódio 4-desilusão 4-entusiasmo 3-dor 2-tristeza 2-desesperança 2-pena 1-raiva 1-asco 1-paixão 1-alegria 1-compaixão 1-admiração

alegria:
100-alegria 86-raiva 65-ódio 64-dor 50-tristeza 47-desespero 45-entusiasmo 44-compaixão 41-paixão 40-susto 40-aflição 38-desilusão 37-pena 35-estimação 34-respeito 34-esperança 31-desesperança 31-admiração 29-asco 28-luto 27-apreensão 24-decepção 13-amor 9-enfado 7-fúria 4-furor 4-inveja 3-ciúme 3-pânico 2-encanto 1-alvoroço 1-agitação 1-medo 1-cólera 1-vergonha 1-excitação 1-comoção 1-indignação 1-inclinação

alvoroço:
100-alvoroço 76-agitação 72-excitação 71-cólera 66-comoção 60-pânico 60-fúria 48-ira 41-indignação 30-enfado 17-susto 15-desespero 15-aflição 9-luto 9-entusiasmo 9-apreensão 8-decepção 7-tristeza 7-ódio 6-dor 5-desesperança 5-desilusão 4-asco 4-pena 1-furor 1-raiva 1-alegria

amor:
100-amor 78-furor 76-inveja 13-luto 13-alegria 13-pena 10-raiva 10-asco 10-aflição 9-desesperança 8-desespero 7-ódio 6-tristeza 6-dor 6-paixão 6-compaixão 6-apreensão 5-respeito 5-susto 4-entusiasmo 4-estimação 3-desilusão 2-esperança 2-decepção 2-admiração 1-enfado 1-indignação

apreensão:
100-apreensão 52-entusiasmo 51-desespero 47-tristeza 42-aflição 41-indignação 37-asco 35-susto 34-ódio 33-desilusão 32-decepção 28-raiva 27-alegria 26-dor 24-pena 22-fúria 21-luto 20-enfado 17-pânico 9-alvoroço 9-desesperança 9-admiração 6-agitação 6-excitação 6-amor 6-compaixão 5-cólera 5-comoção 4-furor 3-respeito 3-paixão 3-inveja 2-esperança 2-estimação 1-ira

asco:
100-asco 42-tristeza 37-apreensão 35-raiva 33-entusiasmo 32-desespero 31-dor 31-ódio 29-alegria 27-aflição 24-susto 24-pena 23-desilusão 17-decepção 15-indignação 14-enfado 13-luto 13-compaixão 10-amor 9-paixão 9-admiração 8-respeito 8-estimação 7-esperança 7-desesperança 7-inveja 6-furor 6-fúria 5-pânico 4-alvoroço 3-ira 3-comoção 1-agitação 1-cólera 1-excitação

ciúme:
100-ciúme 71-encanto 70-inclinação 53-medo 53-vergonha 24-estimação 23-esperança 21-respeito 20-paixão 19-admiração 17-compaixão 5-dor 5-desesperança 3-raiva 3-alegria 2-ódio 2-pena 1-desilusão 1-fúria

cólera:
100-cólera 82-excitação 80-comoção 78-agitação 71-alvoroço 67-fúria 60-ira 47-indignação 45-pânico 32-enfado 14-susto 8-desespero 8-aflição 7-entusiasmo 6-luto 6-decepção 5-dor 5-ódio 5-apreensão 4-desesperança 4-desilusão 3-pena 2-tristeza 1-raiva 1-asco 1-paixão 1-alegria 1-compaixão 1-admiração

comoção:
100-comoção 86-excitação 83-agitação 80-cólera 66-alvoroço 66-ira 62-fúria 50-indignação 37-pânico 34-enfado 7-susto 5-decepção 5-apreensão 3-desespero 3-asco 3-desilusão 3-entusiasmo 2-dor 2-luto 2-ódio 2-aflição 1-tristeza 1-raiva 1-paixão 1-alegria 1-compaixão 1-admiração 1-pena

compaixão:
100-compaixão 91-paixão 80-respeito 77-esperança 77-estimação 58-admiração 52-raiva 44-alegria 40-dor 37-ódio 35-desesperança 20-desilusão 20-pena 19-tristeza 17-ciúme 16-desespero 16-entusiasmo 14-aflição 13-luto 13-asco 11-susto 9-encanto

