Westfälische Wilhelms-Universität Münster
Romanische Philologie

Entwicklung eines korpusgestützten Zusammenhangsmaßes für das semantische Wortnetz EuroWordNet

Inaugural-Dissertation
zur Erlangung des Doktorgrades der
Philosophischen Fakultät der
Westfälischen Wilhelms-Universität zu
Münster, Westfalen

vorgelegt von
Sonja Hillebrand
aus Hagen
2005

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AAAI	American Association for Artificial Intelligence
ACL	Association for Computational Linguistics
ACM	Association for Computing Machinery
AK	gemeinsamer Ahnenknoten zweier Knoten im Netz
ANLP	Association for Neuro-Linguistic Programming
BN	Bayesian Network
DELOS	Network of Excellence on Digital Libraries
EACL	European Chapter of the Association for Computational Linguistics
EWN	EuroWordNet
FWN	französisches Netz innerhalb von EWN
GN	GermaNet
GWN	deutsches Netz innerhalb von EWN (nicht identisch mit GermaNet)
ILI	Interlingualindex
kgAK	kleinster gemeinsamer Ahnenknoten zweier Knoten im Netz
NAACL	North American Chapter of the Association for Computational Linguistics
NLP	Natural language processing
POS	Part of Speech, Wortart
WN	WordNet

Symbolverzeichnis

$=$	Gleichheit
$>$	Vergleichsoperator (größer als)
$<$	Vergleichsoperator (kleiner als)
$\geq$	Vergleichsoperator (größer als oder gleich zu)
$\in$	Elementbeziehung (ist Element von)
$\subseteq$	unechte Teilmenge (ist Teilmenge von oder gleich zu)
$\bigcup$	Vereinigungsmenge Teilmenge (vereinigt mit)
$\sum_{j=1,\ldots,n} x$	Summe (summiere den Ausdruck x über j)
$\max_{k \in M}(x)$	Maximierung (maximiere den Ausdruck x über alle möglichen k aus der Menge M)
$\mathbb{R}$	Menge der reellen Zahlen
$\mathbb{R}_0^+$	Menge der positiven reellen Zahlen inklusive 0
$\mathbb{N}$	Menge der natürlichen Zahlen
(a, b)	Mengendefinition (Zusammenfassung von a und b zu einer Menge)
$[a \ldots b]$	Mengendefinition (Intervall von a bis b inklusive a und b)
$[a \ldots b[$	Mengendefinition (Intervall von a bis b inklusive a und exklusive b)
:	Bedingung (für die gilt)
$\Rightarrow$	Folgerung (folgt)
$\Leftrightarrow$	Äquivalenz (ist äquivalent zu)
$*$	Multiplikation
φ	Verbindungspfad eines Netzwerkes
$\lvert\varphi\rvert$	Länge des Pfades φ (Anzahl der Knoten)
ν	Kante eines Pfades

Kapitel 1

Einleitung

1.1 Problemstellung

Ein wichtiges Problem in vielen Bereichen der Computerlinguistik ist die eindeutige Zuweisung von Lesarten zu einem mehrdeutigen Element. Die eindeutige Bestimmung der Lesart einer mehrdeutigen Lexie (Lesartendisambiguierung) ist für viele elektronische Verarbeitungen von Texten, z.B. Übersetzung, Sprachsynthese, Inhaltsangabe, *Information retrieval* u.a. nötig. Wie kann für den Computer modelliert werden, was passiert, wenn ein menschlicher Leser eine mehrdeutige Lexie sieht und ohne größere Schwierigkeiten aus dem Kontext heraus eine Lesart bevorzugt? Wie vergleicht er die Informationen aus dem Kontext mit den Informationen aus seinem Weltwissen?

Wird ein menschlicher Sprecher gefragt, zwischen welchen Objekten ein stärkerer Zusammenhang besteht, zwischen einem Motorrad und einem Fahrrad oder einem Fahrrad und einer Tasse, wird er sicherlich die Beziehung zwischen Fahrrad und Motorrad als stärker empfinden. Auf welche Information bezieht er sich und wie findet in seinem mentalen Lexikon der Vergleich dieser Konzepte statt?

Der menschliche Sprecher legt für diesen Vergleich ein Maß zugrunde, das sich als Maß des Zusammenhangs bezeichnen läßt. Er bevorzugt die Lesart, bei der die Kontextinformationen mit den Informationen aus seinem Weltwissen nach diesem Maß am besten übereinstimmen. Die Problematik der Modellierung von *relatedness* läßt sich im Bereich der Philosophie bis zu Aristoteles verfolgen, für die Verarbeitung in computerlinguistischen Anwendungen ist die Diskussion über 50 Jahre alt (siehe Quillian 1968 und Collins und Loftus 1975).

Auf die Lesartendisambiguierung durch den Computer übertragen bedeutet es, daß der Computer die für ihn erreichbaren Informationen aus dem Kontext des mehrdeutigen Elementes mit dem ihm zur Verfügung stehenden Wissen vergleichen muß. Mit Hilfe des Zusammenhangsmaßes werden dann aus allen möglichen Lesarten diejenigen herausgesucht, die einen sehr hohen Zusammenhangswert zu den Kontextinformationen besitzen. So kann der Computer die wahrscheinlichste Lesart für diesen Kontext zuweisen.

Als ein Modell des assoziativen menschlichen Gedächtnisses wurde das Netzwerk *WordNet* in Princeton entwickelt. Für dieses semantische Netz sind verschiedene Algorithmen entworfen worden, die durch die Verbindungen der Netzelemente untereinander berechnen können, wie „ähnlich“ sich zwei Elemente sind. Durch die Konstruktion der semantischen Netze (besonders das recht dünn gewebte französische Netz) geben sie allerdings außer Hyperonym- und Hyponymstrukturen nur wenige weitere Verbindungen wieder, also nur einen recht eng gefaßten Begriff von „Ähnlichkeit“ gegenüber dem weiter gefaßten Begriff des Zusammenhangs.

1.2 Zielsetzung der Arbeit

In dieser Arbeit wird zuerst ein Überblick über die Möglichkeiten gegeben, diese „Ähnlichkeit“ mit Hilfe eines semantischen Netzwerkes zu modellieren. Die Modelle arbeiten mit Funktionen, die auf der Grundlage der Netzstruktur die Ähnlichkeit von Konzepten (den Knoten im Netzwerk) als numerischen Wert dem Computer zur Verfügung stellen. Mit dem Ziel, das Zusammenhangsmaß im Rahmen der Lesartendisambiguierung zu verwenden, wird eine Korpusauswertung vorgenommen, um mit diesen Informationen das semantische Netzwerk *FrenchWordNet* (das französische Teilnetz eines europäischen Projekts, erstellt auf der Grundlage von *WordNet* 1.5) anzureichern.

Durch das Einfügen von neuen Kanten, die häufig auftauchende Kollokationen mit dem Ausgangselement (*pivot*) verbinden, wird das semantische Netzwerk *FrenchWordNet*, in dem diese Knoten sich nicht in der unmittelbaren Nähe des *pivot* befinden, verdichtet. Für dieses veränderte Netz wird ein Zusammenhangsmaß erstellt, das die Vorteile der kodierten semantischen Ähnlichkeit im semantischen Netz mit der Information über die Kollokationshäufigkeit verbindet.

Bei der Erweiterung des französischen Teilnetzes durch die korpusbasierten Daten soll auch getestet werden, inwieweit der manuelle Aufwand sich reduzieren läßt, um den Vorgang zu automatisieren. In einem letzten Abschnitt wird dieses neu erstellte Zusammenhangsmaß mit den Ergebnissen des Projektes *ROMANSEVAL*[1] verglichen werden.

1.3 Aufbau der Arbeit

Nach der Klärung der verwendeten Bezeichnungen, der Formalisierung des Prozesses der Lesartendisambiguierung und der wichtigen Rolle des Zusammenhangsmaßes wird das verwendete *FrenchWordNet* (im folgenden *FWN*) dargestellt und mit französischen Lexika verglichen. Darauf werden die Zusammenhangsmaße, die für das semantische Netz *WordNet* (im folgenden *WN*) entwickelt wurden vorgestellt und auf ihre Vor- und Nachteile hin untersucht. Nach einer genaueren Betrachtung des verwendeten Korpus *Frantext* wird durch eine Korpusuntersuchung die Erweiterung des Netzwerkes durch Korpusinformationen (statistische Angaben über Kollokationen) motiviert.

Nach dieser kleineren Korpusuntersuchung wird im achten Kapitel eine genaue Auswertung für alle vier in *FWN* vorhandenen Lesarten des Substantives *maison* vorgenommen, um Knoten, die für die neue Verknüpfung in Frage kommen, für die Erweiterung auszuwählen. Mit diesen Informationen zum kontextuellen Zusammenhang wird eine konkrete Einbindung in die Datenbankstruktur der französischen Seite von *EuroWordNet* (*EWN*) durchgeführt, die die netzbasierte Komponente für die Berechnung von *semantischer Ähnlichkeit* bereitstellt. Dazu wird die erstmalig durchgeführte Transformation des Datenbankformats von *EWN* in das *WN*-Format und die notwendige Korrektur der 1998 erstellten, fehlerhaften Datenbank *FWN* dargestellt. Anschließend werden die Datenbankerweiterung und das neu entwickelte Zusammenhangsmaß beschrieben und ausgewertet.

Im Anhang finden sich die Korrekturen, die für die Datenbank vorgenommen wurden, die Quellcodes der verwendeten Programme, Tabellen, die Kommentare zu der vorgenommenen manuellen semantischen Annotation und die Korpusnachweise.

[1] Ein Projekt aus dem Jahr 1998 `http://aune.lpl.univ-aix.fr:16080/projects/romanseval/` (13.01.2005), das verschiedene Disambiguierungsmodelle für romanische Sprachen verglich. Es wurde für die französische Sprache bisher nicht wiederholt. Vgl. auch Kilgariff und Palmer 2000 und Véronis und Ide 1998.

Kapitel 2

Verwendete Terminologie

Wegen der in der Literatur uneinheitlich verwendeten Terminologie soll hier kurz zusammengestellt werden, wie die Termini in dieser Arbeit verwendet werden.

2.1 Graphie – Lexie – Wort

Durch den Terminus *Lexie* wird ausschließlich die Zusammenfassung aller möglichen Lesarten einer *Graphie* bezeichnet. Sie umfaßt damit die einzelnen lexikalischen Einheiten, die Abgrenzung von *Homonymie* und *Polysemie* wird nicht vorgenommen (siehe auch Abschnitt 2.4).

Der Terminus *Graphie* bezeichnet die als Folge von Graphemen ausgedrückte computerlesbare Repräsentation einer Lexie. Da der Rechner eventuelle graphische Varianten nicht berücksichtigen kann (außer wenn sie in der Datenbank explizit verknüpft sind), ordnet er unterschiedliche Graphien auch unterschiedlichen *Lexien* zu.

In manchen englischen Veröffentlichungen wird mit dem allgemeineren Terminus *word* gearbeitet, der aber in diesem Zusammenhang durch den jeweils konkret passenderen Terminus (*Lexie* oder *Konzept*) ersetzt wird.

2.2 Ähnlichkeit – Zusammenhang – Abstand

In den Publikationen werden von den Autoren (oft auch innerhalb derselben Veröffentlichung) verschiedene Termini verwendet, um das betrachtete Phänomen zu benennen: lexikalische und semantische *Ähnlichkeit*, *Zusammenhang*, *Abstand*,

Übereinstimmung oder in der englischsprachigen Literatur *relatedness*, *similarity*, bzw. *distance*.

Auf die unterschiedlichen Definitionen von *Similarität* aus der Philosphie oder der Psychologie kann hier nicht eingegangen werden, übernommen wird hier für die Verknüpfungspfade innerhalb des semantischen Netzes der Terminus *Assoziationskette*.

Manchmal wird Similarität (in den englischen Publikationen *similarity*) verwendet, um die äußere Ähnlichkeit oder die semantische Verbindung durch Hyperonym- oder Hyponymverknüpfungen zu betonen, während *relatedness* auch andere Verbindungen (Meronymie, Antonymie etc.) miteinbezieht (vgl. z.B. das Maß von Hirst und St.Onge, Abschnitt 5.5). Aber diese Unterscheidung wird nicht durchgehend einheitlich gemacht; das Programm, das die hier untersuchten Programme zusammenfaßt, wird vereinheitlichend `similarity.pl` genannt.

Der Terminus *similarity* oder *Ähnlichkeit* ist aber zumindest bei sehr langen Pfadverläufen (also sehr langen Assoziationsketten) nicht mehr zu begründen. Daher wird in dieser Arbeit von der *semantischen Verbundenheit* oder vom *semantischen Zusammenhang* gesprochen. Diese beiden Termini beziehen das Bild der Verknüpfung im semantischen Netz mit ein und implizieren keine Betonung des einen oder anderen Endes des graduierbaren Zusammenhangs.

2.3 Das *Synset*

In der vorliegenden Arbeit wird der Terminus *Synset* verwendet, um die Gesamtheit der lexikalischen Einheiten, die in einem Konzeptknoten des semantischen Netzes zusammengefaßt sind, zu bezeichnen (z.B. *automobile:1*, *auto:1* und *voiture:1*). Der hier verwendete Terminus *Synonymie* ist allerdings nicht sehr streng gefaßt.

Die Konzeptknoten fassen lexikalische Einheiten zusammen, die verschiedenen Definitionen der *Synonymie* entsprechen (der *complete synonymy*, der *total synonymy*, der partiellen Synonymie oder der Synonymie mit Rücksicht auf die Kontextabhängigkeit).[1] Diese für die Verwendung in einer computerlinguistischen Datenbank zu feinen Unterscheidungen sind für den Terminus *Synset* nicht anzuwenden.

[1] Vgl. besonders die Diskussion bei Geckeler 1971, 234ff. und Lyons 1968, 448, Gauger 1961, 173, Müller 1965, 91ff. oder Ullmann 1967, Reprint der 2. Ausgabe von 1957, 108f..

Miller diskutiert für die Anwendung in *WordNet* Synonymie und legt dar (aus seiner psycholinguistischen Sicht): „synonymy is best thought of as one end of a continuum along which similarity of meaning can be graded“ (Miller et al. 1993b, 7). Damit wird eine Graduierbarkeit der konzeptuellen Ähnlichkeit und auch die nicht punktuelle Definition der Synonymie für das semantische Netz zugrunde gelegt, wobei diese Definition aus sich heraus einer Annäherung entspricht.

Die von *WordNet* vorgenommene Zuordnung zu *Synsets* ist in vielen Fällen diskussionswürdig (besonders hinsichtlich des französischen Teilnetzes von *EWN*), aber für den Rechner sind die einmal definierten Vorgaben der Ausgangspunkt seiner Berechnungen. Er verlangt eine Datenbank, in der die möglichen Lesarten einer als Folge von computerlesbaren Zeichen vorliegenden Lexie enthalten sind, kein wissenschaftlich einwandfreies Modell. Für die verschiedenen Anwendungsbereiche wird auch eine unterschiedliche Feinheit der Unterscheidungen ausreichend sein.

Das semantische Netz *WordNet* beruht auf der Zuordnung der einzelnen lexikalischen Einheiten einer Sprache zu diesen *Synsets*. Dabei werden die durch diese Einheiten repräsentierten Konzepte durch Glossen oder Beispielsätze dargestellt. Die problematische Definition von *Konzept* wird für den Rechner aufbereitet, eine theoretische Definition dieses Terminus ist damit aber nicht gegeben und wird auch nicht angestrebt.

2.4 Polysemie und Homonymie

Die Diskussion in der Fachliteratur zeigt, wie schwierig eine synchrone Abgrenzung von Polysemie und Homonymie ist. Bei der Lesartendisambiguierung wird einem mehrdeutigen Element (ob homonym oder polysem) eine eindeutige Lesart zugewiesen. Daher wird hier darauf hingewiesen, daß für die elektronische Datenverarbeitung eine grundlegende, umfassende und stichfeste Definition von Bedeutung nicht vorausgesetzt wird.

Die exakte Unterscheidung von im *System* festgelegten Lesarten und eventuell in der *Norm* möglichen Nuancen und Varianten ist für den Rechner in der Anwendung irrelevant. Die Lesartendisambiguierung ist bei Homonymen einfacher zu modellieren, Polyseme tauchen in ähnlichen Kontexten auf, so daß dort klare Entscheidungsvorgaben für den Algorithmus definiert werden müssen. Das Modell

ist dann komplexer, der Programmablauf identisch. Die Unterscheidung ist für die elektronische Datenverarbeitung lediglich dann ein Problem, wenn daraus unterschiedliche weitere Programmverläufe folgen.

Die Übersetzung eines mehrdeutigen Elements in eine Sprache, in der diese Lesartenunterschiede unterschiedlich lexikalisiert sind, fordert in der Ausgangssprache die Disambiguierung, damit die der jeweiligen Lesart entsprechende Übersetzung vom Programm ausgewählt werden kann. Dieser Vorgang ist unabhängig vom Status *homonym* oder *polysem*.

Daher wird im folgenden die Unterscheidung der Termini *Polysemie* und *Homonymie* nicht mehr gemacht und ausschließlich der neutralere Terminus *Mehrdeutigkeit* verwendet.

Kapitel 3

Lesartendisambiguierung

Wie in der Einleitung beschrieben, ist in vielen computerlinguistischen Anwendungen die Identifizierung von eindeutigen Elementen beispielsweise die Zuweisung von mehrdeutigen Elementen zu eindeutigen Kategorien ein wichtiger Schritt. Daher wird als Beispiel für die mögliche Anwendung des in dieser Arbeit entwickelten Zusammenhangsmaßes – wie die Maße aus Kapitel 5 – im folgenden Abschnitt die Lesartendisambiguierung formalisiert und die wichtige Funktion des Zusammenhangsmaßes in diesen Vorgang eingeordnet.

3.1 Formalisierung

Das zu erreichende Ziel ist die Zuweisung einer eindeutigen Lesart (auf der Grundlage einer Referenz, z.B. Lexikon, Enzyklopädie oder Wörterbuch als Wissensquelle) zu einer mehrdeutigen Lexie aus einem computerlesbar vorliegenden Textausschnitt.[1] Dieses Problem läßt sich folgendermaßen formalisieren:[2]

Ausgangsmenge

Sei $\mathcal{L} = l_1, l_2, ..., l_n$ $(n \in \mathbb{N})$ eine Menge von zusammenhängenden *Graphien* (Text, Abschnitt, Satz), die aus einem Testkorpus extrahiert wurde. Jedes Element

[1] Wie im vorangegangenen Kapitel dargestellt, ist die Unterscheidung von Polysemen und Homonymen nicht von Belang.

[2] Ansatzweise findet sich diese Formalisierung bei Resnik 1999, 111. Bei Véronis und Ide 1998 und Resnik und Yarowsky 1999 findet sich auch ein (inzwischen veralteter) Vergleich der verschiedenen Methoden der Lesartendisambiguierung, bei dem auf die Projekte *SENSEVAL-1* und *ROMANSEVAL* hingewiesen wird. Neuere Vergleiche finden sich bei den Veröffentlichungen zu den Workshops *SENSEVAL-2* und *SENSEVAL-3*, vgl. `http://www.senseval.org`.

l_i besitzt eine zugehörige Menge von möglichen Lesarten: $B_i = b_{i,1}, ..., b_{i,m_i}$ $(i, m_i \in \mathbb{N})$. Dabei ist die Anzahl der unterschiedlichen Lesarten jeweils abhängig von der i-ten Lexie. Damit ist $\bigcup B_i$ die Menge aller Lesarten, die zu $\mathcal{L}$ gehören. Je nach der verwendeten Wissensquelle enthält diese Menge entweder sehr feine Unterscheidungen oder nur wenige Einträge.

Die Menge der Lesarten B', die ein menschlicher Sprecher dem Textausschnitt zuweisen würde, soll als *ideale* Zuweisung angesehen werden. Sie ist eine Untermenge aller Lesartenkombinationen von $\mathcal{L} : B' \subseteq \bigcup B_i$.

Algorithmus

Die Lesartendisambiguierung führt nun einen Algorithmus durch (d.h. eine Funktion $dis(b_{i,j})$ mit $i, j \in \mathbb{N}$) der – versehen mit den Argumenten $b_{i,j}$ und den Informationen l_i – diejenigen Lesarten $b_{i,j}$ aus $\bigcup B_i$ heraussucht, die mit der größten Wahrscheinlichkeit zur *idealen* Menge B' gehören. Damit wird aus der Menge aller möglichen Zuweisungen diejenige herausgesucht, die der *idealen* Menge möglichst nahe kommt.

Vergleichsfunktion

Dabei ist der Vergleich der möglichen Lesarten $b_{i,j}$ mit den Elementen aus $\mathcal{L}$, die Hinweise auf die Lesart geben, ein zentraler Schritt. Diese Hinweise aus dem extrahierten Textausschnitt können Informationen über die Textdomäne, Elemente aus dem globalen oder lokalen Kontext, aber auch syntaktische Beziehungen sein. Für alle möglichen Lesarten werden ebenfalls aus der Wissensquelle oder einem Trainingskorpus Informationen zur Verfügung gestellt (z.B. Beispielsätze aus dem Wörterbuch, Definitionen aus einem Lexikon). Diese beiden Informationsdatenbanken werden von dem Algorithmus sukzessive miteinander verglichen.

Seien $r_{ij,g} \in \mathcal{R}$ die $g \in \mathbb{N}$ verschiedenen zur Lesart $b_{i,j}$ gehörigen Elemente (aus der Wissensquelle oder dem Trainingskorpus entnommen), außerdem bezeichnen $t_{i,k} \in \mathcal{T}$ die $k \in \mathbb{N}$ verschiedenen aus dem Testkorpus extrahierten und mit l_i verknüpften Elemente. Dabei ist g abhängig von der Kombination von i und j. Stimmen nun einzelne Elemente $t_{i,k}$ aus dem Testkorpus mit den Informationen $r_{ij,g}$ überein oder sind sie ähnlich, kann der Lesart $b_{i,j}$ ein Wert für die Wahrscheinlichkeit, die ideale Lesart des Elementes l_i in diesem Kontext zu sein, zugewiesen werden.

Stellt sich bei dem Vergleich heraus, daß einige Elemente identisch sind, kann ein hoher Wert zugewiesen werden. Sind aber keine – oder nicht ausreichend viele – exakte Übereinstimmungen vorhanden, muß es möglich sein, zwischen den Verknüpfungen $t_{i,k}$ der ambigen Lexie aus dem Testkorpus und den Elementen $r_{ij,g}$ aus dem Trainingskorpus Vergleiche anzustellen.