8-decepção 6-amor 6-apreensão 4-inclinação 3-furor 3-inveja 2-medo 2-vergonha 1-agitação 1-cólera 1-excitação 1-comoção 1-fúria

decepção:
100-decepção 80-desilusão 59-tristeza 45-enfado 37-entusiasmo 32-apreensão 31-ódio 30-desespero 30-admiração 29-dor 26-raiva 24-susto 24-alegria 23-fúria 22-aflição 17-asco 16-indignação 15-luto 11-pânico 9-pena 8-alvoroço 8-desesperança 8-paixão 8-compaixão 7-respeito 7-agitação 7-excitação 7-estimação 6-cólera 6-esperança 5-comoção 2-amor 1-ira

desesperança:
100-desesperança 41-aflição 38-estimação 37-paixão 36-luto 35-compaixão 35-pena 33-desespero 33-ódio 32-respeito 32-raiva 31-esperança 31-alegria 29-susto 25-entusiasmo 21-dor 19-admiração 16-tristeza 12-desilusão 9-amor 9-pânico 9-apreensão 8-decepção 7-asco 6-furor 6-inveja 5-alvoroço 5-ciúme 4-cólera 4-enfado 3-encanto 3-excitação 3-fúria 2-agitação 1-indignação

desespero:
100-desespero 82-aflição 74-entusiasmo 62-ódio 60-susto 55-tristeza 51-apreensão 49-luto 47-raiva 47-alegria 44-pena 42-dor 33-desesperança 32-asco 32-desilusão 30-decepção 27-pânico 21-enfado 17-fúria 16-compaixão 15-alvoroço 13-paixão 12-estimação 11-indignação 10-respeito 10-admiração 9-esperança 8-cólera 8-amor 7-agitação 7-excitação 5-furor 4-inveja 3-comoção 2-ira

desilusão:
100-desilusão 80-decepção 69-tristeza 44-ódio 43-admiração 41-dor 41-raiva 39-entusiasmo 38-alegria 33-apreensão 32-desespero 31-enfado 27-susto 25-aflição 23-asco 22-fúria 20-compaixão 19-paixão 17-luto 16-estimação 15-pena 14-esperança 13-respeito 12-desesperança 10-indignação 9-pânico 5-alvoroço 4-agitação 4-cólera 4-excitação 3-amor 3-comoção 1-ciúme 1-medo 1-vergonha 1-encanto 1-inclinação

dor:
100-dor 72-raiva 64-alegria 59-ódio 55-tristeza 42-desespero 41-desilusão 41-entusiasmo 40-compaixão 35-susto 35-paixão 33-admiração 33-aflição 31-asco 30-estimação 29-respeito 29-esperança 29-decepção 27-pena 26-apreensão 22-luto 21-desesperança 19-enfado 16-fúria 11-pânico 6-alvoroço 6-amor 5-ciúme 5-cólera 5-indignação 4-excitação 3-agitação 3-encanto 2-comoção 1-medo 1-ira 1-vergonha 1-inveja 1-inclinação

encanto:
100-encanto 78-inclinação 71-ciúme 60-vergonha 58-medo 15-respeito 14-esperança 14-estimação 13-paixão 11-admiração 9-compaixão 3-dor 3-desesperança 2-raiva 2-alegria 2-pena 1-ódio 1-desilusão

enfado:
100-enfado 49-fúria 45-decepção 39-indignação 37-excitação 34-agitação 34-comoção 33-tristeza 32-cólera 31-desilusão 30-alvoroço 25-entusiasmo 21-desespero 21-susto 20-ira 20-pânico 20-apreensão 19-dor 17-aflição 15-ódio 14-asco 10-raiva 9-alegria 8-luto 8-admiração 8-pena 4-desesperança 1-amor

entusiasmo:
100-entusiasmo 74-desespero 65-tristeza 63-ódio 57-aflição 52-susto 52-apreensão 47-raiva 45-alegria 41-dor 39-desilusão 37-decepção 36-pena 33-asco 31-luto 25-desesperança 25-enfado 17-pânico 17-fúria 16-compaixão 14-indignação 14-admiração 13-paixão 13-estimação 9-respeito 9-alvoroço 8-esperança 7-cólera 5-excitação 4-agitação 4-amor 3-ira 3-comoção

esperança:
100-esperança 83-paixão 82-respeito 77-compaixão 74-estimação 57-admiração 40-raiva 34-alegria 31-desesperança 29-dor 26-ódio 23-ciúme 17-pena 14-encanto 14-desilusão 11-luto 10-tristeza 10-susto 9-desespero 9-inclinação 9-aflição 8-entusiasmo 7-medo 7-vergonha 7-asco 6-decepção 2-amor 2-apreensão