Zusammenhangsfunktion

Für diesen Vergleich wird eine weitere Funktion *Zus* benötigt, die für die Lesarten $b_{i,j}$ – versehen mit den Informationen $t_{i,k}$ und $r_{ij,g}$ – einen Wert der Ähnlichkeit bzw. des semantischen Zusammenhangs errechnet.[3] Der Wertebereich dieser Funktion liegt im Intervall $[0, 1]$, wobei 1 für die vollständige Übereinstimmung steht und 0 für die maximale Entfernung. Somit kann die Funktion *Zus* als Maß der semantischen Ähnlichkeit oder – wie es im folgenden bezeichnet wird – des semantischen Zusammenhangs angesehen werden.

Mit Hilfe dieser Funktion hat der Rechner ein Kriterium (einen Wert im Intervall [0,1]), um zu bestimmen, ob die Lesart $b_{i,j}$ für den Rechner zur bestmöglichen Näherung der idealen Zuweisung gehört. Dieser Wert wird als Argument $s_{i,j}$ in die Funktion $f(s_{i,j}, l_i)$ aufgenommen. Die Funktion f bewertet nun die Ergebnisse für die verschiedenen Lesarten und weist dem Paar $(s_{i,j}, l_i)$, für das die Funktion *Zus* das absolute Maximum erreicht, den Funktionswert 1 zu, während die anderen Paare den Wert 0 erhalten.

Zuweisungsvorschrift

Insgesamt weist somit die Funktion $dis(b_{i,j})$ derjenigen Lesart, die zum wahrscheinlichsten Paar $(b_{i,j}, l_i)$ gehört, den Wert 1 zu. Ausgehend von den vorangehenden Überlegungen ist die Lesartendisambiguierung folgendermaßen zu formalisieren:

$$dis(b_{i,j}) = f(s_{i,j}, l_i) = f(Zus(r_{ij,g}, t_{i,k}), l_i).$$

[3] Unter dem Oberbegriff des semantischen Zusammenhangs werden die semantische Ähnlichkeit (Verbindung durch Hyperonymie und Hyponymie) und die semantische Verbundenheit (Verbindungen auch durch Meronymie, Antonymie etc.) zusammengefaßt. Siehe auch Kapitel 2.

Diese Funktion bildet das Tripel ($r_{ij,g}$, $t_{i,k}$, l_i) auf die Werte 1 (wahrscheinlichste Lesart, diejenige die der Algorithmus zuweist) oder 0 ab (der Zusammenhangswert war kleiner, d.h. die Bed $b_{i,j}$ wird nicht zugewiesen). Der Definitionsbereich $\mathbb{D}$ von $dis(b_{i,j})$ besteht aus den Kontextinformationen des Texts und der Wissensquelle, der Wertebereich $\mathbb{W}$ ist $\{0, 1\}$.

3.2 Maß des Zusammenhangs

Die Zusammenhangsfunktion *Zus* steht im Mittelpunkt der vorliegenden Arbeit. Als wichtigstes Element der Vergleichsfunktion entscheidet sie über die semantische Nähe der Argumente, damit letztendlich über die Zuweisung der einen oder der anderen Lesart. Diese Zusammenhangsfunktion kann unterschiedliche Formen haben: Sie ist stark abhängig von den zugrundeliegenden Mengen $\mathcal{R}$ und $\mathcal{T}$.[4]

Für das semantische Netzwerk *WordNet* wurden Zusammenhangsmaße (vgl. Kapitel 5) entwickelt, die zwei Konzepten, die im Netz existieren, einen Wert zuordnen, der ihren Zusammenhang wiedergeben soll. Damit werden die Mengen $\mathcal{R}$ und $\mathcal{T}$ aus dem Netz gewählt. Als erster entwickelte Sussna 1993[5] ein Maß, bei dem er den unterschiedlichen Verknüpfungen Gewichte zuteilte und die Tiefe der Knoten berücksichtigte. Danach folgten einige andere, die insgesamt stark abhängig von der Struktur des Netzes sind und für den Vergleich keinerlei Informationen aus der wirklichen Verwendung der Elemente mit einbeziehen. Die Kollokationen in einem Korpus sind meistens nicht die Elemente, die sich in einem assoziativen Netz nahe beieinander befinden (vgl. die Auswertungen in Kapitel 7.3).

In dem Maß, das in Kapitel 10 vorgestellt wird, werden für die Mengen $\mathcal{R}$ und $\mathcal{T}$ Informationen aus einer Korpusuntersuchung auf existierende Knoten im Netz abgebildet. Damit werden im neu entwickelten Maß Verknüpfungen berücksichtigt, die nicht netzimmanent, sondern durch eine Korpusuntersuchung motiviert sind.

[4] Bei Budanitsky 1999, 5ff. findet sich eine Zusammenstellung und Beurteilung der verschiedenen Herangehensweisen an dieses Problem: basierend auf Wörterbüchern, Thesauri und *WordNet*. Für die Lesartendisambiguierung ist die in manchen Teilnetzen sehr detaillierte *Synset*-Einteilung von *WordNet* allerdings zu fein, vgl. Véronis und Ide 1998, 13 und 22ff..

[5] Da aber durch das Programm `Similarity.pl` (vgl. Kapitel 9.1) kein Paket bereitgestellt wird, um dieses Maß in einen Vergleich miteinzubeziehen, wird es hier nicht weiter betrachtet.

Kapitel 4

Die semantischen Netze

4.1 Das semantische Netzwerk *WordNet*

Ein Versuch, den gewaltigen Speicher, den das menschliche Gedächtnis darstellt, zu modellieren, ist 1985 in Princeton entwickelt worden: ein semantisches Netzwerk.[1] Statt eines maschinenlesbaren enumerativen Wörterbuchs wurden – aktuellen psycholinguistischen Theorien zum menschlichen Gedächtnis folgend – die Einträge durch zahlreiche Verbindungen manuell zu einem Netzwerk zusammengefügt. Im Jahr 1993 enthielt *WordNet* nach einem ersten Arbeitsabschnitt 51.500 Einträge in der Datenbank, organisiert in 70.100 Konzepten.

Die einzelnen lexikalischen Einheiten sind durch semantische und thematische Relationen miteinander verknüpft, synonyme lexikalische Einheiten sind in so genannten *Synsets* zusammengefaßt. In der Substantivhierarchie[2] sind neben *Hyponymie* und *Hyperonymie* besonders die Verknüpfungen durch *Antonymie* und *Meronymie* ausgearbeitet, die Verbhierarchie wird durch *Entailment* und verschiedene Troponymietypen strukturiert. Adjektive und Adverbien werden in *WordNet* hauptsächlich durch *Antonymie* strukturiert, da sich keine Beziehungen wie Hyponymie oder Hyperonymie herstellen lassen. Das Datenbankformat von *Wordnet* wird im Abschnitt 9.4 dargestellt. Die Version 2.0 enthält 114.648 Sub-

[1] Vgl. für die Grundlagen vor allem Miller 1986a und Miller 1986b, für die weitere Entwicklung Miller et al. 1993a, Miller und Fellbaum 1991, Miller 1995 und schließlich Fellbaum 1998. Eine ausführliche Bibliographie zu *WordNet* ist unter `http://engr.smu.edu/~rada/wnb/` (13.01.2005) einzusehen. Ein Webinterface für die Version 2.0 ist unbeschränkt unter `http://www.cogsci.princeton.edu/cgi-bin/webwn` (13.01.2005) zu finden.

[2] Zum Aufbau der Substantivhierarchie siehe Miller 1993.

stantive (in 79.689 *Synsets* zusammengefaßt), außerdem 11.306 Verben, 21.436 Adjektive und 4.669 Adverbien.

Direkte Anwendungsgebiete dieses Netzwerkes sind synchrone lexikalische Untersuchungen, maschinelle Übersetzung, *cross-lingual information retrieval*[3], Spracherwerb[4], aber auch die Lesartendisambiguierung. Für diese Anwendung bietet sich die Netzstruktur an, da die unterschiedlichen Lesarten einer Lexie verschiedenen *Synsets* – den Knoten im Netz – zugeordnet sind. Je nach ihrer Lesart sind sie mit anderen Knoten verknüpft, also durch ihre Position im Netzwerk und durch die nähere Netzumgebung eindeutig charakterisiert.

4.2 *EuroWordNet*

In der Folgezeit wurden (im Projekt *EuroWordNet*[5]) auf der Grundlage der Idee aus Princeton ab 1996 Netze für verschiedene europäische Sprachen entwickelt, die als multilinguale semantische Wissensquellen benutzt werden sollten. Untereinander sind die einzelnen Sprachnetze durch einen *Interlingua-Index* (kurz *ILI*) verbunden, der übereinstimmende Konzepte (Netzknoten) aufeinander abbildet. Im Rahmen dieses Projektes ist an der *Université d'Avignon et des Pays du Vaucluse* (Frankreich) der französische Teil dieser Datenbank entstanden.

Unter dem Titel *Global WordNet Association* ist dieses Projekt in vielen Sprachen fortgeführt worden (über 30 indoeuropäische Sprachen, aber auch mehrere andere). Die Resonanz auf dieses Projekt ist groß, es eröffnen sich immer neue Anwendungsgebiete.[6]

4.3 Das französische Teilnetz

Abhängig von den vorhandenen Ressourcen wird sprachintern unterschiedlich gearbeitet. Wie bei vielen anderen Teilnetzen des *EuroWordNet*-Projektes wird

[3] Vgl. Voorhees 1993 und Vossen 1997.

[4] Als Beispiel wird auf ein Produkt für Sprachlerner verwiesen: `http://www.thinkmap.com/` (13.01.2005).

[5] Vgl. Vossen et al. 1998a.

[6] Für weitergehende Informationen zu den vorhandenen Netzen vgl. z.B. `http://www.globalwordnet.org/gwa/wordnet_table.htm` (13.01.2005). Ausführliche Darstellungen zu den Projektzielen der *Global WordNet Association* finden sich unter `http://www.globalwordnet.org` (13.01.2005).

in zwei Arbeitsschritten die sprachspezifische Datenbank erarbeitet und an die *EuroWordNet*-Datenbank angehängt.[7]

Die Grundlage für das französische Teilnetz (*FWN*) bildet ein komplexes Wörterbuch, eine semantisch annotierte multilinguale Datenbank (*Dictionnaire Intégral*), die von *MEMODATA* (Avignon) seit 1989 entwickelt wurde.[8]

Es wurden die *Synsets* aus der Datenbank von *WordNet* 1.5 mit den Daten aus dem *Dictionnaire Intégral* maschinell verglichen und bei genügend hoher Ähnlichkeit übersetzt. Damit wurden die Strukturen von *WordNet* 1.5 beibehalten. Durch die manuelle Nachbearbeitung wurde jedes einzelne *Synset* kontrolliert, evt. ergänzt oder gelöscht und die Ausgangsbasis beinhaltete schließlich 19.000 *Substantiv-* und *Verbsynsets*.[9]

Durch manuelles Hinzufügen von Verzweigungen, evt. fehlenden Konzepten und des ganzen Bereichs der Computerterminologie (316 Lexien in 301 *Synsets*) entstanden insgesamt 22.745 *Synsets*, die Substantive, Verben und zehn unverknüpfte Adjektive enthalten. Das Substantivnetz ist im französischen Teilnetz am besten ausgebaut (siehe Tabelle 4.1).[10]

	Substantiv	Verb	insgesamt
Synsets	17.826	4.919	22.745
Anzahl der Lesarten	24.499	8.310	32.809
Lesarten pro *Synset*	1,37	1,69	1,44
sprachinterne Verknüpfungen	14.879	3.898	18.777
d.h. Verknüpfungen pro *Synset*	2,2	2,1	2,18
Verknüpfungen zum *ILI*	17.810	4.915	22.730
d.h. Verknüpfungen pro *Synset*	1	1	1
Synsets ohne Verbindung zum *ILI*	16	4	20

Tabelle 4.1: *Synsets* in *FrenchWordNet*

[7] Vgl. die Beschreibungen bei Vossen et al. 1998a Das deutsche *GermaNet* entstand zuerst unabhängig und wurde dann in *EWN* eingebunden. Daher ist die grundlegende Struktur eine andere. Vgl. (Hamp und Feldweg 1997).

[8] Durch seine Struktur ist das *Dictionnaire Intégral* nicht mit *WordNet* zu vergleichen. Es ist eher eine semantisch annotierte multilinguale Datenbank. Vgl. Wagner et al. 1999 und Catherin und Wagner 1998.

[9] Vgl. Kunze et al. 1998. Bei der hier dokumentierten Arbeit mit dem französischen Teilnetz fiel auf, daß durch die Übersetzung der englischen Ausgangssynsets in *WordNet* 1.5 (wenn auch durch ein französisches Wörterbuch kontrolliert) die Konzepte der französischen Sprache häufig nicht gut wiedergegeben wurden. Auf dieses Problem wird in Abschnitt 4.6 ausführlicher eingegangen.

[10] Vgl. Catherin 1999, 3. Die in der Tabelle falsch angegebene Anzahl der Verknüpfungen zum *ILI* wurde verbessert.

Die semantischen Verbindungen, die sich in *FWN* finden, sind im Anhang in der Tabelle B.1 aufgelistet. Diese Verbindungen beruhen hauptsächlich auf den Verbindungen aus *WordNet* 1.5 (ausgenommen die manuell hinzugefügten Computertermini). Detailliertere Informationen zum Aufbau des französischen Teils, besonders zur Auswahl der *Synsets*, finden sich in den Veröffentlichungen des Projekts.[11]

Die Zahl der Kopfknoten der Hierachie für das französische Substantivnetz wurde den Projektveröffentlichungen zufolge auf 22 verringert[12] (gegenüber 25 in *WordNet* 1.5). Eine ausführliche Darstellung und Begründung der Auswahl der Kopfknoten für *WordNet* 1.5 findet sich bei Vossen et al. 1997. Allerdings hat sich bei der vorliegenden Arbeit an der französischen Datenbank herausgestellt, daß eine große Zahl der dort vorhandenen Kopfknoten offensichtlich durch Verarbeitungsfehler keine Hyperonyme besitzen. Daher mußte dort stark in die vorgegebene Struktur eingegriffen werden (siehe Abschnitt 9.6).

Ein großer (technischer) Unterschied sind die differierenden Datenbankformate der Projekte *EuroWordNet* und *WordNet*. Die unterschiedlichen Datenbankstrukturen werden im Kapitel 9.3 detailliert beschrieben.

4.4 Zugriff auf *EWN*

Statt einer konsolenbasierten Zugriffsmöglichkeit kann mithilfe des Programms *Periscope* (in der Version 1.3.2.)[13] die Netzstruktur von *EWN* betrachtet werden. Dort können innersprachliche Untersuchungen (Verknüpfungen der Einträge untereinander) vorgenommen werden oder Vergleiche mehrerer Sprachen (z.B. Ausdrucksmöglichkeiten eines Konzepts in verschiedenen Sprachen). Auf die Daten kann wegen des binären Formats der zugrundeliegenden Datenbanken nicht direkt zugegriffen werden. Ein *Screenshot* dieses Programms mit den Angaben zum Substantiv <maison> findet sich in Abbildung 4.1.

[11] Hier wird besonders auf Vossen und Escudero 1999, Kunze et al. 1998, Climent et al. 1996, Vossen et al. 1997 und Vossen et al. 1998b hingewiesen, die sich mit der Erstellung und der Struktur der *Synsets* beschäftigen.

[12] Vgl. Catherin 1999, 5.

[13] Dieses Programm ist frei erhältlich und kann unter `http://www.illc.uva.nl/EuroWordNet/sample.html` (13.01.2005) bezogen werden. Mitgeliefert werden Beispieldateien für die einzelnen semantischen Sprachnetze.

4.5 Das Substantiv *maison* in *FrenchWordNet*

Im französischen Teil des *EuroWordNet* wird die Lexie *maison* in vier *Synsets* (Konzeptknoten) geführt (siehe den *Screenshot* des Anwendungswerkzeugs *Periscope* 4.1). Im folgenden werden diese *Synsets* mit ihren direkt im Netz anliegenden Knoten beschrieben und ihre Verknüpfungen mit den über den *ILI* verknüpften englischen Knoten (die Ausgangspunkte der Übersetzung) und deutschen Knoten (als Vergleich) dargestellt.

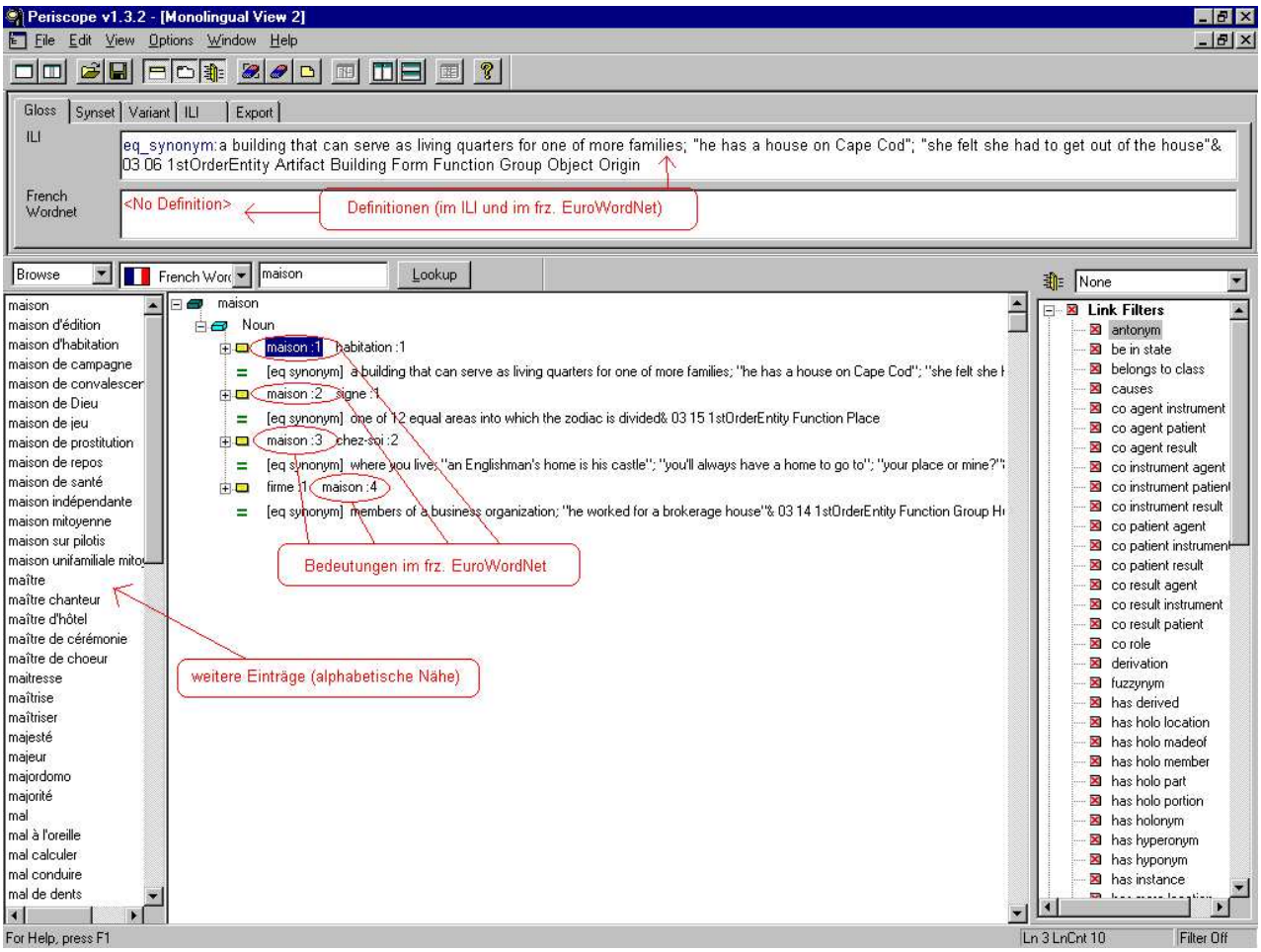


Abbildung 4.1: Das Substantiv *maison* in *Periscope*.

[maison:1, habitation:1]

Die Schreibweise [maison:1, habitation:1] faßt die beiden lexikalischen Einheiten *maison:1* und *habitation:1* zu einem *Synset* zusammen. Dieser Konzeptknoten

wird mit der Glosse <*a building that can serve as living quarters for one of* [sic] *more families*> versehen. Diese Erläuterung auf der Grundlage der Glossen zum *WordNet 1.5* wird ergänzt durch (englische) Beispielsätze:[14] „he has a house on Cape Cod"; „she felt she had to get out of the house", die zum ursprünglichen englischen *Synset* [*house:2*] gehören.

Direkt wird das französische Konzept mit dem Hyperonymknoten [*pension:1, logis:1, logement:1*] verknüpft, mit den Hyponymknoten [*pension:2*], [*bungalow:2, maison de campagne:2*], [*propriété:3, maison de campagne:1*], [*maison indépendante:1*], [*chambre d'ami:1*], [*résidence:3, propriété:2*], [*abri:7*], [*manoir:1, résidence:1*] und den Meronymknoten [*porche:1, véranda:2*], [*étage mansardé:2, grenier:2, mansarde:2*], [*cabinet de travail:1*] (vgl. Abbildung 4.2).

Der deutsche Teil des *EWN* hat mehrere Konzeptknoten über den *ILI* mit diesem französischen Konzept verknüpft: [*Haus:1*], [*Altbau:1*], [*Fachwerkhaus:1*] und [*Neubau:1*]. Diese letzten Konzepte werden im Französischen durch komplexe Lexien wiedergegeben.

[maison:2, signe:1]

Dieses Konzept aus der Astrologie wird beschrieben durch <*one of 12 equal areas into which the zodiac is divided*>. Es ist als Hyponym eingeordnet unter dem Konzeptknoten [*région:5, étendue:4*] und besitzt das Holonym [*zodiaque:1*]. Dieses Konzept ist ebenfalls als Hyponym unter dem Konzept [*région:5, étendue:4*] eingeordnet. Als weitere Holonymverbindung von [*zodiaque:1*] existiert das Konzept [*firmament:1, ciel:2*], dessen Hyperonyme sich durch einige Zwischenebenen in der Hierarchie wieder bis zum Knoten [*région:5, étendue:4*] verfolgen lassen. Die Abbildung 4.3 soll diese Verknüpfungen verdeutlichen.

Die deutsche Entsprechung ist im deutschen Teil des *EWN* nicht enthalten. Dagegen ist diese Lesart in *GermaNet* berücksichtigt, denn sie ist wie im Französischen als Bezeichnung in der Astrologie gebräuchlich.[15]

[14] Zu den englischen Glossen und Beispielsätzen siehe auch den Abschnitt 4.6.
[15] Vgl. Brockhaus 2000.

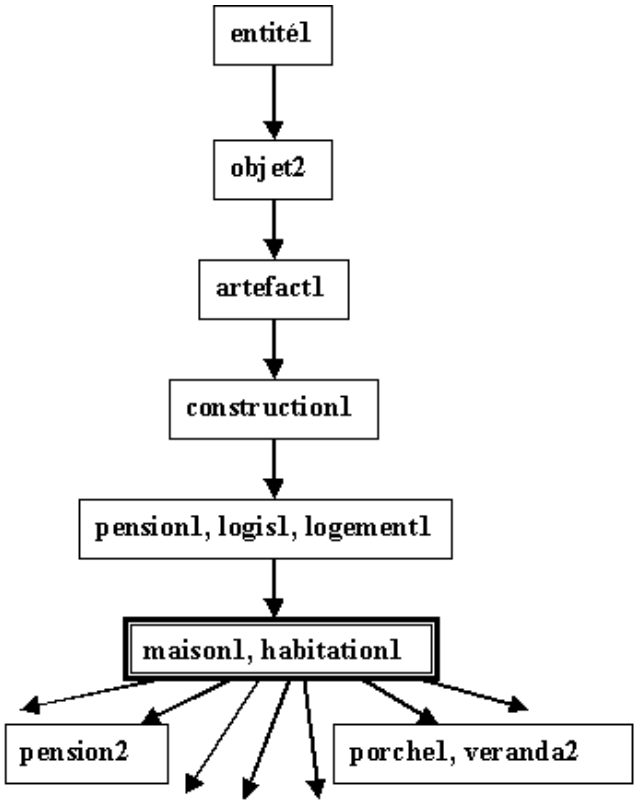


Abbildung 4.2: Teilhierarchie für [*maison:1, habitation:1*].

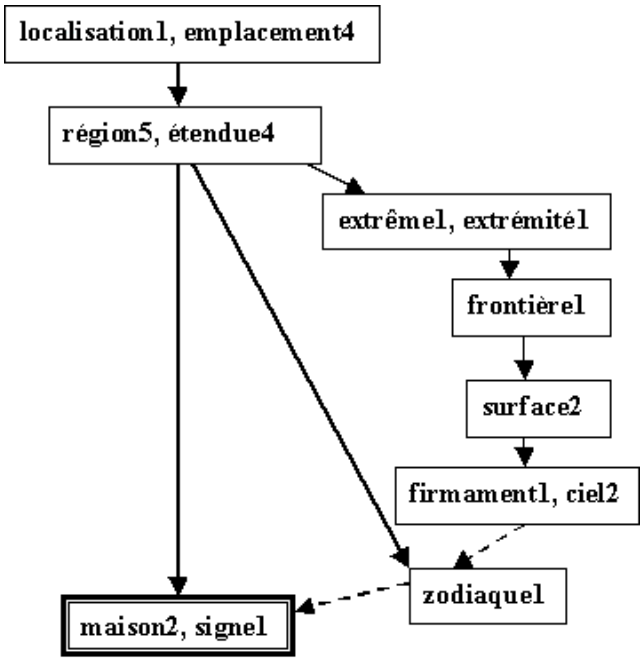


Abbildung 4.3: Hyperonyme und Meronyme für [*maison:2, signe:1*].

[maison:3, chez-soi:2]

Das dritte Konzept wird durch die Glosse <*where you live*> erläutert. Damit wird das französische Konzept nur annähernd beschrieben. Näher liegt die deutsche Entsprechung – vernetzt durch den *ILI*: [*Zuhause:1, Heim:1*] und [*Wohnort:1*]). Die von *WN* gegebenen englischen Glossen können das französische Konzept auch nicht erläutern: „an Englishman's home is his castle"; „you'll always have a home to go to"; „your place or mine?", sie beschreiben eher das englische Konzept [*place:8, home:5*]. Der Hyperonymknoten im französischen Netz ist [*résidence:5, demeure:1*].

In der Abbildung 4.4 ist die Einordnung der beiden Konzepte [*maison:2*] und [*maison:3*] gut zu erkennen. Sie sind beide unter einem gemeinsamen Kopfknoten [*localisation:1, emplacement:4*] eingehängt, aber durch den weiteren Hierarchieverlauf getrennt.

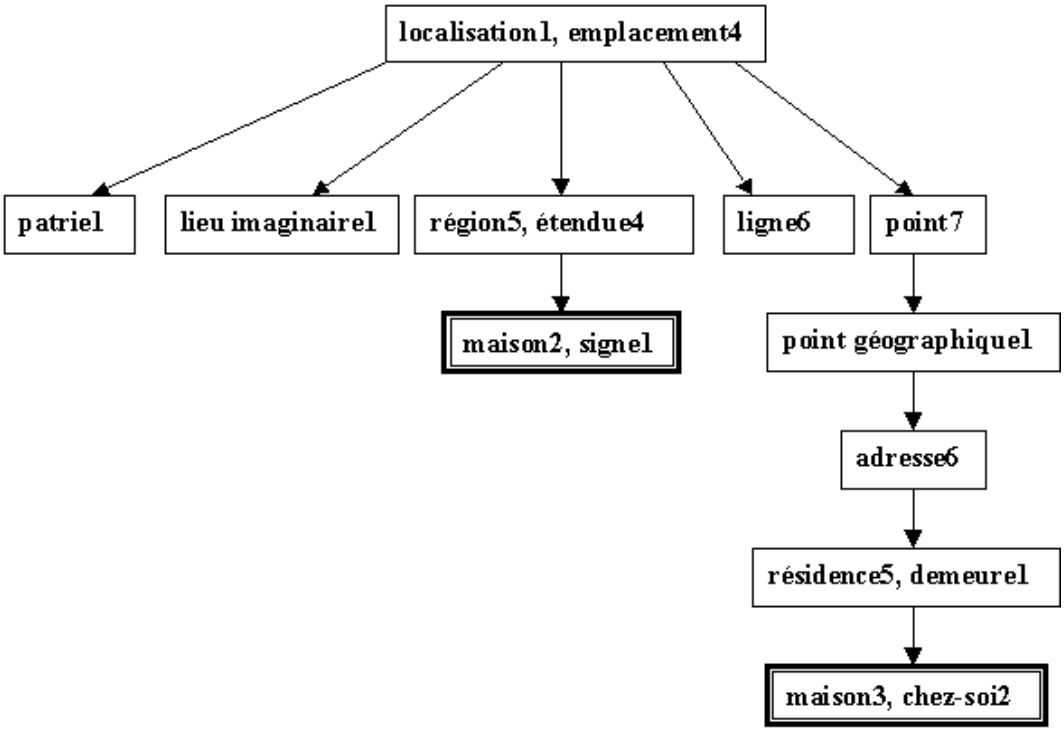


Abbildung 4.4: Teilhierarchie für [*maison:2, signe:1*] und [*maison:3, chez-soi:2*].

[maison:4, firme:1]

Den Ausgangspunkt für die französische Übersetzung bildet das englische *Synset* [*house:5, firm:1, business firm:1*]. Die Glosse ist <*members of a business organization*> und als Beispielsatz gibt *EWN* „he worked for a brokerage house“ an. Dieses Konzept wird über den *ILI* mit dem deutschen Konzeptknoten [*Firma:1*] wiedergegeben. Der Hyperonymknoten ist [*entreprise:1, société*], die zwei Hyponymknoten sind [*maison d'édition:1, société d'édition:1, éditeur:1*] und [*marchand:5, négociant:1*]. Der oberste Kopfknoten dieses Ausschnitts ist [*groupe:2, groupement:3*], daher ist keine Verbindung zu den anderen Konzepten herstellbar (vgl. Abbildung 4.5).

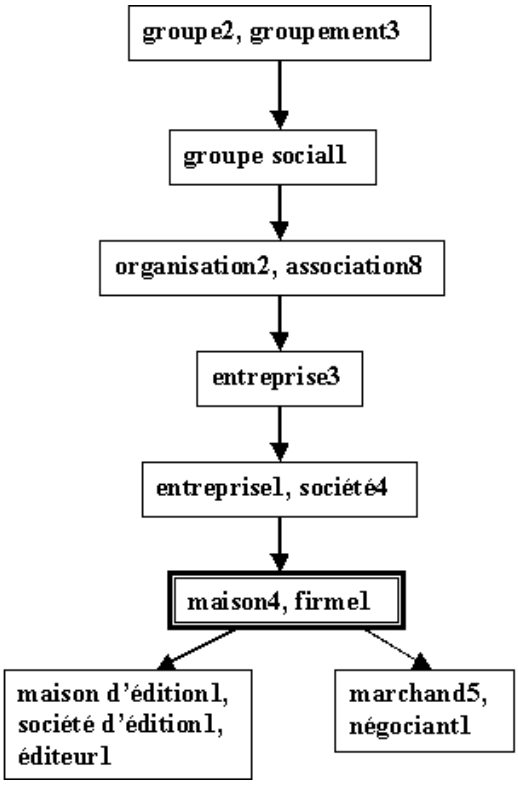


Abbildung 4.5: Visualisierung des Konzeptknotens [*maison:4, firme:1*].

4.6 Anmerkungen

Auf die formalen Fehler wird in Abschnitt 9.6 genauer eingegangen. Inhaltlich ist am französischen Teilnetz zu kritisieren, daß die Strukturen aus *WordNet*

durch die Übersetzung der Konzeptknoten in die französische Sprache auf das französische Netz übertragen wurden. Zwar wurde dieses entstandene Basisnetz danach mit Hilfe einer französischen Datenbank kontrolliert, aber in vielen Fällen ist die Struktur (d.h. die Verteilung der *Lexien* auf die einzelnen *Synsets*) und ein Teil der vorhandenen Knoten nicht an die französische Sprache angepaßt.

Die englischen Glossen sind die Erläuterungen der amerikanischen *Synsets* aus *WN*, die durch die Kontrolle durch die französische Datenbank als diejenigen identifiziert wurden, die dem französischen *Synset* am ähnlichsten sind.[16] Damit kann aber nicht die französische lexikalische Einheit beschrieben werden. Außerdem fehlen Aspekte, die bei einem aufwendigen Vergleich mit verschiedenen semantischen Datenbanken (Wörterbüchern, Enzyklopädien oder korpusbasierten semantischen Clustermodellen) bei den verschiedenen Lesarten aufgefallen wären.[17] Als Einblick in die übersetzungsbedingten Probleme wird für das französische Substantiv *maison* ein kurzer Vergleich mit dem ursprünglichen *WN* 1.5 und den Lemmata französischer Wörterbücher vorgenommen.

4.7 Die englischen Quellen in *WN* für *maison*

Die Grundlage für drei der Lesarten von *maison* bildet die Übersetzung des Substantivs *house* des amerikanischen *WN* in der Version 1.5. Dort werden neun Lesarten des Substantivs *house* unterschieden. In der erweiterten Version *WN* 2.0 unterscheidet man 12 Lesarten, die zum Teil mit einigen aus der älteren Version übereinstimmen.[18]

Drei der Lesarten von *house* werden in der Übersetzung übernommen:

[*house:2*]: <*building that can serve as living quarters for one of* [sic] *more families* „he has a house on Cape Cod„ „she felt she hat to get out of

[16] Es wird nicht erläutert, wie die Ähnlichkeit, d.h. die Distanz zwischen dem französischen und dem amerikanischen *Synset* hergestellt wird.

[17] Ein besseres Beispiel stellt dahingegen das deutsche Teilnetz dar, das unabhängig entwickelt wurde und dann an das *EWN*-Format angeglichen wurde. Daher besitzt die deutsche Ausgangsbasis *GermaNet* deutschsprachige Glossen, da diese inhaltlich selbständig entstanden, vgl. die Literatur zur Entstehung von *GermaNet*: Hamp und Feldweg 1997 und Wagner und Kunze 1999.

[18] Es sollte nebenbei bemerkt werden, daß für das semantische Netz für das britische Englisch (Teil des *EWN*) gar kein Eintrag für das Substantiv *house* existiert, es ist lediglich als Verb aufgeführt.

the house“>, über den *ILI* verknüpft mit [*maison:1, habitation:1*] (in der erweiterten Version Lesart Nr. 1)

[*house:5*, *firm:1*, *business firm:1*]: <*members of a business organization*, „he worked for a brokerage house“>, mit *ILI* zu [*maison:4, firme:1*] (als Lesart Nr. 6)

[*house:9*, *sign of the zodiac:1*, *sign:7*, *planetary house:1*, *mansion:2*]: <*one of 12 equal areas into which the zodiac is divided*> mit *ILI* zu [*maison:2, signe:1*] (als Lesart Nr. 11)

Für das vierte Konzept ([*maison:3*, *chez-soi:2*]) ist das folgende englische *Synset* der Ausgangspunkt:

[*place:8*, *home:5*]: <*where you live*; „an Englishman's home is his castle“; „you'll always have a home to go to“; „your place or mine?“>

Beim Vergleich der weiteren Lesarten fällt auf, daß eine Lesart von *house* nicht übernommen wurde, obwohl diese Lesart (wenn auch im *Grand Robert* als figuratives Konzept aufgeführt) durchaus existiert:

[*house:3*]: <*aristocratic family line* „the House of York“> (bei 2.0 Lesart Nr.7)

4.8 Vergleich mit dem *Grand Robert*

Um einen kurzen Vergleich mit einem einsprachigen französischen Standardwörterbuch durchzuführen, werden die Konzepte des französischen Netzes mit dem *Grand Robert* verglichen.

Der Vergleich der Lesartenunterscheidungen zum Substantiv *maison* innerhalb des *FWN* mit den Lesartendifferenzierungen des *Grand Robert* zeigt die Mängel des französischen Teilnetzes. Auch dort werden vier Lesarten zu *maison* unterschieden, allerdings sind die Übereinstimmungen mit den Lesartenunterscheidungen von *FWN* nur sehr gering (siehe Abschnitt 4.8.2).[19]

[19] Vgl. Eintrag zum Substantiv *maison* in Rey 2003.

4.8.1 Einträge im *Grand Robert*

Die **erste Lesart** wird in fünf Konzepte eingeteilt:

1. Das erste Konzept wird durch die Glosse <*Bâtiment d'habitation,* spécialement *Bâtiment construit pour loger une seule famille, ou* maison individuelle *(opposé à* immeuble, appartement*)*> beschrieben. Durch diverse Verweise auf andere Lemmata (z.B. *habitation*, *bâtiment* oder *construction*) liegt der Schwerpunkt auf dem Gebäude als greifbares Objekt, in dem gewohnt wird.

2. Das zweite Konzept wird dargestellt durch <*Habitation, logement (qu'il s'agisse ou non d'un bâtiment entier).*>, aber auch <*L'intérieur d'un logement, son aménagement.*> und es wird auf andere Konzepte verwiesen (z.B. *demeure*, *domicile* und *intérieure*). Unter dieses Konzept fällt auch der Ausdruck À LA MAISON: chez soi. Damit liegt hier der Schwerpunkt auf dem Wohnungsaspekt.

3. Beim dritten Konzept liegt dagegen der Arbeitsort von Hausangestellten im Vordergrund: <*Spécialement Lieu où travaille un domestique.*> und ein Verweis geht zum Lemma *place*.

4. Das vierte Konzept beschränkt die Lesart auf den religiösen Bereich: eine Glosse fehlt, das Konzept wird durch das Beispiel (<*La maison du Seigneur, de Dieu: le temple de Jérusalem*> und die Verweise auf *église*, *sanctuaire* und *temple* erläutert.

5. Auf die Verwendung in der Domäne der Astrologie verweist das fünfte Konzept (ohne Glosse): *Les douze maisons du ciel: les douze fuseaux par lesquels les astrologues divisent le ciel, pour analyser son état au moment de la naissance de qqn.*

Im **zweiten Lemma** liegt der Schwerpunkt auf der Funktion des Gebäudes. Vier Konzepte umfassen *maison* in der Verwendung von festen Ausdrücken:

1. <*Établissement de détention*> mit den Beispielen *maison de correction*, *maisond'arrêt* oder *maison centrale*,

2. <*Établissement public ou privé à un ou plusieurs bâtiments où l'on reçoit des usagers, qu'on les loge ou non.*> mit einer Auswahl aus den Beispielen *maison de santé*, *maison de retraite* und *maison d'éducation*,

3. <*Spécialement Lieu de plaisir.*> mit z.B. *maison de jeux*, *maison de passe* und

4. <*Entreprise commerciale, industrielle.*>, wobei hier auf die Konzepte <*établissement*> und <*firme*> verwiesen wird, z.B. *maison de commerce*, *maison*

mère oder als Spezialbedeutung <*L'établissement où l'on travaille (maison de commerce, administration, etc.).*>.

Als **figuratives Konzept** werden drei Unterkonzepte aufgeführt: Im übertragenen Sinn wird *maison* auch für die Bewohner desselben verwendet <*Les personnes qui vivent ensemble, habitent la même maison.*>, es kann im (veralteten) Konzept auch den weiteren Personenkreis eines Hauses umfassen: < *Les gens attachés au service d'une maison.*> (Verweis auf das Lemma *domesticité*) oder insgesamt eine Familie durch die Generationen beschreiben: <*Descendance, lignée des familles nobles.*>, z.B. *Maison d'Autriche* oder *Maison de Lorraine.*

Als **invariable Apposition** finden sich im vierten Lemma die Glossen <*Qui a été fait à la maison, sur place (opposé à* de série, industriel*).*> (z.B. *Pâté maison*), als ironische, umgangssprachliche Apposition auch <*Particulièrement réussi, soigné.*> (z.B. *Une engueulade maison* und als letzte <*Particulier à (un groupe, une société).*> (z.B. *Esprit maison*, illustriert durch einen Beispielsatz von Simone de Beauvoir: *„Elle a vite attrapé le genre maison.“*).

4.8.2 *Grand Robert* und *FWN*

Das erste Konzept des ersten Lemma zu <maison> im *Grand Robert* läßt sich durch seine Beschreibung am ehesten mit dem ersten Konzept von *FrenchWordNet* [*maison:1, habitation:1*] vergleichen. Als Verweis auf ein Synonym gibt der *Grand Robert* das im *Synset* von *FWN* enthaltene <habitation> an, der Schwerpunkt liegt auf dem Gebäude als Konstruktion (betont durch die Verweise im *Grand Robert* und die Hyperonyme, die den Aspekt des Bewohnens gegenüber dem greifbaren Objekt zurückstellen).

Für [*maison:2, signe:1*] findet sich im *Grand Robert* eine genaue Entsprechung. Es ist auch das Konzept, das sich am besten von den anderen abgrenzt und sich im *FWN* erst durch entferntere Verknüpfungen (vgl. Abbildung 4.4) mit den anderen in Beziehung setzen läßt.

Das *Synset* [*maison:3, chez-soi:2*] findet sich in Ansätzen im zweiten Konzept im *Grand Robert* wieder: Der Verweis auf die Synonyme <demeure> und <foyer>,[20] beschreibt das in *FWN* durch die Glossen und Beispielsätze nur unzureichend dargestellte Konzept. Allerdings bringen die Verweise auf <logis>

[20]Das französische Substantiv <foyer> wird in *FWN* nicht mit der Lesart <*lieu où habite une famille*> wie bei Rey 2003 geführt.

und <domicile> das Konzept in die Nähe des *Synsets* [*maison:1, habitation:1*], da <logis> in *FWN* nur in einem *Synset* auftaucht: dem unmittelbaren Mutterknoten von [*maison:1, habitation:1*] und <domicile> ebenfalls nur einmal belegt ist: als Tochterknoten von [*pension:1, logis:1, logement:1*], also als Schwesterknoten von [*maison:1, habitation:1*]. Die Unterscheidung dieser beiden Konzepte ist nicht einfach, wie die Überschneidung der in *FWN* und *Grand Robert* unterschiedenen Konzepte zeigt.

Das *Synset* [*maison:4, firme:1*] findet sich im *Grand Robert* im zweiten Lemma als viertes Konzept: die Glosse des *Grand Robert* <*Entreprise commerciale, industrielle*> entspricht in etwa der englischen Glosse <*members of a business organization*>, wobei in der letzteren der Schwerpunkt auf den Angehörigen der Firma liegt und nicht auf der Firma selbst. Die Verweise auf die Synonyme im *Grand Robert* (<*établissement*> und <*firme*>) bestätigen einerseits diese Einordnung und widersprechen ihr andererseits. Das Substantiv <*firme*> findet sich im *FWN* nur im oben genannten *Synset*. Das Substantiv <*établissement*> findet sich zweimal: als [*établissement:2, institut:1*] (<*a building or complex of buildings where an organization for the promotion of some cause is situated*>), ein Unterknoten von [*établissement:1*] (<*a public or private structure (business or governmental or educational) including buildings and equipment for business or residence*>), der als unmittelbarer Tochterknoten von [*construction:1*] eingehängt ist, also als Schwesterknoten des Mutterknotens von [*maison:1, habitation:1*].

Das vierte Lemma des *Grand Robert* findet sich in *FWN* nicht. Das Substantiv <temple> ist in keinem *Synset* vertreten, <église> findet sich (durch schlichte Übersetzung des englischen *Synsets* [*kirk:1*], daher fragwürdig) in zwei *Synsets* wieder: als [*église:1*] mit der Glosse <*a Scottish church*>[21] als direkter Tochterknoten von [*église:2*] mit der Glosse <*for public (especially Christian) worship*>.

Die lexikalische Einheit *la maison de Dieu* ist als Mutterknoten [*maison de Dieu:1*][22] von [*église:2*] eingehängt. Eine Verbindung zu einem der *Synsets* von *maison* ist nur über einen Ahnenknoten ([*construction:1*] als Großmutterknoten von [*maison:1, habitation:1*] und Großmutterknoten von [*église:2*]) weit verzweigt vorhanden.

[21] Die Stellung der schottischen Kirche in Frankreich ist nicht so wichtig wie die Präsenz des *Synsets* im Netz suggeriert.

[22] Es gibt keinen Knoten, in dem <la maison du Seigneur> belegt ist.

Bei der folgenden Untersuchung wird das vierte Lemma (<maison> als Apposition) getrennt betrachtet. Es läßt sich nicht eindeutig einer der vier von *FWN* unterschiedenen Lesarten zuweisen: Es stehen sich Argumente für die vierte Lesart als Produktionsort (für das erste Konzept) oder für die dritte Lesart als zugehörig zu einer Gruppe (bei dem dritten Konzept) gegenüber.