estimação:
100-estimação 83-paixão 77-compaixão 75-respeito 74-esperança 54-admiração 41-raiva 38-desesperança 35-alegria 32-ódio 30-dor 24-ciúme 20-pena 16-desilusão 15-luto 14-tristeza 14-encanto 13-aflição 13-entusiasmo 12-desespero 11-susto 9-inclinação 8-asco 7-medo 7-vergonha 7-decepção 4-amor 2-apreensão 1-furor 1-inveja

excitação:
100-excitação 89-agitação 86-comoção 82-cólera 72-alvoroço 65-fúria 58-ira 51-indignação 43-pânico 37-enfado 11-susto 7-desespero 7-decepção 7-aflição 6-apreensão 5-ódio 5-entusiasmo 4-dor 4-luto 4-desilusão 3-desesperança 3-pena 2-tristeza 1-raiva 1-asco 1-paixão 1-alegria 1-compaixão 1-admiração

fúria:
100-fúria 67-cólera 65-excitação 62-agitação 62-comoção 60-alvoroço 58-indignação 50-pânico 49-enfado 41-ira 23-susto 23-decepção 22-desilusão 22-apreensão 20-tristeza 17-desespero 17-entusiasmo 16-dor 13-ódio 12-aflição 10-admiração 7-alegria 6-raiva 6-luto 6-asco 6-pena 3-desesperança 1-ciúme 1-paixão 1-compaixão

furor:
100-furor 90-inveja 78-amor 9-pena 8-luto 6-desesperança 6-asco 6-aflição 5-desespero 4-alegria 4-apreensão 3-paixão 3-compaixão 2-respeito 2-raiva 2-susto 1-alvoroço 1-ódio 1-indignação 1-estimação

inclinação:
100-inclinação 78-encanto 73-vergonha 71-medo 70-ciúme 10-respeito 9-esperança 9-estimação 8-paixão 7-admiração 4-compaixão 1-dor 1-raiva 1-ódio 1-alegria 1-desilusão 1-pena

indignação:
100-indignação 58-fúria 51-excitação 50-comoção 47-cólera 46-agitação 41-alvoroço 41-apreensão 39-enfado 35-ira 28-pânico 16-decepção 15-asco 14-entusiasmo 13-tristeza 11-desespero 10-desilusão 9-susto 9-aflição 5-dor 3-admiração 3-pena 2-luto 2-ódio 1-furor 1-raiva 1-desesperança 1-amor 1-alegria

inveja:
100-inveja 90-furor 76-amor 9-pena 7-luto 7-asco 6-desesperança 5-aflição 4-desespero 4-alegria 3-raiva 3-paixão 3-compaixão 3-apreensão 2-respeito 2-susto 1-tristeza 1-dor 1-ódio 1-estimação

ira:
100-ira 66-comoção 60-cólera 58-agitação 58-excitação 48-alvoroço 41-fúria 35-indignação 27-pânico 20-enfado 4-susto 3-asco 3-entusiasmo 2-desespero 1-dor 1-luto 1-ódio 1-decepção 1-aflição 1-apreensão 1-pena

luto:
100-luto 59-aflição 49-desespero 36-desesperança 36-susto 35-ódio 31-entusiasmo 30-raiva 30-pena 28-alegria 22-dor 21-tristeza 21-apreensão 17-desilusão 15-decepção 15-estimação 13-amor 13-asco 13-pânico 13-compaixão 12-paixão 11-esperança 10-respeito 9-alvoroço 8-furor 8-enfado 8-admiração 7-inveja 6-cólera 6-fúria 5-agitação 4-excitação 2-comoção 2-indignação 1-ira

medo:
100-medo 91-vergonha 71-inclinação 58-encanto 53-ciúme 8-respeito 7-esperança 7-estimação 6-paixão 6-admiração 2-compaixão 1-dor 1-raiva 1-ódio 1-alegria 1-desilusão 1-pena