4.9 Ergebnis des lexikographischen Vergleichs

Bei diesem Vergleich (der nur für eine einzelne französische Lexie durchgeführt wurde) ist gut zu erkennen, daß durch die Übersetzung des englischen *WordNet*, d.h. durch die Übertragung der Struktur mit der Ersetzung der englischen *Synsets* durch französische, ein Netz entstanden ist, das den Eigenheiten der französischen Sprache (trotz Verwendung der Datenbank *Dictionnaire Intégral*, die diese maschinelle Übersetzung überwachen sollte) nicht Rechnung tragen kann. Um diese Mängel auszugleichen, müßte aufwendig ein Vergleich zwischen verschiedenen semantischen Datenbanken stattfinden, um ein Netzwerk zu erhalten, in dem die spezifisch französischen Lesarten (in einer ausreichend feinen Lesartendifferenzierung) repräsentiert sind.

Ein Beispiel für die „Eigentümlichkeiten“ die sich während der Arbeit an der Datenbank ergaben (siehe auch Abschnitt 9.6) sind die beiden Einträge für die Konzeptknoten <*barrette:1*> und <*barette:1*, *balai:2*>:

[*barrette:1*]: <two-part cylindrical tumblers held in place by springs; when they are aligned with a key, the bolt can be thrown> (als Übersetzung von engl. <pin:6> in *WN*-1.5).

[*barette:1*, *balai:2*]: <an implement that has hairs or bristles firmly set into a handle> (als Übersetzung von engl. <brush:5> in *WN*-1.5).

In den Wörterbüchern *Petit Robert* und *Le Petit Larousse* findet sich kein Eintrag zu <barette>. Die jeweiligen Glossen in den Wörterbüchern zu <barrette> sind:

1. Toque carree à trois ou quatre cornes, des ecclésiastiques. 2.1. Petite barre portée comme ornement vestimentaire. 2.2. Pince à cheveux, souvent munie d'un système de fermeture. 2.3. Bride décorative. 2.4. Petite portion allongée de haschisch. (Rey und Rey-Debove 1993)

1. Bonnet carre, à trois ou quatre cornes, des ecclésiastiques, noir pour les prêtres, violet pour les évêques, rouge pour les cardinaux. 2.1. Épingle à fermoire pour les cheveux. 2.2. Broche (bijou) longue et étroite. 2.3. Ruban de décoration monté sur un support. (Legrain und Garnier 2000)

Ein Zusammenhang mit dem im französischen Teilnetz eingehängten Konzept ist nicht zu sehen. Bei einer Überprüfung der Einträge hätte das auffallen können.

Die aufwendige Korrektur der offensichtlichen Fehler des französischen Teilnetzes war nicht als Schwerpunkt für die vorliegende Arbeit gedacht, ergab sich aber durch die immer wieder auftretenden Fehler bei der Transformation des Datenbankformats. In Abschnitt 9.6 findet sich eine genaue Beschreibung der vorgenommenen Verbesserungen. Die häufigen Rechtschreibfehler hätten beispielsweise durch einen Abgleich mit einer französischen Datenbank vermieden werden können.

Kapitel 5

Zusammenhangsmaße

In diesem Kapitel werden die verschiedenen Maße vorgestellt, die bei der Berechnung des semantischen Zusammenhangs die Taxonomie *WordNet* zugrunde legen. Dazu gehören einfache Maße, die entweder nur auf statistischen Werten beruhen oder schlicht die Kanten auf dem Verbindungspfad zählen, ein Maß, das mit gewichteten Kanten arbeitet, und komplexere Maße, die Information verwerten, die ein gemeinsamer Ahnenknoten besitzt. Diese Maße werden mit dem in Kapitel 9.1 vorgestellten Programm berechnet.

Zuerst werden einige mathematische Elemente, die in den Zusammenhangsmaßen verwendet werden, zusammengestellt: die *Metrik*, die korpusbasierte und die netzimmanente Wahrscheinlichkeit und der *Informationsgehalt* eines Knotens in einer Netzhierarchie.

5.1 Mathematische Voraussetzungen

5.1.1 Abstandsmaß als Metrik

Die Funktion *Zus* des semantischen Zusammenhangs soll alle Eigenschaften einer Metrik, d.h. einer Abstandsfunktion besitzen:
$d : M \times M \rightarrow \mathbb{R}_0^+$, so daß für $x, y \in M$ gilt:

1. $d(x, y) = 0 \Leftrightarrow x = y$ (Nulleigenschaft). Ein Element hat zu sich selbst den Abstand 0. Auf den Zusammenhang übertragen bedeutet es, daß zwei Elementen desselben *Synsets* der maximal vergebene Wert zugeordnet wird. Bei dem Wertebereich $\mathbb{W} = [0, 1]$ ist dies 1.

2. $d(x, y) = d(y, x)$ (Symmetrieeigenschaft). Der Abstand von Element x zu Element y ist derselbe wie von y zu x.

3. $d(x, y)+d(y, z) \geq d(x, z)$ (Dreiecksungleichung). Der Abstand zwischen zwei Elementen x und z ist immer kürzer oder gleich dem Abstand über einen Umweg y. Diese letzte Eigenschaft garantiert immer die Berechnung des minimalen Wertes für den Abstand, unabhängig von der Zahl der möglichen Verbindungspfade im Netz. Diese Bedingung stellt insbesondere sicher, daß der Abstand immer ≥ 0 ist. Für $x = z$ folgt nämlich:

$$d(x, y) + d(y, x) \geq d(x, x) \overset{(1)}{=} 0 \overset{(2)}{\Leftrightarrow} 2d(x, y) \geq 0 \Leftrightarrow d(x, y) \geq 0.$$

5.1.2 Verwendete Wahrscheinlichkeitsfunktionen

Eine netzbasierte Wahrscheinlichkeitsfunktion $p_N(k)$ wird bei einigen Zusammenhangsmaßen verwendet. Sie bezieht sich auf die Netzstruktur. Jeder Knoten bekommt einen Wert zugewiesen, der berechnet wird durch:

$$p_N(k) = \frac{|Subbaum(k)|}{AZ}.$$

Als Subbaum eines *Synsets* wird die gesamte Unterknotenhierarchie dieses Knotens bezeichnet. Bei dieser Wahrscheinlichkeitsfunktion wird die Anzahl der Elemente aus dem Subbaum dividiert durch die Anzahl der Knoten im gesamten Netz. Damit wird dem Wurzelknoten der Wert 1 zugewiesen. Jeder andere Knoten k_i erhält als Wert den Wahrscheinlichkeitswert, daß bei zufälliger Auswahl eines Knotens aus dem Netz ein Knoten aus dem eigenen Subbaum (k_i mitgerechnet) gewählt wird. Diese Funktion ist monoton fallend: Je tiefer ein Knoten in der Hierarchie liegt, desto kleiner ist sein Wahrscheinlichkeitswert.

Je nach Netzstruktur wirkt dieser Wert verfälschend. Die Subdomänen sind unterschiedlich gut ausgebaut, und ein Knoten mit einem umfangreichen Subbaum wird gegenüber einem hochangesiedelten Knoten mit einem schwach ausgebauten Subbaum stärker gewichtet.

Der imaginäre Kopfknoten im französischen Teil des *EWN*, der alle existierenden Kopfknoten zusammenfaßt, wird für diese Wahrscheinlichkeitsberechnung als Divisor in die Rechnung miteinbezogen.[1]

Dieser Wahrscheinlichkeitswert unterscheidet sich von einer anderen Wahrscheinlichkeitsfunktion, die häufig bei der Lesartendisambiguierung verwendet wird: Die korpusbasierte Wahrscheinlichkeitsfunktion berechnet die Wahrscheinlichkeit, daß die Lexie l_i in der Lesart $b_{i,k}$ in dem zugrunde liegenden Korpus vorkommt:

$$p_K(b_{i,k}) = \frac{Anzahl(b_{k,i})}{\sum_{j=1,\ldots,n} b_{j,i}}.$$

Hier wird die Häufigkeit der Lesart $b_{i,k}$ durch die Häufigkeit der zugehörigen Lexie l_i im Korpus dividiert. Diese Wahrscheinlichkeit ist unabhängig von eventuellen Kollokationen oder semantischen Konstruktionen. Sie stellt lediglich die unterschiedliche Verteilung der Lesarten im Korpus dar.

5.1.3 Informationsgehalt eines Konzeptknotens

Eine wichtige Funktion ist der Informationsgehalt, der einem Konzeptknoten in der Hierarchie zugeordnet wird. Basierend auf der Wahrscheinlichkeit $p_N(k)$ wird jedem Konzept ein Wert zugewiesen, der wiedergeben soll, wie allgemein oder wie speziell dieses Konzept beschrieben wird. Diese Funktion ist monoton steigend in der Hierarchie: Je tiefer der Konzeptknoten in der Hierarchie angesiedelt ist, desto speziellere Information enthalten die lexikalischen Einheiten, die sich in den Knoten befinden. Dem Wurzelknoten wird der Wert 0 zugeordnet:[2]

$$IG(k) = -\log p_N(k).$$

Diese Funktion verhindert bei Zusammenhangsmaßen die Überbewertung von Konzepten, die hoch in der Hierarchie angesiedelt sind und versieht die Blattknoten am Ende der Verästelungen mit einem zusätzlichen Gewicht.[3]

Es folgen jetzt die in der Literatur vorgestellten Maße, die auf der Grundlage von *WordNet* einen Zusammenhangswert von zwei Konzepten berechnen.

[1] Vgl. die Berechnungen von Resnik, Jiang und Conrath und Lin in Kapitel 5.
[2] Vgl. z.B. Jiang und Conrath 1997, 20.
[3] Vgl. zur weiteren Ausführungen Resnik 1999, 103.

5.2 Einfaches Wahrscheinlichkeitsmaß

In diesem einfachen knotenbasierten Maß wird der Zusammenhang zweier Konzepte (k_1 und k_2) berechnet, indem der niedrigste Ahnenknoten (AK) bestimmt wird, der unter sich die Konzepte k_1 und k_2 vereinigt (durch die Maximierung von $[1 - p_{N(k)}]$).[4]

$$Zus_{p(k)}(k_1, k_2) = \max_{k \in AK(k_1,k_2)} [1 - p_N(k)].$$

Der Wert des Zusammenhangs ist die Differenz $1 - p_N(AK)$. Diese liegt im Intervall $[0, 1[$. Der Wert 1 wird nicht erreicht, da die netzbasierte Wahrscheinlichkeit $p_N(k)$ nie den Wert 0 annimmt. Je niedriger der Ahnenknoten in der Hierarchie liegt, desto niedriger ist $p_N(AK)$, desto höher die Differenz, also der Zusammenhang der Konzepte.

Dieses Zusammenhangsmaß ist (wie andere auch) ein grobes Maß, da es vielen Konzeptknotenpaaren denselben Wert zuweist: sämtliche Unterknoten eines Konzeptpaares erhalten denselben Zusammenhangswert, da sie denselben niedrigsten Ahnenknoten haben. Die jeweilige Tiefe der verglichenen Knoten wird nicht berücksichtigt. Auch werden hier nur die Hyperonym- und Hyponymverbindungen miteinbezogen, die Dichte der Taxonomie (Feinheit des Netzes in dieser lexikalischen Domäne) wird vernachlässigt.

5.3 Einfaches knotenzählendes Maß

Dieses Maß zählt die Knoten, die auf dem kürzesten Weg zwischen zwei Konzepten durchlaufen werden. Durch die Inversion dieser Zahl wird ein nichtlineares Zusammenhangsmaß mit Werten aus dem Intervall $]0, 1]$ berechnet:[5]

$$Zus_{Knoten}(k_1, k_2) = \frac{1}{|\varphi(k_1, k_2)|}.$$

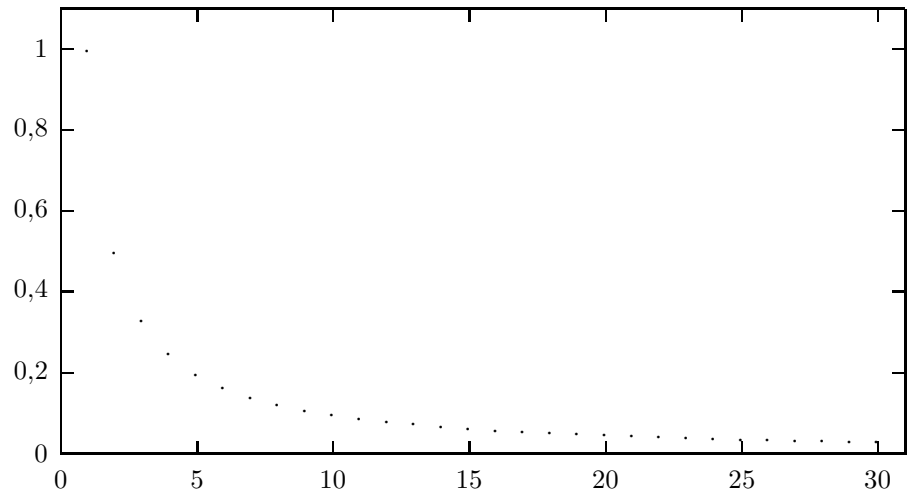
Mit $|\varphi(k_1, k_2)|$ wird die Anzahl der durchlaufenen Knoten auf dem Verbindungspfad zwischen zwei Knoten ausgedrückt. Dieser Wert ist immer $\neq 0$,

[4] Vgl. Resnik 1995b, 3. Die Berechnung dieses Wertes wird mit dem Perl-Package `WordNet::Similarity::einfWahr` realisiert.

[5] Vgl. Resnik 1995b, 3. Die Berechnung erfolgt durch `WordNet::Similarity::edge`.

da mindestens der Ausgangsknoten gezählt wird, falls er identisch mit dem Zielknoten ist. Der höchste Zusammenhangswert 1 wird vergeben, falls k_1 und k_2 identisch sind.

Graph 5.1: Kurvenverlauf



Durch die Verwendung des Inversen einer linearen Funktion verläuft die Kurve (vgl. Graph 5.1) derart, daß sie hohe Werte für nahe beieinander liegende Konzepte vergibt, dann aber schnell geringere Werte annimmt. Damit wird modelliert, daß über lange Pfade (Assoziationsketten) der Zusammenhang nicht linear abnimmt.

5.4 Einfaches kantenzählendes Maß

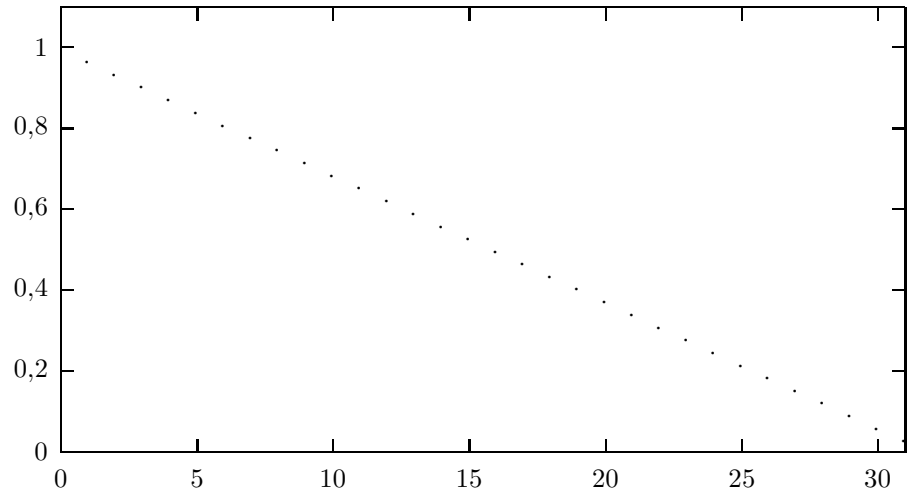
Dieses Maß zählt die Kanten, die auf dem kürzesten Weg von einem Knoten zum anderen durchlaufen werden. Dabei werden alle möglichen Kanten ohne Unterschied (ohne verschiedene Gewichtung) durchlaufen. Durch die Subtraktion der Pfadlänge φ von der maximalen Tiefe der Taxonomie d_{max} entsteht ein Wert aus dem Intervall $[0, 32]$ (die maximale Tiefe in *FrenchWordNet* beträgt 16 Ebenen, damit ist der maximale Abstand 32), der den Zusammenhang der Elemente repräsentiert. Aus der Anzahl der durchlaufenen Knoten ($|\varphi(k_1, k_2)|$) läßt sich durch die Subtraktion von 1 die Anzahl der durchlaufenen Kanten errechnen:[6]

[6] Vgl. Resnik 1995b, 3. Die Berechnung ist durch das neu erstellte Perl-Package `WordNet::Similarity::edge2` möglich.

$$Zus_{Kanten}(k_1, k_2) = 2 * d_{max} - (|\varphi(k_1, k_2)| - 1).$$

Wie bei dem einfachen knotenzählenden Maß spielt hier die Verbindungsart keine Rolle; auch wird die Dichte des Netzes nicht berücksichtigt. Diese beeinflußt aber im Netz die Anzahl der Kanten, die auf dem Weg zwischen zwei Konzeptknoten liegen: in einem sorgfältig und detailliert ausgebauten Teilnetz ist die Dichte höher und damit die Kantenanzahl größer als in einem nur oberflächlich angelegten Teilnetz. Eine Vergleichbarkeit von Werten aus unterschiedlich dichten Teilnetzen ist damit kaum gegeben.

Graph 5.2: Kurvenverlauf



Die Funktion berechnet den kürzesten Pfad von Konzept 1 zu Konzept 2, wobei die Kanten jeweils mit dem Gewicht 1 belegt werden. Der längste Abstand ist der von einem Blattknoten über den Kopfknoten bis zu einem anderen Blattknoten. Dann wird als Wert des Zusammenhangs 0 vergeben. Der höchste Wert 32 wird vergeben, wenn die Ausgangsknoten im selben *Synset* sind. In diesem Fall ist der Kurvenverlauf einfach linear und berücksichtigt daher nicht eventuelle große Sprünge in der Assoziationskette (vgl. Graph 5.2). Außerdem ist das Netz nur an wenigen Stellen 16 Ebenen tief, so daß die vergebenen Werte sich hauptsächlich im Intervall $[0.6, 1]$ befinden, also eine Vergleichbarkeit mit den anderen Maßen nicht möglich ist.

5.5 Semantische Verbundenheit

Semantische Ähnlichkeit liegt vor, wenn die Verbindung innerhalb des Netzwerks nur über Verknüpfungen durch *Hyperonymie* oder *Hyponymie* hergestellt wird. Wenn auch die anderen Verknüpfungen des Netzes miteinbezogen werden, kann eher von *semantischer Verbundenheit* gesprochen werden. Hirst und St.Onge[7] lassen Verbindungspfade über alle möglichen Verknüpfungen innerhalb des Netzwerks zu. Für ihren Disambiguierungsalgorithmus entwerfen sie ein Maß, das zwei lexikalischen Einheiten (in Abhängigkeit von der Art ihrer Verknüpfungen) einen *Zusammenhangswert* zuweist.

Sie beschränken sich vorerst auf die Verwendung der Substantivhierarchie, da die anderen Taxonomien noch nicht über ein sehr differenziertes System von Verknüpfungen verfügen. Einer Verbindung, die nicht durch *Synonymie* oder direkte Verknüpfung charakterisiert ist, außerdem über eine Sequenz von zwei bis fünf Kanten (Festlegung durch Hirst und St.Onge) verläuft und nur eine bestimmte Folge von Richtungswechseln besitzen darf (Anzahl als $d \in \mathbb{N}_0$ in der Gleichung aufgenommen), ordnen sie einen Zusammenhangswert zu (abhängig von zwei Konstanten $C, c \in \mathbb{R}_0^+$):[8]

$$Zushg_{HS}(k_1, k_2) = C - Pfadlänge - c * d.$$

Daraus ist zu erkennen, daß bei langem Weg und häufigem Richtungswechsel das Gewicht der Verbindung sehr niedrig ausfällt. Im verwendeten Perl-Paket werden die beiden Variablen $c = 1$ und $C = 8$ gesetzt. Falls mehr als 5 Kanten durchlaufen werden, wird der Wert des Zusammenhangs 0 gesetzt. Auch in dem Fall, daß kein Pfad existiert (nur über den imaginären Kopfknoten, der alle Teilhierarchien zusammenfaßt), wird der Zusammenhang auf 0 gesetzt.

Eine sinnvolle Überlegung, die diesem Algorithmus zugrunde liegt, ist die Unterscheidung der verschiedenen Pfadverläufe.[9] Diese spiegelt zwar noch nicht die unterschiedlichen Verbindungsarten wieder (Hirst und St.Onge beziehen ohne differenzierte Wertung alle Verbindungstypen mit ein), vermeidet aber zu weite

[7] Vgl. Hirst und St.Onge 1997, 306f..

[8] Vgl. zu den verwendeten Konstanten die detaillierten Ausführungen bei Hirst und St.Onge 1997, 308. Mit Hilfe des Perl-Package `WordNet::Similarity::hso` kann das Maß verwendet werden. Vgl. Abschnitt 9.1.

[9] Vgl. Hirst und St.Onge 1997, 308.

Pfade zwischen den Konzeptknoten und eine zu häufige Pfadrichtungsänderung, erreicht somit also eine konzeptuell nahe Verwandtschaft. Durch die Grenze der maximalen Pfadknotenanzahl wird das Problem der Assoziationskette gelöst und durch das Ausschließen von zu häufigen Richtungswechseln eine zu willkürliche Assoziationsreihe vermieden.

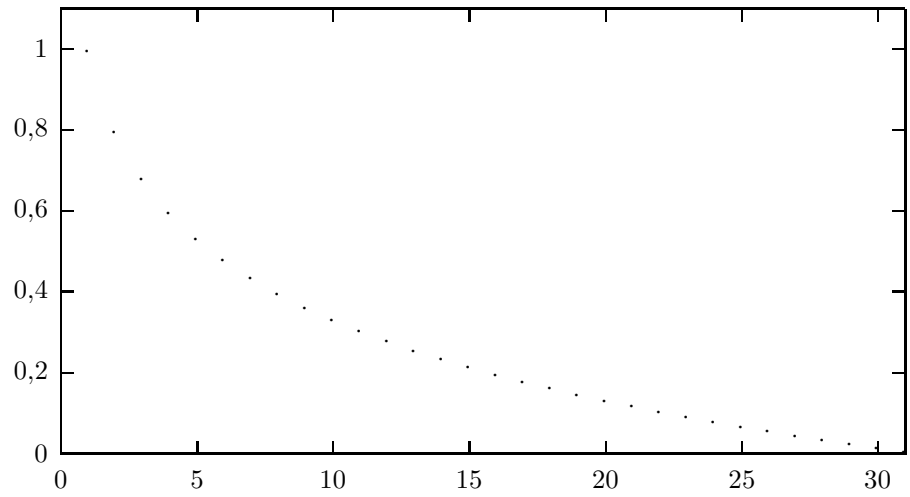
5.6 Pfad und Tiefe der Taxonomie

Leacock und Chodorow entwickeln ein knotenbasiertes Zusammenhangsmaß, das die Anzahl der Knoten auf dem Verbindungspfad mit der maximalen Tiefe der Taxonomie verrechnet:[10]

$$Zus_{LC}(k_1, k_2) = -\log\left(\frac{AZ}{d_{max}}\right).$$

Die Anzahl der Knoten, die auf dem Verbindungspfad von Knoten k_1 nach k_2 liegen (hierbei werden Anfangs- und Endknoten mitgezählt), wird mit AZ und die maximale Tiefe der Taxonomie mit d_{max} (in diesem Fall ist $d_{max} = 16$) bezeichnet. Die verschiedenen Teiltaxonomien des Substantivnetzes in *WordNet* werden durch einen Kopfknoten zu einer einzigen Hierarchie zusammengefügt.

Graph 5.3: Kurvenverlauf bei Leacock und Chodorow



[10] Vgl. Leacock und Chodorow 1998, 275. Mit dem Perl-Package `WordNetSimilarity::lch` kann dieses Maß berechnet werden. Vgl. Abschnitt 9.1.

Auch dieses Maß kann als grob bezeichnet werden, da zwei Konzeptpaaren, die in unterschiedlichen Taxonomieebenen (z.B. durch einen flach verlaufenden Pfad im Gegensatz zu einem sehr steilen Pfad) durch einen gleichlangen Pfad verbunden werden, der gleiche Wert zugeordnet wird (vgl. Graph 5.3). Damit wird die Tatsache, daß Verbindungen in einer hohen Ebene der Hierarchie einen viel allgemeineren Zusammenhangsbegriff darstellen, nicht berücksichtigt.

Durch die Normierung mit der Tiefe der Taxonomie wird versucht, das Problem der unterschiedlichen Dichte im Netz zu neutralisieren. Allerdings ist die Dichte des Netzes in den verschiedenen Teilnetzen unterschiedlich, so daß eine globale Normierung nicht die Lösung sein kann.

5.7 Verhältnis der Knotentiefe

Mit dem Verhältnis der Knotentiefe des kleinsten gemeinsamen Ahnenknotens zu den einzelnen Knoten bestimmen Wu und Palmer ein Maß des Zusammenhangs. Die Tiefe des kleinsten gemeinsamen Ahnenknotens wird dividiert durch die Summe der Tiefe der zu vergleichenden Knoten. Damit entsteht ein Wert, der innerhalb des Intervalls $]0, 1]$ liegt und für Elemente des gleichen *Synsets* den Wert 1 liefert. Der geringste Wert ist der Kehrwert der Tiefe der Taxonomie (immer > 0). Dieses Maß wird durch die Verwendung des *Informationsgehalts* durch Lin noch verfeinert, da dann nicht die Tiefe der Taxonomie, sondern die Anzahl der jeweiligen Unterknoten berücksichtigt wird (siehe Abschnitt 5.10).[11]

$$Zus_{WP}(k_1, k_2) = \frac{2 * t_{AK}}{t_{k1} + t_{k2}}$$

Hier bezeichne t_{AK} die Tiefe des kleinsten gemeinsamen Ahnenknotens und t_{k1} und t_{k2} die Tiefe der zu vergleichenden Knoten. Die Verrechnung der unterschiedlichen Tiefen berücksichtigt den unterschiedlichen Informationsgehalt der Konzeptknoten. Allerdings wird durch die Nichtbeachtung der Pfadlänge die Assoziationskette unberücksichtigt gelassen, damit wird allen Knoten aus derselben Ebene, die einen gemeinsamen niedrigsten Ahnenknoten haben, derselbe Wert zu geordnet.

[11] Vgl. Wu und Palmer 1994, 136. Der Wert kann mit dem Perl-Package `WordNetSimilarity::wup` errechnet werden, vgl. Abschnitt 9.1.

5.8 Der informativste Ahnenknoten

Die Idee, das Maß des Zusammenhangs auf der Grundlage eines Ahnenknotens zu definieren, hat Resnik weiterentwickelt.[12] Er definiert den semantischen Zusammenhang zweier Konzepte in *WordNet* durch den Informationsgehalt des in der Hierarchie am niedrigsten liegenden Konzepts, das diese beiden Konzepte unter sich vereinigt:[13]

$$Zus_R(k_1, k_2) = \max_{k \in AK(k_1,k_2)}[IG(k)] \quad = \max_{k \in AK(l_1,l_2)}[-\log p_N(k)].$$

Aus der Menge der möglichen Ahnenknoten (AK) wird der sogenannte „informativste“ Ahnenknoten (engl. *most informative subsumer*) ausgewählt, dessen Informationsgehalt (IG) (vgl. Abschnitt 5.1.3) maximal ist.

In einem Netzwerk, das (durch Mehrfachvererbung) nicht nur einen einzigen eindeutigen Ahnenknoten zur Verfügung stellt, versieht Resnik die verschiedenen Ahnenknoten mit Gewichten.[14]

Dieses Maß vereinheitlicht vieles, was an Informationen bei den verglichenen Konzeptknoten vorhanden ist. Wie bei den Maßen aus Abschnitt 5.2 und 5.7 wird allen Knoten, die einen gemeinsamen niedrigsten Ahnenknoten haben, derselbe Wert zugeordnet. Durch die Berechnung des Logarithmus hat die Funktion allerdings die wünschenswerte Eigenschaft, daß Wahrscheinlichkeitswerte, die nahe bei 1 liegen, ein größeres Gewicht erhalten und diejenigen, deren Ahnenknoten schon einen niedrigen Wahrscheinlichkeitswert besitzt, zusätzlich geschwächt werden.

5.9 Gewichtung durch Informationsgehalt

Ausgehend vom einfachen knotenbasierten Ansatz entwickeln Jiang und Conrath ein Maß, bei dem zusätzlich ein Gewicht aus dem Informationsgehalt des Knotens erstellt wird: Die Wahrscheinlichkeit, nach einem Mutterknoten (MK) einen der

[12] Vgl. Resnik 1995b, Resnik 1995a, Resnik und Yarowsky 1997, Resnik 1998, Resnik und Yarowsky 1999 und Resnik 1999.

[13] Vgl. Resnik 1995b, 449. Dieses Maß wird als Perl-Package `Wordnet::Similarity::res` zur Verfügung gestellt. Vgl. Abschnitt 9.1.

[14] Zur genaueren Darstellung siehe Resnik 1999, 103.

angefügten Kinderknoten (KI) anzutreffen, wird beschrieben durch die bedingte Wahrscheinlichkeit $p(KI|MK)$:[15]

$$p_N(KI \mid MK) = \frac{p_N(KI \cap MK)}{p_N(MK)} = \frac{p_N(k_i)}{p_N(MK)}.$$

Um daraus eine Gewichtung (GW) zu definieren, wird wieder der Logarithmus verwendet:

$$GW(KI, MK) = -\log(p_N(KI \mid MK) =$$

$$-\log(\frac{p_N(KI)}{p_N(MK)}) = IG(KI) - IG(MK).$$

Damit ergibt sich als Gewichtung durch die Umrechnung des Logarithmus die Differenz des Informationsgehalts von Mutter- und Kindknoten in dem betrachteten Netzwerk. Insgesamt ergibt sich so als Maß für den Abstand zwischen zwei Knoten:[16]

$$Abst_{JC}(l_1, l_2) = IG(k_1) + IG(k_2) - 2 * IG(minimAK(k_1, k_2)).$$

Dies bedeutet, daß der Abstand zweier Konzepte in einem Netzwerk, bei dem jeder Knoten einen numerischen Wert (z.B. Informationsgehalt) zugewiesen bekommen hat, durch die Differenz zwischen der Summe dieses Informationsgehalts und dem doppelten Wert des Informationsgehalts des gemeinsamen Ahnenknotens dargestellt werden kann. Aus dieser Formel für den Abstand zweier Konzepte kann durch Ermittlung des maximalen Abstands in der Taxonomie eine Formel für den Zusammenhang berechnet werden.

Dieser Zusammenhangswert berücksichtigt durch die additive Verrechnung des Informationsgehalts die Tiefe in der Taxonomie. Die Verbindungspfadlänge der Konzepte wird nur indirekt in der Rechnung berücksichtigt. Trotzdem erhält dieses Maß bei dem Vergleich mit menschlichen Annotatoren den besten Korrelationswert (vgl. Abschnitt 5.12).

[15] Vgl. Jiang und Conrath 1997, 26.

[16] Vgl. Jiang und Conrath 1997, 26. Die Berechnung erfolgt mit Hilfe des Perl-Packages `WordNet::Similarity::jcn`. Vgl. Abschnitt 9.1.

5.10 Konvergenz und Divergenz

Ein Zusammenhangsmaß, das sich aus den jeweiligen Gemeinsamkeiten und den trennenden Charakteristika der verglichenen Elemente berechnet, wird von Lin beschrieben.[17] Er ordnet zwei Einträgen in *WordNet* einen Wert zu, der sich folgendermaßen berechnet:[18]

$$Zus(k_1, k_2) = \frac{\log p_N(\text{Gemeinsamkeiten}(k_1, k_2))}{\log p_N(\text{Definition}(k_1, k_2))}.$$

Hierbei bezeichne *Gemeinsamkeiten*(k_1, k_2) die Information, die von k_1 und k_2 geteilt wird (die Konvergenz) und *Definition*(k_1, k_2) die vollständige Information, die benötigt wird, um k_1 und k_2 jeweils von allen anderen Einträgen in der Hierarchie abzugrenzen (enthält also zusätzlich die Divergenz). Die Konvergenz wird im Netz durch den gemeinsamen Ahnenknoten dargestellt, die Divergenz ist in der jeweiligen Definition der Konzepte enthalten. Damit wird ein Wert im Intervall $[0, 1]$ als Zusammenhangsmaß berechnet.[19]

Insgesamt ergibt sich, bezogen auf die hierarchische Struktur von *WordNet*:

$$Zus_L(k_1, k_2) = \frac{2 * \log p_N(MK)}{\log p_N(k_1) + \log p_N(k_2)}.$$[20]

[17] Er benutzt dieses Zusammenhangsmaß für seinen etwas anderen Ansatz der Lesartendisambiguierung: „Two different words are likely to have similar meanings if they occur in identical local contexts." Lin 1997, 3. Mit diesem Ansatz umgeht er das als *bottleneck* bekannte Phänomen, da er aus einem ähnlichen Kontext zweier lexikalischen Einheit schließt, daß die gleiche Lesart vorliegt. Vgl. Lin 1997.

[18] Vgl. Lin 1997, 4. Mit dem Perl-Package `WordNet::Similarity::lin` kann auch dieses Maß berechnet werden. Vgl. Abschnitt 9.1.

[19] Die wichtigsten Elemente seiner theoretischen Herleitung sind:
1. Der Zusammenhang von zwei Konzepten k_1 und k_2 soll sich aus dem Informationsgehalt der geteilten und der nicht geteilten Information ableiten (siehe auch Abschnitt 5.1.3): $Zus(k_1, k_2) = f(IG(Gemeinsam(k_1, k_2)), IG(Definition(k_1, k_2)))$.
2. Für zwei identische Einträge soll die Funktion den Wert 1 vergeben, da der Informationsgehalt dieser Einträge übereinstimmt. Es gilt: IG(Gemeinsamkeiten(k_1, k_2)) = IG(Definition(k_1, k_2)): $Zus(A, B) = \frac{\log p_N(Gemeinsamkeiten(A,B))}{\log p_N(Definition(A,B))} = \frac{\log p_N(Gemeinsamkeiten(k_1,k_2))}{\log p_N(Gemeinsamkeiten(k_1,k_2))} = 1$.
3. Als letzte Forderung soll die Funktion stetig sein. $Zus(k_1, k_2)$ erfüllt diese Bedingung als Komposition stetiger Funktionen. Dies bedeutet, daß die Funktion keinen Sprung in der graphischen Darstellung aufweist, also kleine Änderungen der Ausgangswerte auch nur kleine Änderungen im Funktionswert hervorrufen. Eine ausführlicher Beweis findet sich in Lin 1997, 3f..

Auch hier wird die Verbindungspfadlänge nur ungenau berücksichtigt. Die multiplikative Verrechnung des Informationsgehalts der Konzeptknoten kommt dem logarithmischen Wachstum eher entgegen als die additive Verwendung in 5.9. In der Untersuchung von Budanitsky und Hirst 2001 bekommt dieses Maß aber im Vergleich mit menschlichen Annotatoren gegenüber dem Maß von Jiang und Conrath geringfügig schlechtere Werte, bildet aber mit diesem Maß im Anwendungsfall die zwei besten Zusammenhangsmaße (vgl. Abschnitte 5.12 und 11.2).

5.11 Vergleich der Maße

Als ein Problem bei der Entwicklung eines Maßes für *WordNet* hat sich die unterschiedliche Dichte des Netzes herausgestellt. Einige Subhierarchien (besonders fachspezifische Domänen in der Substantivhierarchie) sind sehr fein aufgegliedert und sind im Zahlenverhältnis anderen weniger ausgebauten Bereichen überlegen. Diese Überlegenheit hat einen starken Einfluß auf die Dichte in dieser Subhierarchie, damit auch einen Einfluß auf die Maße des Zusammenhangs, die hierauf basieren.[21]

Ein einzelner Mutterknoten (bei einigen Ansätzen künstlich hinzugefügt), der in der Substantivhierarchie die getrennten Subhierarchien zusammenfaßt, verhindert die Zuweisung des Wertes 0, falls die zu vergleichenden Konzeptknoten in verschiedenen Subhierarchien auftauchen.

Einzelne Verbindungstypen sind noch nicht vollständig implementiert. In vielen Teilnetzen von *EuroWordNet* ist bislang nur der Typ *Hyponomie - Hyperonymie* verwirklicht. Eine Verbesserung und Erweiterung der Netzstruktur in diese Richtung kann die Ergebnisse der Maße, die mit verschiedenen Verbindungstypen arbeiten, entscheidend beeinflussen. Je nach Verbindung kann ein Gewicht in die Formel miteinbezogen werden.

Die fehlende Verknüpfung der einzelnen Hierarchien (Substantive, Verben, Adjektive, Adverbien) durch spezielle Verbindungen verhindert bislang die Berücksichtigung aller Satzelemente im Korpus. Für die Lesartendisambiguierung

[21] Ein Ausweg aus diesem Problem ist der Begriff der *local density*, der die Dichte in der Taxonomie mit berücksichtigt. Vgl. Agirre und Rigau 1996, Agirre und Rigau 1995 und Richardson et al. 1994.

wäre diese zusätzliche Information wichtig, allerdings müßte dann der Zusammenhangsbegriff weiter gefaßt werden.

Insgesamt stellt sich bei diesem Vergleich heraus, daß die Netzstruktur als Darstellung von Weltwissen für die Lesartendisambiguierung nicht ausreichend ist. Es fehlt die korpusbasierte Komponente gegenüber der psycholinguistisch motivierten – theoretischen – Herangehensweise. Auf dieser Grundlage wird in Kapitel 10 ein Maß vorgestellt, das die Informationen aus dem Verbindungspfad und der Tiefe in der Taxonomie berücksichtigt und auf Informationen aus einer Korpusuntersuchung zurückgreift.

Diese statistische Korpusuntersuchung wird in den folgenden Kapiteln durchgeführt, um die Basis für die Erweiterung des semantischen Netzes zu bilden.

5.12 Die Evaluation von Hirst und Budanitsky

Fünf der vorangegangenen Maße wurden von Budanitsky und Hirst[22] auf der Grundlage von *WordNet* 1.5 mit den Werten menschlicher Annotierer verglichen und durch die Einbindung in ein konkretes *NLP*-Programm getestet.

Bei dem Vergleich mit den Ergebnissen von Sprechern hatten Resnik und Hirst und St.Onge jeweils die schlechtesten Korrelationswerte, Lin und Leacock-Chodorow hatten Werte in [0.816, 0.829] bei beiden Tests, Jiang und Conrath erreichten bei dem Vergleich mit den Werten von Miller und Charles den höchsten Wert 0.850, aber bei Rubenstein und Goodenough nur 0.781. Diese Untersuchung wird in Abschnitt 11.1 noch einmal aufgegriffen.[23]

[22] Vgl. Budanitsky und Hirst 2001. Sie berufen sich auf die Untersuchungen von Rubenstein und Goodenough (Rubenstein und Goodenough 1965) und Miller und Charles (Miller und Charles 1991). Die detaillierte Beschreibung des *NLP*-Programms findet sich ebenfalls in Budanitsky und Hirst 2001. Ein ausführlicher Vergleich und die genauen Werte sind zu finden in Budanitsky 1999, 33f..

[23] Vgl. Budanitsky und Hirst 2001, 2.

Kapitel 6

Das Korpus *Frantext*

6.1 Umfang des Korpus

Das französischsprachige Korpus *Frantext*[1] umfaßt insgesamt 3.679 Texte (mit insgesamt 216.926.960 *Graphien*[2]) aus dem Zeitraum von 1507 (*Le Voyage de Gênes*, ein Werk von Jean Marot) bis 1998 (*Chéri, tu m'écoutes?: alors répète ce que je viens de dire ...* von Nicole de Buron). Im wortartannotierten Teil sind es 1.892 Texte (124.629.091 *Graphien*) aus dem Zeitraum 1830 (mehrere Autoren) bis 1997 (*La Bataille* von Patrick Rambaud). Der Schwerpunkt des Korpus liegt klar auf literarischen Textgattungen: Nur 20% der Texte entstammen verschiedenen wissenschaftlichen Disziplinen, den weitaus größeren Teil konstituieren literarische Werke unterschiedlicher Stilrichtungen (80%).

Im Gesamtkorpus finden sich 1054 *romans* (90.780.696 *Graphien*), 469 Texte der Kategorie *poésie* (10.086.949 *Graphien*), die Gattung *éloquence* ist durch 50 Texte (670.101 *Graphien*) und *pamphlet* durch 20 Texte (1.391.807 *Graphien*) vertreten. Im wortartannotierten Teil finden sich dagegen z.B. nur 735 *romans* (60.637.204 *Graphien*) und 181 Texte der Kategorie *poésie* (4.461.955 *Graphien*). Manche Kategorien sind dort trotz Angabe im Filter nicht oder nur sehr gering im wortartannotierten Teil durch Texte vertreten: *éloquence* durch zwei Texte (85.829 *Graphien*) und *pamphlet* gar nicht.

[1] Zertifizierter Onlinezugriff unter `http://atilf.atilf.fr/frantext.htm` (13.01.2005), Auflage vom März 2001.

[2] Innerhalb von *Frantext* wird die Bezeichnung *mot* verwendet. Diese wird hier durchgehend durch *Graphie* ersetzt.

Aus dem gesamten Korpus kann je nach Untersuchungsziel durch Filter ein Subkorpus als konkretes Untersuchungskorpus zusammengestellt werden. Für die Berücksichtigung der diachronen Unterschiede können für einen synchronen Schnitt Zeiträume ausgewählt werden; ein einzelnes Element, ein einzelner Autor, ein einzelnes Werk oder eine Gattung kann diachron betrachtet werden. Hier wird zwischen den Kategorien *correspondance*, *éloquence*, *mémoire*, *pamphlet*, *poésie*, *récit de voyage*, *roman*, *théâtre*, *traité* und *essai* unterschieden, die im Korpus unterschiedlich gewichtet sind.

6.2 Zugangswerkzeuge

Frantext stellt folgende Untersuchungsmöglichkeiten zur Verfügung:[3]

Einfache Suche

Alle Korpusbelege eines einzelnen Ausdrucks (*expression de séquence*) können aufgelistet werden. Dabei kann es sich um einzelne Wörter, Wortlisten, flektierte Formen oder reguläre Ausdrücke für die Suche nach leicht variierten Formen handeln.

Die Suche nach „le roi est“ liefert 384 Korpusbelege im gesamten *Frantext*-Korpus: z.B. „ [...] le roi est mort [...] “, „ [...] le roi est la malignité de l'homme [...] “, „ [...] le roi est présent [...] “ oder „ [...] le roi est allé [...] “.

Einfache und mehrfache Kookkurrenz

Im Korpus kann nach Kookkurrenzen von bis zu drei Ausdrücken gesucht werden. Dabei kann einzeln gewählt werden, ob die Kookkurrenz miteinbezogen oder ausgeschlossen werden soll. Das Fenster kann ein einzelner Satz oder auch ein größerer Abschnitt sein bis zu einer Distanz von 300 *Graphien*. Außerdem kann noch angegeben werden, in welcher Reihenfolge die einzelnen Ausdrücke auftauchen sollen.

Die Suche nach der Kookkurrenz von <roi>, <reine> und <fils> ergab z.B. 7.487 Textstellen (Textfenster), in denen die drei Elemente gemeinsam auftauchen (darunter 217 Textstellen aus Theaterwerken von Corneille).

[3] Bei der Überschreitung der maximalen Trefferzahl von 50.000 wird die Liste der Belege abgebrochen.

Reguläre Ausdrücke

Mit der Möglichkeit, ein sprachliches Phänomen durch einen regulären Ausdruck zu beschreiben, stellt *Frantext* ein sehr flexibles Werkzeug zur Verfügung. Mit einfachen Vorschriften können Regeln aufgestellt werden, so daß schablonenartig nach Zusammensetzungen eines bestimmten Typs im Korpus gesucht werden kann.

Als Beispiel wird zur *Graphie roi* ein Substantiv gesucht, das als erstes oder zweites Element innerhalb eines Satzes folgt. Die Suche ergab 2.309 Resultate: z.B. „ [...] le sang du roi martyr [...] “, „ [...] le roi d'Angleterre [...] “ oder auch „ [...] précédant le roi une cire à la main [...] “.