ódio:
100-ódio 73-raiva 65-alegria 63-tristeza 63-entusiasmo 62-desespero 59-dor 52-aflição 50-susto 46-pena 44-desilusão 37-compaixão 35-luto 35-paixão 34-apreensão 33-desesperança 32-estimação 31-asco 31-decepção 29-admiração 26-respeito 26-esperança 15-enfado 13-fúria 12-pânico 7-alvoroço 7-amor 5-cólera 5-excitação 4-agitação 2-ciúme 2-comoção 2-indignação 1-furor 1-medo 1-ira 1-vergonha 1-encanto 1-inveja 1-inclinação

paixão:
100-paixão 91-compaixão 87-respeito 83-esperança 83-estimação 63-admiração 48-raiva 41-alegria 37-desesperança 35-dor 35-ódio 20-ciúme 20-pena 19-desilusão 14-tristeza 13-desespero 13-encanto 13-entusiasmo 12-luto 12-aflição 10-susto 9-asco 8-decepção 8-inclinação 6-medo 6-vergonha 6-amor 3-furor 3-inveja 3-apreensão 1-agitação 1-cólera 1-excitação 1-comoção 1-fúria

pânico:
100-pânico 60-alvoroço 50-fúria 47-agitação 45-cólera 43-excitação 37-comoção 28-indignação 27-desespero 27-ira 26-aflição 23-susto 20-enfado 17-entusiasmo 17-apreensão 13-tristeza 13-luto 12-ódio 11-dor 11-decepção 9-desesperança 9-desilusão 9-pena 5-asco 4-raiva 3-alegria 2-admiração

pena:
100-pena 46-ódio 44-desespero 43-aflição 41-raiva 40-susto 37-alegria 36-entusiasmo 35-desesperança 30-luto 29-tristeza 27-dor 24-asco 24-apreensão 20-paixão 20-compaixão 20-estimação 17-respeito 17-esperança 15-desilusão 13-amor 11-admiração 9-furor 9-pânico 9-inveja 9-decepção 8-enfado 6-fúria 4-alvoroço 3-cólera 3-excitação 3-indignação 2-agitação 2-ciúme 2-encanto 1-medo 1-ira 1-vergonha 1-comoção 1-inclinação

raiva:
100-raiva 86-alegria 73-ódio 72-dor 56-tristeza 52-compaixão 48-paixão 47-desespero 47-entusiasmo 41-desilusão 41-estimação 41-pena 40-respeito 40-esperança 40-aflição 38-susto 36-admiração 35-asco 32-desesperança 30-luto 28-apreensão 26-decepção 10-amor 10-enfado 6-fúria 4-pânico 3-ciúme 3-inveja 2-furor 2-encanto 1-alvoroço 1-agitação 1-medo 1-cólera 1-vergonha 1-excitação 1-comoção 1-indignação 1-inclinação

respeito:
100-respeito 87-paixão 82-esperança 80-compaixão 75-estimação 57-admiração 40-raiva 34-alegria 32-desesperança 29-dor 26-ódio 21-ciúme 17-pena 15-encanto 13-

desilusão 10-tristeza 10-desespero 10-luto 10-inclinação 9-susto 9-aflição 9-entusiasmo 8-medo 8-vergonha 8-asco 7-decepção 5-amor 3-apreensão 2-furor 2-inveja

susto:
100-susto 60-desespero 55-aflição 52-entusiasmo 50-ódio 46-tristeza 40-alegria 40-pena 38-raiva 36-luto 35-dor 35-apreensão 29-desesperança 27-desilusão 24-asco 24-decepção 23-pânico 23-fúria 21-enfado 17-alvoroço 14-cólera 11-excitação 11-compaixão 11-estimação 10-agitação 10-esperança 10-paixão 9-respeito 9-indignação 9-admiração 7-comoção 5-amor 4-ira 2-furor 2-inveja

tristeza:
100-tristeza 69-desilusão 65-entusiasmo 63-ódio 59-decepção 56-raiva 55-dor 55-desespero 50-alegria 47-apreensão 46-susto 42-asco 42-aflição 33-enfado 29-pena 27-admiração 21-luto 20-fúria 19-compaixão 16-desesperança 14-paixão 14-estimação 13-pânico 13-indignação 10-respeito 10-esperança 7-alvoroço 6-amor 2-agitação 2-cólera 2-excitação 1-comoção 1-inveja

vergonha:
100-vergonha 91-medo 73-inclinação 60-encanto 53-ciúme 8-respeito 7-esperança 7-estimação 6-paixão 6-admiração 2-compaixão 1-dor 1-raiva 1-ódio 1-alegria 1-desilusão 1-pena

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig verfasst habe und dabei keine andere als die angegebene Literatur verwendet habe. Alle Zitate und sinngemäßen Entlehnungen sind als solche unter genauer Angabe der Quelle gekennzeichnet.

Karlsruhe, 10.1.2006