Kontext

Die in unmittelbarer Nachbarschaft (*voisinage des occurrences*, umfaßt ein Kontextfenster von bis zu drei Sätzen) auftauchenden Elemente können berechnet werden. Angegeben werden dann die *Graphie* mit ihrer jeweiligen absoluten Häufigkeit. Eine Unterscheidung nach Wortarten kann nicht vorgenommen werden, außerdem wird der Abstand zum untersuchten Wort (*pivot*) nicht angezeigt.

Eine Untersuchung der Satzumgebung von *roi* ergab im wortartannotierten Korpus aus den Jahren 1980 - 2000 insgesamt 596 Belegstellen mit der Ausgabe sämtlicher *Graphien* aus der Satzumgebung (5653 verschiedene auf 28.213 insgesamt). Die häufigste *Graphie* ist „david“[4] mit 102 Okkurrenzen (allein 96 Okkurrenzen stammen aus dem Werk *La Horde d'Or* von Jacques Lanzmann, in dem nach dem „König David“ gesucht wird).

Frequenz

Die absolute Frequenz einer Graphie kann berechnet werden. In dem gesamten *Frantext*-Korpus taucht die *Graphie* „roi“ 72.773-mal und die *Graphie* „reine“ 19.044-mal auf.

6.3 Annotation des Korpus

Das Korpus *Frantext* stellt die Annotation verschiedener Kategorien zur Verfügung (Wortarten, aber auch nicht erkannte Elemente). In der vorliegenden

[4] Groß- und Kleinschreibung wird nicht berücksichtigt.

Untersuchung liegt die Hauptaufmerksamkeit auf den als Substantiv markierten Elementen des Korpus.

Frantext stellt auf der im Internet bereitgestellten Oberfläche die Annotation der Kategorien durch die Zusammenfassung der Segmente in eckigen Klammern und die Angabe der Annotation durch ein Kürzel vor der schließenden Klammer dar (*labelled bracketing*).

```
[Périgord Np] [lui Per] [apporta V] [son D] [aide S], [puisqu' Cs] [il
Per] [était V] [revenu APs] [prendre Inf] [ses D] [quartiers S] [dans
Pp] [la D] [maison S] [rose A], [avec Pp] [son D] [gros A] [valet S]
[et Cc] [sa D] [giberne S] [en Pp] [vermeil S] [qui P] [contenait V]
[un D] [nécessaire S] [de Pp] [toilette S], [du Dg] [gratte S]-[langue
S] [aux Dg] [fards S].
```
[5]

Durch diese Annotation ermöglicht *Frantext* die Filterung nach Suchkriterien, die durch die Kombination von grammatischen Kategorien zusammengestellt wurden (z.B. Suche nach einer Verbform gefolgt von einem Adverb, die Kombination eines Adjektives mit verschiedenen Substantiven).

Das Problem der Homographenunterscheidung wird bei *Frantext* nicht gelöst.[6] Eine Markierung zeigt diejenigen Elemente an, bei denen die Entscheidungsregeln, die dem Annotationsprogramm des *Frantext*-Korpus (*catégorisateur*) zur Verfügung standen, nicht zu einer eindeutigen Kategorienzuweisung führten.[7] Dabei wird zwischen den Nuancen *certain* und *incertain* unterschieden und bei der Annotation der Kleinbuchstabe <c> für *certitude* und <i> für *incertitude* vorangestellt.

Der *catégorisateur* unterscheidet zwischen 26 grammatischen Kategorien und weist nicht bearbeiteten oder nicht erkannten *Graphien* jeweils eine weitere Kategorie zu (vgl. Tabelle B.2 im Anhang auf Seite 209).[8]

[5] Dies ist ein Ausschnitt aus der Datei `maison_1980_2000Kategorie.txt`. Vgl. die Abbildung A.1, den *Screenshot* dieses Auszugs, im Anhang auf Seite 128.

[6] Siehe auch das Beispiel in Abschnitt 6.5.

[7] Es können z.B. Konflikte durch den Widerspruch mehrerer Zuweisungen auftreten. *Frantext* weist dann die wahrscheinlichste Annotation zu (genauere Angaben werden hierzu nicht gemacht). Aber auch die sichere Zuweisung einer grammatischen Kategorie kann sich in einem speziellen Kontext als falsch erweisen.

[8] Die französischen Termini werden hier beibehalten, da sich die Bezeichnungen nicht übertragen lassen.

6.4 Zusammenstellung des Untersuchungskorpus

Mit *Frantext* wird ein Untersuchungskorpus definiert, das kategorisierte Texte aus den Jahren 1980 bis 2000 umfaßt. Es besteht aus 87 Texten (6.404.598 *Graphien*). Dabei überwiegt der Anteil der *romans* (81 Texte, 5.920.533 *Graphien*), 6 Texte werden der Kategorie *essai* zugeordnet (484.065 *Graphien*), davon ein Text, der auch unter der Kategorie *traité* (225.412 *Graphien*) geführt wird. Die bibliographischen Nachweise der Texte, die in diesem Untersuchungskorpus enthalten sind, befinden sich im Anhang D. Die verschiedenen *Tags* können im Anhang in der Tabelle B.2 eingesehen werden.

6.5 Falsche Annotation in *Frantext*

Einige *Graphien* in diesem Subkorpus werden durch den *catégorisateur* von *Frantext* nicht erkannt (vgl. Tabelle B.3 im Anhang) oder nicht bearbeitet: Im Korpus taucht 530mal <que> und 419mal <qu'> auf, die nicht getaggt wurden. <Où> (253 Belege), <comme> (154 Belege) und <sinon> (3 Belege) werden nicht bearbeitet.

Häufig bleiben im erstellten Korpus einzelne Satzzeichen stehen, die ohne Bezug zum folgenden Textausschnitt sind (Trennstriche z.B. 121mal), oder der Annotationsteil einer *Graphie* wird übernommen, während die dazugehörige *Graphie* innerhalb der eckigen Klammern abgeschnitten wird, da dort die Grenze des Kontextfensters verläuft (in 222 Fällen). Andere werden falsch annotiert, was nicht ohne Einfluß auf die vorliegende Auswertung bleibt.

Es folgt ein Beispiel für die häufige falsche Annotation durch den Tagger von *Frantext*. Dieser Ausschnitt ist durch eine Umformatierung in das `.son`-Format[9] übersichtlicher als die Ausgabe durch das Webinterface von *Frantext*.

[9]Siehe Beschreibung der Formate im Anhang unter A.1.1 (S. 127).

```
 2 - R670 - BENOZIGLIO.J-L - CABINET PORTRAIT - 1980 - page 93
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
[Dans Pp] [la D] [maison S] [paternelle A], [la D] [future S] [mariée A]
[pleure V] [la D] [fin S] [de Pp] [sa D] [jeunesse S], [s' Per]
[accroche V] [aux Dg] [jupes S] [de Pp] [sa D] [mère S], [clame V] [sa D]
[peur S], [son D] [dégoût S] [même Adv], [des Dg] [Np] [étrangers A]
[R] [que X] [sont V] [le D] [futur A] [mari S] [et Cc] [ses D]
[parents S].
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
```

Die *Graphie* <future> wurde als Substantiv identifiziert und die folgende *Graphie* <mariée> als Adjektiv. Diese Kombination ist grammatisch möglich (durch die feminine Form des Adjektivs <mariée> zugehörig zum Substantiv <future>), aber in diesem Fall nicht richtig.[10] Das Substantiv <jeunesse> wird korrekt erkannt, ebenfalls <jupes>, <mère>, <peur>, <dégoût>, <mari> und <parents>. Aber das Zeichen „≪“ wird als *Eigenname* getaggt (im Untersuchungskorpus insgesamt 223-mal), das schließende Zeichen „≫“ wird überhaupt nicht erkannt.

Weitere Fehler in anderen Korpusbelegen sind z.B. das Taggen von `[dans Pp] [un D] [bus APs] [bondé APs]`, in dem das Substantiv <bus> als *Adjectif/participe passé* getaggt wird (eine Verwechslung mit der graphisch identischen Form des *participe passé* im Plural von <boire>). Auch eine falsche Grenzziehung führt zu falscher Annotation: `[Nous Per] [habitons V] [maintenant Adv] [tous les trois un Adv] [peu Adv] [à l'extérieur Adv], [une D] [petite A] [maison S] [en Pp] [banlieue S].`

Während der vorliegenden Untersuchung wurden die gröbsten Annotationsfehler herausgelöscht, aber eine manuelle Korrektur des gesamten Untersuchungskorpus (6.404.598 *Graphien*) war nicht möglich.

[10] Um solche falschen Annotierungen zu vermeiden wäre eine Disambiguierung der *Graphie* <future> durchzuführen. Bei einer maschinellen Annotation des Korpus sind solche Fehler nur durch ein sehr verläßliches Disambiguierungsprogramm zu vermeiden.

6.6 Kontextauswertung durch *Frantext*

Die Werkzeuge für das Korpus *Frantext* bieten ebenfalls eine (rudimentäre) statistische Untersuchung des Kontexts. Dabei werden die Graphien, die sich in einem einstellbaren Kontextfenster um das untersuchte Wort befinden, auf ihre Häufigkeit untersucht. Diese Untersuchung und ihr Ergebnis wird im Abschnitt 7.1 genauer dargestellt.

Kapitel 7

Motivation

Für die Motivation des Ansatzes, das semantische Netz *FrenchWordNet* mit den statistischen Auswertungen einer Korpusuntersuchung anzureichern, wird in diesem Kapitel beispielhaft der Satzkontext des Substantivs *maison* analysiert. Daraufhin wird untersucht, wie weit diese Kollokationen im semantischen Netzwerk von den Konzeptknoten entfernt sind. Durch die Konzeption des semantischen Netzes (nur Knoten, die durch lexikalische und semantische Relationen verbunden sind, sind in der Netzstruktur benachbart), werden die Kollokationen innerhalb eines Textfensters nicht in der Nähe des Ausgangsknotens im Netz zu finden sein. Daher liegt die Vermutung nahe, daß eine Ergänzung des semantischen Netzes durch Informationen aus der Korpusanalyse eine gute Wissensdatenbank für die Lesartendisambiguierung darstellt.

7.1 Korpusuntersuchung

Ausgehend von der *Graphie* <maison> (ohne Berücksichtigung der möglichen *Synset*-zugehörigkeit) werden mit Hilfe der Kontextsuche von *Frantext* die Elemente eines Kontextfensters von einem einzelnen Satz innerhalb des Untersuchungskorpus[1] zusammengestellt.

Hierbei bieten die Werkzeuge von *Frantext* allerdings nicht die Möglichkeit, lediglich die Substantive innerhalb dieses Kontextfensters zu betrachten, daher ist die Berechnung der relativen Frequenz im Verhältnis zu der Gesamtzahl der

[1] Vgl. Abschnitt 6.4.

138	mère	89	maintenant	69	côté	56	campagne
131	rue	89	nuit	68	aller	56	cour
120	grand	87	bois	68	parents	56	femmes
119	temps	82	soir	67	ans	56	heures
117	jour	80	verte	66	air	56	homme
113	enfants	78	place	64	fille	55	hommes
109	grande	77	famille	64	sorte	55	murs
104	père	77	femme	64	tête	55	personne
102	porte	75	enfant	62	seule	54	autour
96	vie	75	jardin	60	jours	54	chambre
95	petit	75	vieille	59	long	54	dessus
92	ville	74	monde	57	petits	54	eau
90	petite	73	presque	57	tard	53	arbres

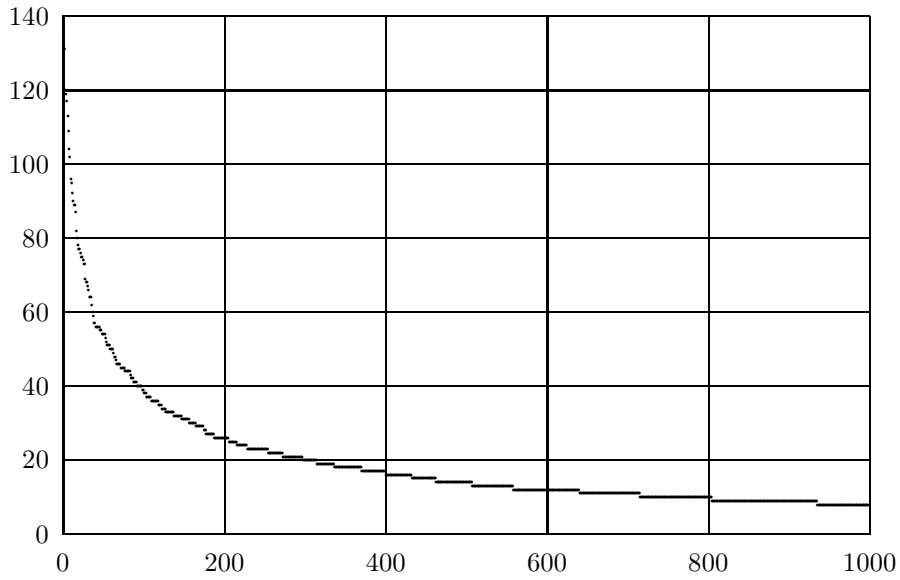
Tabelle 7.1: Umgebung von *maison* und *maisons*.[3]

Graphien verfälschend.[2] *Frantext* geht von 2.868 Belegstellen für Singular und Plural von *maison* aus und gibt an, daß der Singular insgesamt 2.587 Belegstellen im Korpus besitzt, der Plural 565. Die von *Frantext* angegebene Summe 3.152 erklärt sich dadurch, daß bei mehrfachem Auftauchen innerhalb eines Satzes jedes Element pro Belegstelle gezählt wird (also mehrfach). In dem Korpus finden sich insgesamt in der Satzumgebung von <maison>/<maisons> 136.577 *Graphien*, die *Frantext* auf insgesamt 15.637 Graphien zurückführt (flektierte Formen werden von *Frantext* leider einzeln aufgeführt). Die 52 häufigsten Elemente werden in Tabelle 7.1 wiedergegeben. Die durch Flexion unterschiedenen Elemente werden getrennt aufgeführt (z.B. <femme>/<femmes>). An allen Graphien (15.637) hat <mère> (als häufigste Graphie nach den Angaben von *Frantext*) mit 138 einen Prozentanteil von 0,88.

Der folgende Graph zeigt den Kurvenverlauf bei der Gegenüberstellung der einzelnen Graphien und der jeweiligen Anzahl der Belegstellen im Kontextfenster (nach den ersten 1.000 Graphien wird die x-Achse abgeschnitten). Auf der x-Achse sind die einzelnen Graphien aufgeführt (nicht getrennt aufgezählt, es zählt der Rang) und die y-Achse zeigt die Anzahl der Belegstellen im Kontextfenster. Dabei ist zu sehen, daß nur die ersten 91 Graphien (von 15.637) mehr als 40 Belegstellen haben. Über 14.800 Graphien haben weniger als 10 Belegstellen.

[2]Diese Relation wird im nächsten Abschnitt mit eigenen Werkzeugen aus den Ausgangsdateien (vgl. Abschnitt A.1 auf S. 127) berechnet.

[3]Das als *Graphie* gekennzeichnete <..> wurde hier nicht berücksichtigt.



Graphie und Anzahl der Belegstellen im Satzkontext

Für die weitere Arbeit werden diese Auswertungen detaillierter für das extrahierte Subkorpus (für die Erstellung dieses Untersuchungskorpus vgl. A.1) durchgeführt.

7.2 Erste Auswertung des Untersuchungskorpus

7.2.1 Probleme mit *Frantext*

Die vorangegangene Auswertung wurde mit den Werkzeugen von *Frantext* ebenfalls ausschließlich für den Singular <maison> durchgeführt: Dabei ergeben sich bei der Kontextuntersuchung durch das Untersuchungswerkzeug *Etude du voisinage d'un mot* 133 Belege für <mère> und sieben Belege für <mères> im Satzkontext. Wird nun aber mit dem Untersuchungswerkzeug *Lancer une recherche* nach der Kookkurrenz <maison>-<mère> und getrennt davon nach <maison>-<mères> gesucht, ergeben sich 125 Belege für die erste Kookkurrenz und sieben Belege für die zweite. Damit ist innerhalb der Werkzeuge von *Frantext* schon keine Übereinstimmung gegeben.[4]

[4]Ein Grund für die fehlende Übereinstimmung für <mère> konnte nicht ermittelt werden, da das Tool *Etude du voisinage d'un mot* keine Satzbelege angibt, sondern lediglich eine Liste der Kookkurrenzen, bei der auch die Reihenfolge der Graphien keine Rolle spielt.

Die Ergebnisse von *Frantext* stimmen auch nicht mit den Ergebnissen aus den vorliegenden Untersuchungen überein, die für <mère, mères> insgesamt nur 89 Belege und zusätzlich 17 Belege für Komposita (die ja von *Frantext* bei der Frequenzuntersuchung getrennt betrachtet werden) aufzählen. Es müssen von *Frantext* mehr Belegstellen gezählt worden sein, als in der aus *Frantext* extrahierten Basisdatei für die durchgeführten Untersuchungen vorliegen.

Ein Grund kann in der fehlerhaften Annotation des Korpus liegen. Die Ausgabe der Belegstellen durch das *Webinterface* von *Frantext* ist häufig bruchstückhaft, ein Teil der *Graphien* im Korpus wird durch die Annotationswerkzeuge von *Frantext* nicht erkannt oder falsch zugeordnet.[5] Ein Vergleich der eigenen Untersuchungen mit den Ergebnissen der Tools von *Frantext* ist daher so nicht möglich und für die weitere Arbeit werden ausschließlich die Zahlen aus dem eigenen umformatierten Korpus zugrunde gelegt.

7.2.2 Substantive im Satzkontext von *maison*

Mit den Angaben aus den Dateien (vgl. Abschnitt A.1) kann der Anteil einzelner Substantive an der Gesamtzahl der Substantive im Satzkontext berechnet werden.

Mit der Hilfe von kleineren Programmen wurden aus der manuell leicht verbesserten `.son`-Datei alle mit dem Annotationstag <S> versehenen Elemente herausgesucht: 14.275 Substantive (ohne Eigennamen), die sich 4.330 Graphien zuordnen lassen. Von diesen Substantiven sind 2.537 *Graphien* <maison, maisons> zuzuordnen. Damit bleiben 11.738 andere Substantive, die im weiteren betrachtet werden.

Insgesamt finden sich in diesem Untersuchungskorpus 64.080 *Graphien*, die weder als Satzzeichen betrachtet, noch nicht berücksichtigt oder nicht erkannt wurden, wobei auch hier Singular und Plural (falls graphisch unterschieden) einzeln gezählt werden. Damit ergibt sich für <mère, mères> ein Prozentanteil von 0,62 (mit Komposita 0,74%), so daß <mère, mères> das häufigste Substantiv in der Satzumgebung von <maison> darstellt.

Werden nun die unterschiedlichen Lesarten (hier mit Bezug auf die *Synsets* von <maison> in *FWN*) berücksichtigt, ergibt sich ein deutlicheres Bild:[6]

[5] Vgl. Abschnitt 6.5.

[6] Eine genauere Auswertung findet sich in 8.4.1.

- bei [*maison:1*, *habitation:1*] ist <mère, mères> das dritthäufigste Substantiv (0,47)%),
- bei [*maison:2*, *signe:1*] taucht es nicht auf,
- bei [*maison:3*, *chez-soi:2*] in 1,34% der Fälle (fünfthäufigstes Substantiv)
- und bei [*maison:4*, *firme:1*] in 1,15% (ist mit anderen Substantiven am häufigsten belegt).

In den Dateien, in denen <maison> als Teil eines Kompositum, als Eigenname oder in nicht eindeutig zuweisbarer Kategorie auftaucht, ist <mère/mères> mit 23 Belegen (zuzüglich 3 Komposita) vorhanden (bei 2.659 Substantiven ein Prozentsatz von 0,87%).

7.3 Kontextfenster und semantisches Netz

Um zu zeigen, daß sich die im Netz nahe bei dem Ausgangsknoten liegenden *Synsets* im Korpus nicht im betrachteten Kontextfenster befinden, werden die Hyponym- und Meronymknoten aus *FWN* mit der entstandenen Frequenzliste verglichen. Da die Kontextsuche von *Frantext* Komposita in ihre Einzelteile zerlegt, war es nicht ohne weiteres möglich, nach z.B. <cabinet de travail> oder <étage mansardé> zu suchen, daher wurden diese Belegstellen nicht berücksichtigt (angegeben durch ø).

Außerdem gibt *Frantext* keine Lesartendifferenzierung aus, so daß die Frequenzen in diesem Fall ungeachtet der möglichen *Synset*-zugehörigkeit ermittelt wurden (bei Auftreten als Kollokation bei mehreren Elementen des *Synsets* wird die Frequenz mehrfach angegeben). Die Zahlen geben nacheinander die Anzahl der Kollokationen bei den Elementen des *Synsets* an (Singular und Plural zusammengezählt).

Die nächstliegenden *Synsets* des ersten Konzepts [*maison:1*, *habitation:1*] haben folgende Frequenzen:

Hyperonym [*pension:1*, *logis:1*, *logement:1*]: 1+7+4,
Hyponyme [*abri:7*]: 17, [*pension:2*]: 1, [*propriété:2*, *résidence:3*]: 4+2, [*résidence:1*, *manoir:1*]: 2+2, [*maison de campagne:2*, *bungalow:1*]: ø+1, [*maison indépendante:1*]: ø, [*propriété:3*, *maison de campagne:1*]: 4+ø, [*chambre d'ami:1*] ø,

Meronyme [*porche:1*, *véranda:2*]: 8+9, [*étage mansarde:2*, *grenier:2*, *mansarde:2*] ø+16+0, [*cabinet de travail:1*]: ø.

Für die verknüpften *Synsets* von [*maison:2*, *signe:1*] gilt:

Hyperonym [*région:5*, *étendue:4*]: 17+8,
Hyponyme gibt es nicht.

Das Synset [*maison:3*, *chez-soi:2*] ist nur mit seinem **Hyperonym** verknüpft: [*résidence:5*, *demeure:1*]: 5+6.

Die Belegstellen für die verknüpften *Synsets* von [*maison:4*, *firme:1*] sind:

Hyperonym [*entreprise:1*, *société:4*]: 5+12,
Hyponyme [*marchand:5*, *négociant:1*]: 6+1 und [*maison d'édition:1*, *société d'édition:1*, *éditeur:1*]: ø+ø+4.

Umrechnung des Anteils der häufigsten Substantive

Der Vergleich mit den oben untersuchten Hyperonym-, Hyponym- und Meronymknoten der *Synsets* von <*maison*> zeigt, daß nur ein *Synset* unter die ersten 215 Graphien eingeordnet wird ([*région:5*, *étendue:4*], 25 Belege), daß die anderen *Synsets* aber nicht unter den ersten 370 Graphien zu finden sind ([*abri:7*]: 17 Belege, [*porche:1*, *véranda:2*]: insgesamt 17 Belege, [*entreprise:1*, *société:4*]: 17 Belege, [*étage mansarde:2*, *grenier:2*, *mansarde:2*]: 16 Belege).

Damit hat das *Synset* mit der größten Anzahl von Belegstellen lediglich einen Prozentanteil von 0,16 während <mère, mères> in der Umgebung einer einzelnen Lesart einen dreimal höheren Prozentsatz besitzt: 0,47%.

7.4 Schlußfolgerung

Damit ist gut zu erkennen, daß die *Synsets*, die sich in unmittelbarer Umgebung eines Konzepts befinden, innerhalb eines Korpus nicht in der Satzumgebung auftauchen, aber andere Elemente, die sich häufig in der Umgebung einer bestimmten Lesart befinden, im vorliegenden semantischen Netz nicht in der unmittelbaren Umgebung des Ausgangselementes vorkommen.

Diese kleine Korpusuntersuchung motiviert den Ansatz, das Netzwerk mit den Informationen, die sich durch die Korpusuntersuchung (insbesondere durch

die prozentuale Verteilung der verschiedenen Graphien auf die unterschiedlichen *Synsets*) ergeben, anzureichern. Damit würde für ein Maß des Zusammenhangs, das auf der Grundlage des Netzes erstellt wird, zusätzlich das miteinbezogen, was der menschliche Sprecher wirklich in der Rede assoziiert.

Kapitel 8

Auswertung des Korpus

Motiviert durch die Beobachtungen aus dem vorangegangenen Kapitel soll für ein Zusammenhangsmaß eine Kombination des semantischen Netzwerkes *FWN* mit einer statistischen Korpusauswertung erstellt werden.

Die Ergänzung des Netzwerkes durch eine statistische Korpusuntersuchung soll in den ersten Schritten manuell erfolgen. Zusätzlich zu den Knoten des semantischen Netzwerkes werden die häufig mit einer Lesart zusammen auftauchenden Elemente manuell in das Netz an das Ursprungswort angehängt und wie neue Knoten im Netz behandelt.

Daher wird zuerst eine Statistik erstellt, die die Elemente im Kontext einer *Graphie* betrachtet. In dem vorliegenden Fall wird für das Substantiv <maison>[1] ein Subkorpus mit *Frantext* erstellt (vgl. Abschnitt A.1), das die Satzkontexte für die 2.361 Belege von <maison> enthält. In diesem Subkorpus werden manuell die von *FWN* unterschiedenen Lesarten disambiguiert.

Nach der Trennung dieses großen Subkorpus in die Teilkorpora, die jeweils nur die Belege für eine der unterschiedenen Lesarten enthält, können für die jeweilige Lesart charakteristische Strukturen herausgearbeitet werden. In der vorliegenden Arbeit werden die Substantive, die sich innerhalb eines Satzkontexts um <maison> befinden, betrachtet. Diese Auswahl bietet sich an, da *FWN* (mit Einschränkungen) ein gut ausgebautes Substantiv-Teilnetz besitzt[2] und die Verbindung mit den anderen Teilnetzen (z.B. zum Netz der Verben oder der Adjektive) nicht gegeben ist.

[1] Die Statistik behandelt nur den Singular, die 409 Belege für den Plural werden nicht betrachtet.

[2] Siehe aber auch den Abschnitt 9.6.

Die flektierten Formen der Substantive lassen sich manuell (da *Frantext* diese Möglichkeit nicht bietet) auf eine einzige Form zurückführen (vgl. Abschnitt 8.4).

Eine ausführliche statistische Auswertung der Kollokationen stellt für jede unterschiedene Lesart häufige Kombinationen heraus. Dabei muß berücksichtigt werden, ob eine häufige Kollokation bei der einen Lesart nicht auch bei einer anderen häufiger auftaucht. Für jede Lesart wird am Ende der Auswertung eine Liste von aussagekräftigen Kollokationen zusammengestellt, die dann durch eine neue Verbindung im semantischen Netzwerk mit dieser Lesart verknüpft werden. Diese neue Kante verknüpft die jeweiligen Konzeptknoten der Lexie *maison* mit den Kollokationen, die schon als Knoten im Netz existieren (vgl. Abschnitt 8.9).

Mit diesen neuen Kanten kann ein Zusammenhangsmaß modelliert werden, das die ursprünglichen und die neuen Verbindungen miteinbezieht (vgl. Abschnitt 10) und je nach Art der Verknüpfung (ursprüngliche Netzstruktur oder Erweiterung) eine Gewichtung vornimmt.

8.1 Manuelle Disambiguierung

Das aus *Frantext* extrahierte Teilkorpus mit den 2.361 Belegen für <maison> wurde in ein besser lesbares Format umformatiert. Darin werden die Belegstellen für <maison> manuell disambiguiert und mittels einer Annotation im Textformat kenntlich gemacht.[3]

```
37 - R762 - SABATIER.R - LES FILLETTES CHANTANTES - 1980 - page 134
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
[La D] [maison S], [toute A] [en Pp] [longueur S] [entre Pp]
[deux Dca] [jardins S], [ne Avn] [portait V] [qu' X] [un D] [étage S]
[mansardé A].
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
```

wird annotiert zu:

[3] Vgl. Abbildungen im Abschnitt A.1.1 im Anhang auf Seite 130.

```
37 - R762 - SABATIER.R - LES FILLETTES CHANTANTES - 1980 - page 134
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
[La D] [§1§maison S], [toute A] [en Pp] [longueur S] [entre Pp]
[deux Dca] [jardins S], [ne Avn] [portait V] [qu' X] [un D] [étage S]
[mansardé A].
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
```

Die Grenzen der einzelnen Lesarten wurden anhand der in Abschnitt 4.5 ermittelten Unterscheidungen getroffen. Neben den vier von *FWN* unterschiedenen Lesarten werden noch drei weitere Kategorien verwendet: <maison> als Teil eines Eigennamens (z.B. *Maison-Blanche*, *maison Tellier*), als Teil eines Kompositum (z.B. *maîtresse de maison*) oder die Restgruppe, in der die Belege gesammelt werden, die nicht eindeutig einer der sechs anderen Kategorien zugeordnet werden können.

In vielen der Fälle ist für den menschlichen Annotator die Kategorienzuweisung klar, in wenigen Belegstellen ist die Zuweisung aufgrund der Lesartenabgrenzung (besonders zwischen Lesart 1 und Lesart 3) nicht sofort eindeutig, sondern kann erst nach einer weiteren Eingrenzung der Lesart, die für weitere Fälle übernommen wird, festgelegt werden. Diese weitere Eingrenzung ergänzt die Konzeptumschreibung der von *FWN* vorgegebenen Glosse um einige Aspekte.[4]

Die Lesart [*maison:1*, *habitation:1*] („Gebäude“) hat in dem Subkorpus 1.562 Belege, die Lesart [*maison:2*, *signe:1*] ist zweimal belegt, die Lesart [*maison:3*, *chez-soi:2*] wurde 360mal zugewiesen und die Lesart [*maison:4*, *firme:1*] in 49 Fällen. In 133 Textstellen war <maison> Teil eines Eigennamens, in 237 Fällen Teil eines Kompositums und in 18 Belegen war die Belegstelle zu klein, um eine eindeutige Zuweisung vorzunehmen. Die folgenden Beispiele verdeutlichen diese Zuweisung.

Wenn der Gebäudeaspekt im Vordergrund stand, wurde die erste Kategorie zugeordnet:

```
10 - R670 - BENOZIGLIO.J-L - CABINET PORTRAIT - 1980 - page 257
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
[Juste A] [devant Pp] [la D] [§1§maison S].
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
```

[4]Diese Unterscheidungen sind im Anhang C aufgelistet.

Ganz klar war jeweils die Zuweisung zur zweiten Lesart in den beiden Belegstellen des Subkorpus:

```
113 - R544 - MATZNEFF.G - IVRE DU VIN PERDU - 1981 - page 327
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
[Le D] [Soleil S] [en Pp] [§2§maison S] [neuf A], [Jupiter Np]
[près du Pp] [milieu S] [du Dg] [ciel S], [conjonction S] [Mars S]-
[Pluton Np], [votre D] [ascendant S] [serait V] [tombé APs] [dans Pp]
[le D] [Sagittaire S] [et Cc] [non Adv] [dans Pp] [le D]
[Scorpion S], [Nil Np] [écoute S] [les D] [explications S] [de Pp]
[l' D] [astrologue S] [sans Pp] [les Per] [comprendre Inf] [ni Cc]
[les Per] [retenir Inf].
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
```

Wenn *maison* ohne Einschränkung durch das andere Element des *Synsets* (*chez-soi* ersetzt werde konnte, wurde die dritte Lesart zugewiesen:

```
118 - R757 - MORDILLAT.G - VIVE LA SOCIALE - 1981 - page 18
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
[Filant Pr], [éplorés APs], [mon D] [oncle S] [et Cc] [mon D]
[père S] [coururent V] [à Pp] [la D] [§3§maison S] [raconter Inf]
[leur D] [malheur S].
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
```

Die Zuweisung zur vierten Kategorie (die französische Entsprechung zum deutschen Konzeptknoten [*Firma:1*]) war auch mit einem kleinen Kontextausschnitt einfach:

```
123 - R757 - MORDILLAT.G - VIVE LA SOCIALE - 1981 - page 102
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
[La D] [§4§maison S] [ne Avn] [faisait V] [pas Avn] [crédit S].
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
```

Als Eigenname tauchte die *Graphie* <maison> in rund 5% der Fälle auf (vgl. die Liste der gefundenen Eigennamen in Anhang C):[5]

[5]Der Eigenname <Lalaire> wurde vom Annotationsprogramm von *Frantext* nicht erkannt und daher auch nicht getaggt.

```
558 - R769 - HANSKA.E - LES AMANTS FOUDROYES - 1984 - page 107
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
[Au cours d' Pp][une P] [des Dg] [sauteries S] [de Pp] [la D]
[§===§maison S] Lalaire, [j' Per][ai V] [demandé Ps] [à Pp] [un P]
[des Dg] [convives S] [l' D][adresse S] [d' Pp][un D]
[psychanalyste S].
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
```

In etwa 10% der Belegstellen war <maison> Element eines Kompositum. Eine Liste der gefundenen Komposita findet sich im Anhang C.

```
93 - R544 - MATZNEFF.G - IVRE DU VIN PERDU - 1981 - page 18
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
[Le D] [malheureux S] [a V] [fait Ps] [deux Dca] [tentatives S]
[de Pp] [suicide S], [un D] [séjour S] [en Pp] [§!!!§maison S]
[de Pp] [repos S] ;
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
```

Manchmal war das im Untersuchungskorpus gegebene Textfenster zu klein, um eine Zuordnung vorzunehmen. Um hier eine genaue Zuweisung vorzunehmen, hätte im Korpus ein größeres Textfenster betrachtet werden müssen.

```
1131 - R725 - BIENNE.G - LE SILENCE DE LA FERME - 1986 - page 114
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
[Np] [Et Cc] [la D] [§???§maison S] ?
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
```

8.2 Auswertung der Kollokationen

Mit Hilfsprogrammen wird die einzelne Datei, die alle Belegstellen umfaßt, in sieben Dateien aufgesplittet, in denen sich nur die Belege befinden, die zu einer der sieben Kategorien gehören.[6]

Damit können diese Dateien einzeln automatisch auf Charakteristika untersucht werden, die die einzelnen Kategorien voneinander unterscheiden. Durch die POS-Annotation von *Frantext* (die allerdings fehlerhaft ist, vgl. Abschnitt

[6] Vgl. auch die Beschreibungen im Abschnitt A.1.2 auf Seite 131.

6.5) können die unterschiedlichen Wortarten im Kontext des *pivot* <maison> betrachtet werden, um aussagekräftige Häufigkeiten zu ermitteln. Für die Kombination mit der Struktur des *FWN* bieten sich allerdings nur die dort vorhandenen Wortartteilnetze an (Substantive, Verben, Adjektive und Adverbien), die unterschiedlich gut ausgebaut sind. In *FWN* sind bis jetzt nur das Substantiv- und das Verbteilnetz in sich durch semantische Verknüpfungen verbunden.[7]

Daher bietet sich für das französische semantische Netz an, das Substantivnetz zu ergänzen, also die Substantive aus dem Satzkontext und ihre Häufigkeiten in Bezug auf eine Lesart zu untersuchen. Da die Wortart-Teilnetze untereinander nicht verbunden sind, kann auch keine wortartenübergreifende Untersuchung stattfinden.

Damit ergibt sich als Fragestellung für die folgende Untersuchung: Welche Substantive befinden sich besonders häufig innerhalb eines Kontextfensters um das untersuchte *pivot* <maison>? Da *Frantext* außer der POS-Annotation keine weitere Annotation (z.B. syntaktische Verbindungen) zur Verfügung stellt, kann keine weitere Aussage zum Substantiv im Satzkontext gemacht werden.

8.3 Auswertung der Substantive im Kontext

Die Ergebnisse für die Belege von <maison> als Teil eines Kompositums oder als Eigenname werden hier nicht mit ausgewertet. Die Komposita sind größtenteils als Netzknoten nicht in *FWN* vorhanden und ihr jeweiliger Kontext ist für diesen Ansatz zu heterogen.

Für jede Lesart werden einzeln Auswertungen durchgeführt: Zuerst wurden alle Substantive herausgefiltert, um die frequentesten zu bestimmen.[8] Die Substantive, die in mindestens 1% der Fälle auftreten, sollen im folgenden betrachtet werden. Diese Grenze ist wegen des manuellen Aufwandes willkürlich gezogen, bei einer möglichen Automatisierung könnte eine größere Menge betrachtet werden. Es ergeben sich schon bei einem ersten Blick auf die maschinell erstellten Listen deutliche Unterschiede. Für die erste Lesart sind es folgende (von insgesamt 3.335 verschiedenen) Wortformen:

[7] Vgl. Abschnitt 4.3.
[8] Vgl. die Beschreibung der Programme zu dieser Auswertung im Anhang A.1.2.

113 maison	32 soir	21 voiture	17 quartier
51 porte	32 homme	21 palais	17 mer
46 famille	30 vie	21 escalier	17 hommes
45 mère	28 temps	21 bras	17 fond
42 jour	28 pièce	20 sorte	17 campagne
41 jardin	26 terre	20 fille	17 arbres
41 bois	26 fois	20 étage	16 yeux
38 rue	25 ville	20 enfant	16 table
38 cour	25 parents	19 nom	16 soleil
37 père	24 ans	19 main	16 silence
36 nuit	24 air	19 jours	16 salle
35 enfants	23 eau	18 fleurs	16 odeur
35 chambre	23 années	17 vacances	16 fenêtres
34 femme	22 mari	17 tête	16 femmes

Bei der zweiten Lesart werden hier wegen der sehr geringen Belegzahl alle aufgeführt:

2 mars	1 sagittaire	1 explications	1 astrologue
1 suite	1 mots	1 écoute	1 ascendant
1 soleil	1 milieu	1 conjonction	
1 scorpion	1 judaïsme	1 ciel	

Für die dritte Lesart werden wieder nur die ersten 57 herausgezogen (von 823 verschiedenen):

23 père	7 temps	5 mort	4 papa
17 soir	7 maman	5 moment	4 mot
17 mère	7 maison	5 mois	4 monde
15 vie	7 lit	5 fois	4 médecin
15 enfants	7 jours	5 façon	4 matin
13 jour	7 famille	5 envie	4 mari
13 heures	7 classe	5 dimanche	4 main
12 rue	6 voiture	5 bureau	4 frères
12 école	6 vacances	5 amis	4 copains
11 enfant	6 place	4 téléphone	4 chemin
10 retour	6 frère	4 table	4 chambre
9 parents	6 fille	4 silence	4 ans
8 porte	6 années	4 semaine	
8 nuit	5 voix	4 reste	
8 femme	5 tête	4 paris	

Da bei der vierten Lesart nur wenige Belege mehrfach auftauchen, werden hier nur diejenigen aufgeführt, die mindestens zweimal belegt sind (von 166 verschiedenen).[9]

21 disques	3 mère	2 maison	2 corps
4 paris	3 directeur	2 guerre	2 contrat
3 production	2 semaine	2 esprit	2 arrêt
3 patron	2 nom	2 duc	2 a
3 papa	2 matin	2 correction	

8.4 Bestimmung der neuen Verbindungen

Nach der rein maschinellen Auswertung der Kontexte steht die Überlegung, welche dieser Substantive als aussagekräftig im Hinblick auf die Lesart, in der sie vorkommen, bezeichnet werden können (Signifikanz). Nur diese kommen für eine neue Verbindung im semantischen Netzwerk in Frage.

Als erstes müssen daher die Listen für die verschiedenen Lesarten gegeneinander abgeglichen werden. Falls ein Substantiv bei mehreren Lesarten frequent ist, verfälscht die Verknüpfung mit dem Ausgangssubstantiv bei einer Einbindung ins Netz die Ergebnisse. Auch kann ein insgesamt im ganzen Korpus sehr frequentes Substantiv auf eine einzelne Lesarten keinen Hinweis geben. Ist für den menschlichen Sprecher keine klare Assoziation gegeben, sollte dieses Substantiv ebenfalls wegfallen (manueller Eingriff).

Danach werden die anderen Elemente aus dem betreffenden *Synset* betrachtet (*habitation:1*, *signe:1*, *chez-soi:2* und *firme:1*). Als letzter Schritt werden diejenigen Substantive aus diesen Listen zusammengefaßt, die eine Verbesserung für das Zusammenhangsmaß versprechen (manuelle Filterung). Da entschieden werden muß, welche der Lesarten der Substantive mit dem Ausgangsknoten ([*maison:1*, *habitation:1*]) verbunden wird, werden die Substantive aus der Liste nach ihrer Lesart in *FWN* disambiguiert.

Bei dieser Arbeit hat sich herausgestellt, daß der manuelle Eingriff an vielen Stellen erfolgen mußte, eine Automatisierung dieses Vorganges insgesamt daher

[9] Es ist hier gut zu sehen, daß einige Elemente in *FWN* falsch getaggt wurden: z.B. <paris> als Substantiv statt als Eigenname (Troponym, die Einordnung in die Kategorie *Substantiv* ist gegenüber der existierenden Kategorie *Eigenname* nicht zu halten) und die *Graphie* <a> als Substantiv. Es ist im Korpus der zweite Teil einer Abkürzung, deren erstes Element als Eigenname (*Np*) getaggt wurde.

sehr erschwert wird. Eine Automatisierung kann bei der Zusammenfassung der Singular- und Pluralformen (durch einen Tagger) erfolgen und bei eindeutigen Zahlenverhältnissen.

8.4.1 Grundlagen für die Auswertung

Für die weiteren statistischen Auswertungen werden folgende Daten verwendet:

- Insgesamt enthält das Untersuchungskorpus der 87 Texte 6.404.598 *Graphien*. Der Kontext eines Belegs besteht aus dem Satz, in dem dieser Beleg steht.
- Der Kontext der ersten Lesart besteht aus insgesamt 43.928 *Graphien* mit 9.631 Substantiven (3.335 verschiedene).
- Bei der zweiten Lesart finden sich 17 Substantive auf 61 *Graphien* (15 verschiedene Substantive).
- In den Satzkontexten der dritten Lesart finden sich 9.113 *Graphien* mit 1.341 Substantiven (823 verschiedene).
- In der vierten Lesart sind 260 Substantive auf 1.127 *Graphien* (166 verschiedene Substantive).

8.4.2 Auswertung für [*maison:1*, *habitation:1*]

maison:1

Bei insgesamt 1.561 Belegen für die erste Lesart von <maison> sollten diejenigen Substantive betrachtet werden, die in mindestens 16 Satzkontexten belegt sind. Die oben schon aufgeführte Liste läßt sich durch die Zusammenfassung von Singular- und Pluralformen verkürzen, dabei wird dann jeweils nur die Singularform aufgeführt. Hier kann mit geeigneten Datenbanken (mit der Einbeziehung von Flexionsformen) eine Automatisierung erfolgen. Auch wird <maison>, wenn es ein zweites Mal im Satz auftaucht, nicht betrachtet. Andere Formen, die durch die Zusammenführung von Singular und Plural auf mehr als 16 Belege kommen, werden in die Liste mit aufgenommen.

Die entstandene Liste enthält 44 Substantive, die insgesamt 12,75% aller Substantive aus dem Kontext stellen (1.228 von 9.631). Die Umrechnung der absoluten Frequenz in Prozentangaben relativiert diese Zahlen gegenüber den Werten der anderen Lesarten. In der folgenden Tabelle ist die Zahl vor dem Substantiv die absolute, die nachfolgende Zahl die relative Häufigkeit auf alle *Graphien* im Kontext der Lesart. Die Zahl in Klammern gibt den Anteil an den Substantiven im Kontext an.

61 jour	0,14%	(0,63%)	30 vie	0,07%	(0,31%)	22 sorte	0,05%	(0,23%)
60 porte	0,14%	(0,62%)	28 main	0,06%	(0,29%)	21 bras	0,05%	(0,22%)
55 enfant	0,13%	(0,57%)	28 temps	0,06%	(0,29%)	21 nom	0,05%	(0,22%)
50 femme	0,11%	(0,52%)	27 oeuil	0,06%	(0,28%)	20 odeur	0,05%	(0,21%)
49 homme	0,11%	(0,51%)	26 fois	0,06%	(0,27%)	19 fond	0,04%	(0,20%)
48 rue	0,11%	(0,50%)	26 ville	0,06%	(0,27%)	19 fleur	0,04%	(0,20%)
47 famille	0,11%	(0,49%)	26 étage	0,06%	(0,27%)	19 quartier	0,04%	(0,20%)
45 mère	0,10%	(0,47%)	25 parents	0,06%	(0,26%)	18 mur	0,04%	(0,19%)
44 chambre	0,10%	(0,46%)	24 air	0,05%	(0,25%)	17 salle	0,04%	(0,18%)
41 bois	0,09%	(0,43%)	24 eau	0,05%	(0,25%)	17 vacances	0,04%	(0,18%)
41 père	0,09%	(0,43%)	24 an	0,05%	(0,25%)	17 tête	0,04%	(0,18%)
41 nuit	0,09%	(0,43%)	24 fenêtre	0,05%	(0,25%)	17 table	0,04%	(0,18%)
39 pièce	0,09%	(0,40%)	23 année	0,05%	(0,24%)	16 soleil	0,04%	(0,17%)
34 fille	0,08%	(0,35%)	23 voiture	0,05%	(0,24%)	16 silence	0,04%	(0,17%)
34 soir	0,08%	(0,35%)	22 mari	0,05%	(0,23%)			

Der gesamte Verlauf der Kurve wird in der Abbildung 8.1 deutlich. Die Zahl der Graphien mit einer Häufigkeit unter 1% ist sehr hoch (3.295), damit verläuft die Kurve sehr schnell sehr flach. In der Abbildung 8.2 ist im Ausschnitt dann der schnell fallende Kurvenverlauf bei den 40 häufigsten Substantiven zu sehen.
In der folgenden Tabelle sind diese Substantive zusammengestellt mit den Prozentzahlen, die den Anteil in den anderen Satzkontexten darstellen. Die Zahl vor dem Substantiv zeigt den Anteil am Satzkontext bei [*maison:1*, *habitation:1*]. In den drei folgenden Spalten findet sich der Anteil, den dieses Substantiv in den drei anderen Lesarten hat (jeweils gerundet). Die vorletzte Spalte zeigt das Verhältnis im Gesamtkorpus an und die letzte Spalte vergleicht diesen Wert mit der ersten Spalte (Verhältnis Korpus/Lesart:1, Rechnung mit genaueren Werten).[10]

[10]Das Verhältnis im Gesamtkorpus ist nur um einen winzigen Bruchteil abweichend von dem Verhältnis des Komplements (Korpus ohne erste Lesart, Anzahl der Belege des Substantivs ohne die, die bei der ersten Lesart vorkommen). Für die bessere Lesbarkeit wird bei x=0 der Wert ganz weggelassen.

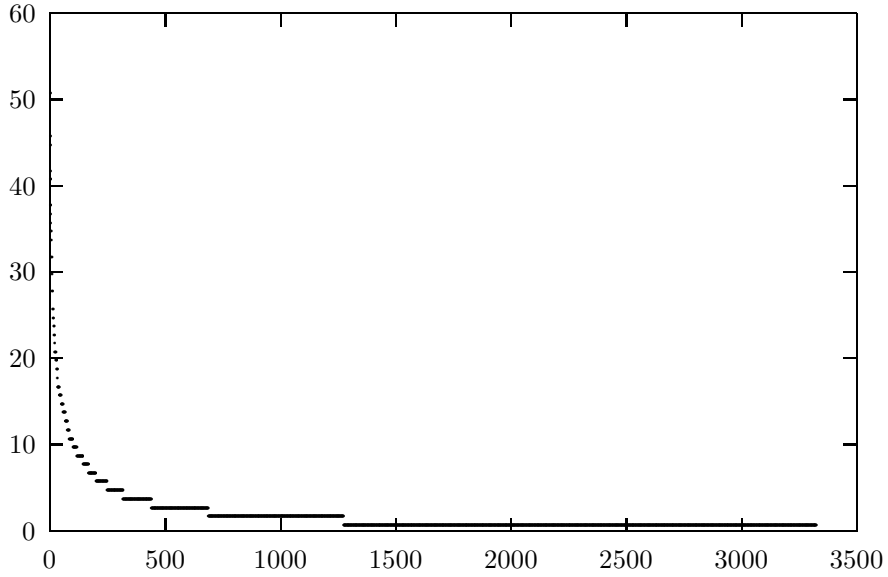


Abbildung 8.1: Frequenz aller Substantive im Kontext von <maison:1>

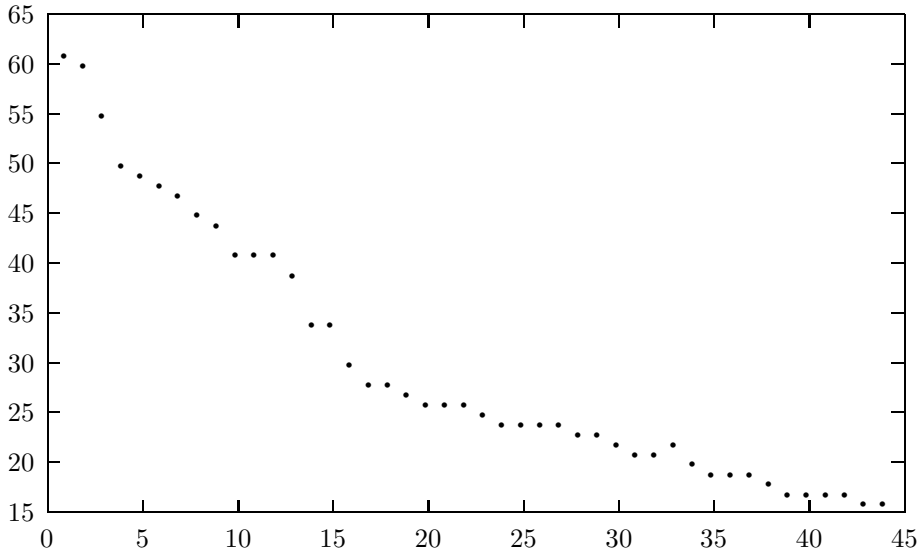


Abbildung 8.2: Frequenz der häufigsten Graphien im Kontext von <maison:1>

bei 1		bei 2	bei 3	bei 4	im Korpus	Korpus/Lesart:1
0,06%	étage		0,02%		0,01%	14,67%
0,04%	vacances		0,07%		0,01%	18,92%
0,09%	bois		0,02%		0,02%	24,09%
0,04%	quartier				0,01%	24,73%
0,05%	mari		0,07%	0,09%	0,01%	25,10%
0,11%	famille		0,08%	0,09%	0,03%	27,04%

0,10%	chambre		0,04%	0,03%	29,74%
0,09%	pièce		0,01%	0,03%	32,20%
0,04%	fleurs		0,01%	0,02%	36,10%
0,05%	odeur		0,01%	0,02%	36,83%
0,05%	fenêtre		0,01%	0,02%	37,32%
0,06%	parents		0,11%	0,02%	41,15%
0,05%	voiture		0,07%	0,02%	41,72%
0,14%	porte		0,09%	0,09% 0,06%	42,11%
0,04%	salle		0,01%	0,02%	44,70%
0,11%	rue		0,13%	0,05%	46,95%
0,05%	sorte		0,02%	0,03%	54,84%
0,06%	ville		0,02%	0,03%	57,53%
0,08%	soir		0,19%	0,05%	58,97%
0,04%	mur			0,02%	59,37%
0,05%	eau		0,01%	0,04%	64,39%
0,09%	père		0,29%	0,06%	64,76%
0,09%	nuit		0,09%	0,06%	66,05%
0,04%	silence		0,04%	0,03%	71,07%
0,04%	table		0,04%	0,09% 0,03%	72,42%
0,11%	femme		0,12%	0,08%	72,96%
0,04%	fond			0,09% 0,03%	74,11%
0,05%	année		0,10%	0,04%	74,46%
0,04%	soleil	1,64%	0,01%	0,03%	75,23%
0,08%	fille		0,09%	0,18% 0,06%	77,59%
0,10%	mère		0,20%	0,27% 0,08%	78,94%
0,14%	jour		0,23%	0,09% 0,11%	80,42%
0,13%	enfant		0,29%	0,10%	82,82%
0,05%	nom		0,01%	0,18% 0,04%	83,55%
0,05%	bras		0,02%	0,04%	87,27%
0,11%	homme		0,07%	0,09% 0,10%	92,69%
0,05%	air		0,03%	0,05%	100,60%
0,05%	an		0,05%	0,06%	116,97%
0,06%	oeuil		0,02%	0,08%	123,71%
0,07%	vie		0,16%	0,09% 0,09%	128,17%
0,06%	main		0,07%	0,09% 0,09%	145,99%
0,06%	fois		0,05%	0,09% 0,09%	159,28%
0,06%	temps		0,08%	0,09% 0,12%	189,45%
0,04%	tête		0,07%	0,08%	194,75%

Als erstes können aus der Liste der häufigsten Substantive im Kontext der ersten Lesart diejenigen Substantive herausgesucht werden, die im Gesamtkorpus (oder im Komplement) gleich häufig vorkommen. Ein Prozentwert über 100 in der letzten Spalte bedeutet, daß eher die Abwesenheit dieses Substantivs für diese Lesart charakteristisch ist (soweit der Wert sehr deutlich über 100% liegt). Da

diese Eigenschaft eines Substantivs aber für das Ziel, einen neuen Knoten an den *pivot*-Knoten zu knüpfen, nicht wesentlich ist, werden diese Substantive für die Erweiterung (zumindest für diese Lesart) ebenfalls nicht in Frage kommen.

Substantive, deren relative Korpusfrequenz in der letzten Spalte nahe bei 100% liegt, kommen in den anderen Lesarten ebenfalls in ähnlicher relativer Frequenz vor, daher können auch sie für diese Lesart nicht charakteristisch sein.

Außerdem werden die Substantive gestrichen, die in anderen Lesarten häufiger vorkommen.[11] Beträgt die Häufigkeit des Substantivs bei anderen Lesarten mehr als das zweifache des Wertes bei der ersten Lesart, wird das Substantiv aus dieser Liste gestrichen. Beispiele dafür sind hier <porte>, <vacances>, <mari>, <parents>, <soir>, <ville>, <père>, <fond>, <année>, <soleil>, <fille> und <mère>.

Andererseits dürfen sich die relativen Frequenzen auch nicht zu nahe stehen, da dies bedeutet, daß das Substantiv in mehreren Lesarten mit gleicher Häufigkeit vorkommt, also keinen Hinweis auf die Lesart geben kann. Damit fallen aus der vorstehenden Liste weitere Substantive heraus: <famille>, <fleur>, <voiture>, <rue>, <eau>, <nuit>, <silence>, <table> und <femme>. Dieser Auswahlprozeß kann automatisiert werden, da lediglich numerische Werte verglichen werden.

Als weiteres hat sich herausgestellt, daß Substantive, die als Zeitangabe verwendet werden, in vielen Kontexten vorkommen, daher werden sie aus allen Listen gestrichen (z.B. <an>, <année>, <dimanche> etc.). Hier ist eine Automatisierung nur mit Listen auszuschließender Substantive vorzunehmen.

habitation:1

Für das zweite Element aus dem *FWN-Synset* [*maison:1*; *habitation:1*] wurde das Ausgangskorpus wie für die Belegstellensuche für <maison> durchsucht, die insgesamt 23 Satzkontexte von <habitation, habitations> wurden manuell lesartendisambiguiert und die Substantive aus den jeweiligen Lesartenkontexten herausgesucht. Für die erste Lesart nach *FWN* finden sich 17 Belege (10mal Singular, 7mal Plural), die zweite Lesart [*habitation:2*] (<„aristocratic family line“>) ist nicht belegt, die dritte Lesart [*habitation:3*, *occupation:3*] (<„the act of

[11] Dieser Schritt ist sehr abhängig von der Ausgewogenheit des Korpus. In diesem Fall ist *Frantext* als Ausgangskorpus nicht ideal, da sich zumindest für die zweite Lesart und nahezu auch für die vierte Lesart nicht genügend Belege fanden, um etwaige Unregelmäßigkeiten des Korpus auszugleichen.

dwelling in a place“>) taucht einmal auf, die vierte [*habitation:4*] (keine Glosse) ist ebenfalls nicht belegt, die letzte [*habitation:5*] (<„the native habitat or home of an animal or plant“>) ist auch nur einmal vorhanden. Als Eigenname ist es einmal belegt, unklar blieben drei Belege.

Folgende Substantive kommen im Satzkontext der Lesart *habitation:1* mindestens zweimal vor:

4 chambre	3 fenêtre	2 population	2 guerre
4 rue	2 soir	2 plaine	2 gare
4 mur	2 siècle	2 ouverture	2 demeure
3 pièce	2 quartier	2 nuit	2 chinois
3 ordure	2 porte	2 mère	2 chansons
3 lit	2 port	2 maison	

Ein Vergleich mit den Substantiven der Satzkontexte bei den übrigen Lesarten von <habitation:1> ergibt keine Übereinstimmungen. Allerdings ist bei der sehr geringen Anzahl von Belegen in der jeweiligen Lesart keine Aussage zu treffen. Es folgt der Vergleich mit den Lesarten von <maison>. Wie bei der Auswertung von *maison:1* gibt die Zahl vor dem Substantiv die absolute, die nachfolgende Zahl die relative Häufigkeit auf alle *Graphien* im Kontext der Lesart an. Die Zahl in Klammern zeigt den Anteil an den Substantiven im Kontext.

habitation:1	maison:1	maison:2	maison:3	maison:4
0,40% (1,94%) chambre	0,10%		0,04%	
0,40% (1,94%) rue	0,11%		0,13%	
0,40% (1,94%) mur	0,04%			
0,30% (1,46%) fenêtre	0,05%		0,01%	
0,30% (1,46%) pièce	0,09%		0,01%	
0,30% (1,46%) ordure			0,01%	
0,30% (1,46%) lit			0,08%	
0,20% (0,97%) soir	0,08%		0,21%	
0,20% (0,97%) quartier	0,04%			
0,20% (0,97%) nuit	0,09%		0,09%	
0,20% (0,97%) mère	0,10%		0,20%	0,27%
0,20% (0,97%) siècle				
0,20% (0,97%) porte	0,14%		0,09%	0,09%
0,20% (0,97%) port				
0,20% (0,97%) population				
0,20% (0,97%) plaine				
0,20% (0,97%) ouverture				

```
0,20% (0,97%) maison
0,20% (0,97%) guerre                              0,01%     0,18%
0,20% (0,97%) gare
0,20% (0,97%) demeure
0,20% (0,97%) chinois
0,20% (0,97%) chansons
```

Bei einem Vergleich der obigen Tabelle mit der Auswertung der Substantive im Satzkontext von *maison:1* bestätigt sich, daß Substantive wie <mère> und <soir> wegen des häufigeren Auftauchens in anderen Kontexten nicht charakteristisch für diese Lesart sind. Außerdem sollte bei den Substantiven, bei denen die Werte für eine Signifikanz sprechen (<chambre>, <rue>, <mur>, <fenêtre>, <pièce> und <quartier>) nur <rue> (wegen des höheren Wertes bei der Lesart von <maison:3>) herausfallen.

Ergebnis für die erste Lesart

Damit ergibt sich die folgende Liste: Die Zahlen geben das Verhältnis des Auftauchens des Substantivs in der entsprechenden Lesart oder im Gesamtkorpus relativ zum Auftauchen in der ersten Lesart an (in diesem Fall < 1, da in den anderen Lesarten das Substantiv weniger häufig vorkommt). Das Substantiv <bois> kommt im Satzkontext der ersten Lesart jeweils etwa viermal so häufig vor wie bei der dritten Lesart oder im Korpus insgesamt (Kehrwert von $0,24$). Das Substantiv <pièce> kommt als Kollokation von <habitation> sogar 33mal häufiger vor als bei der dritten Lesart.[12]

	maison:2	maison:3	maison:4	im Korpus
bois (m)		0,24		0,24
chambre (m)		0,44		0,30
chambre (h)		0,10		0,07
étage (m)		0,37		0,15
fenêtre (m)		0,20		0,37
fenêtre (h)		0,03		0,07
odeur (m)		0,24		0,37
pièce (m)				0,32
pièce (h)		0,03		0,07
quartier (m)		0,12		0,25
quartier (h)		0,05		0,05
salle (m)		0,28		0,45
mur (m)				0,59
mur (h)				0,06

[12] Die Angaben (m) und (h) beziehen sich auf das jeweilige Element des *Synsets*.

Damit können die vorliegenden Substantive auf der Basis dieses Korpus als charakteristisch für die Lesart [*maison:1*, *habitation:1*] angesehen werden.

8.4.3 Auswertung für [*maison:2*, *signe:1*]

maison:2

Die Auswertung ist wegen der sehr geringen Anzahl der Belegstellen schwierig. Die absolute Frequenz ist für die einzelnen Substantive gegenüber der absoluten Anzahl bei den anderen Lesarten gering.

bei 2	bei 1	bei 3	bei 4
2 mars	2		
1 soleil	16	1	
1 ciel	10		
1 mot	9		
1 suite	1	1	
1 explication	1	1	
1 milieu	1		
1 ascendant	1		

Die Gegenüberstellung der absoluten Frequenz wird relativiert, wenn die absolute Frequenz dem relativen Auftauchen gegenüber gestellt wird. Das 16fache Auftauchen von <soleil> in der ersten Lesart entspricht nur 1,15%.

signe:1

Bei der Untersuchung des Korpus nach Kollokationen von <signe>[13] (insgesamt sechs Belege mit insgesamt 178 *Graphien* im Subkorpus) zeigt sich, daß nur für die erste Lesart zwei Substantive mehr als einmal vorkommen (<zodiaque>, das als Element der lexikalischen Einheit <signe du zodiaque> die Lesart eindeutig macht und <jour>, das aber schon bei anderen Lesarten sehr häufig auftauchte). Bei den nur einmal belegten Substantiven finden sich vier, die auch schon bei der zweiten Lesart von <maison> auftauchten: <soleil>, <mars>,

[13] Es werden insgesamt acht Lesarten von <signe> unterschieden: [*signe:1*, *maison:2*] (<„one of 12 equal areas into which the zodiac is divided“>), [*signe:2*, *symbole:2*, *marque:4*] (<„a distinguishing symbol “>), [*signe:3*] (<„a signal for attracting attention“>), [*signe:4*, *signal:1*] (<„any communication that encodes a message“>), [*signe:5*] (<„a character indicating a relation between quantities“>), [*signe:6*] (<„gesture“>), [*signe:7*, *symbole:1*] (<„an individual instance of a type of symbol“>) und [*signe:8*, *geste:3*] (<„the act of signaling by a movement of the hand“>).

<ciel> und <ascendant>. Unter den übrigen finden sich noch zwei weitere Substantive, die wegen der eindeutigen Zugehörigkeit zur Lesart <signe:1> auch im folgenden betrachtet werden: <astrologie> und <astre>. Ein zweifaches Auftauchen bedeutet ein Auftauchen in 1,12% der Fälle, ein einfaches Auftauchen lediglich 0,56%.

2 zodiaque	1 merde	1 imaginaire	1 attraction
2 jour	1 mémoire	1 harpe	1 astrologie
1 zone	1 mars	1 écharpe	1 astre
1 vie	1 mandarine	1 corps	1 ascendant
1 soleil	1 maison	1 ciel	1 absence
1 semaine	1 lueur	1 chant	
1 religion	1 lignes	1 boulette	
1 présence	1 influence	1 bélier	

Ergebnis für die zweite Lesart

Damit ergibt sich folgende Tabelle, in der wegen fehlender Zahlen der Vergleich mit der vierten Lesart weggelassen wird, da dort keine Belege vorhanden sind. Die Zahlen in Klammern geben wieder das Verhältnis der Relationen an (wegen der geringen Werte nicht als Prozentzahlen):

	bei 1	bei 3	im Korpus
3,28% mars (m)	0,0046% (0,0014)		0,00212% (0,00065)
1,12% mars (s)	0,0046% (0,0041)		0,00212% (0,00190)
1,64% soleil (m)	0,0364% (0,0222)	0,0110% (0,0670)	0,02740% (0,01671)
0,56% soleil (s)	0,0364% (0,0650)		0,02740% (0,04893)
1,64% ciel (m)	0,0228% (0,0139)		0,02462% (0,01501)
0,56% ciel (s)	0,0228% (0,0407)		0,02462% (0,04397)
1,64% mot	0,0205% (0,0125)		0,05229% (0,03188)
1,64% suite	0,0023% (0,0014)	0,0110% (0,0670)	0,02076% (0,01266)
1,64% explication	0,0023% (0,0014)	0,0110% (0,0670)	0,00368% (0,00225)
1,64% milieu	0,0023% (0,0014)		0,02161% (0,01318)
0,56% astrologie (s)			0,00008% (0,00014)
0,56% astre (s)			0,00085% (0,00002)
1,64% ascendant (m)	0,0023% (0,0014)		0,00064% (0,00039)
0,56% ascendant (s)	0,0023% (0,0041)		0,00064% (0,00114)

Die Unterschiede zwischen den Häufigkeiten sind sehr groß. Der höchste Wert wird bei 0,0670 erreicht, was bedeutet, daß etwa 15mal so häufig (Kehrwert von 0,0670) die Lesart <signe:1> in dieser Kollokation auftaucht. Die anderen Werte sind im Durchschnitt viel geringer. Als manueller Eingriff werden hier

<explication> und <mot> herausgenommen, da in *FWN* sehr viele Lesarten unterschieden werden und keine sich eindeutig der zweiten Lesart von <maison> zuordnen läßt. Bei den anderen ist gut zu erkennen, daß die Kollokation von <maison> oder <habitation> mit einem Substantiv dieser Liste ein sehr starker Hinweis auf diese Lesart ist.

8.4.4 Auswertung für [*maison:3*, *chez-soi:2*]

maison:3

Unter der Berücksichtigung der Pluralformen läßt sich mit manuellem Aufwand die folgende Liste der häufigsten Substantive erstellen. Ein mindestens vierfaches Auftreten entspricht einem Auftreten in 1% der Fälle (insgesamt 360 Belegstellen). Die erste Zahl ist die absolute Beleganzahl des folgenden Substantivs, die Prozentzahl gibt den Anteil an den *Graphien* im Kontext an, die Zahl in Klammern entspricht dem Anteil an den Substantiven im Kontext.

26	père	0,29%	(1,94%)	7	famille	0,08%	(0,52%)	5	mort	0,05%	(0,37%)
26	enfant	0,29%	(1,94%)	7	lits	0,08%	(0,52%)	5	main	0,05%	(0,37%)
20	jour	0,22%	(1,49%)	7	temps	0,08%	(0,52%)	5	fois	0,05%	(0,37%)
19	soir	0,21%	(1,42%)	6	vacances	0,07%	(0,45%)	4	copains	0,04%	(0,30%)
18	mère	0,20%	(1,34%)	6	voiture	0,07%	(0,45%)	4	papa	0,04%	(0,30%)
15	heure	0,16%	(1,12%)	6	garçons	0,07%	(0,45%)	4	téléphone	0,04%	(0,30%)
15	vie	0,16%	(1,12%)	6	place	0,07%	(0,45%)	4	chemin	0,04%	(0,30%)
12	rue	0,13%	(0,89%)	6	tête	0,07%	(0,45%)	4	travail	0,04%	(0,30%)
11	femme	0,12%	(0,82%)	6	homme	0,07%	(0,45%)	4	silence	0,04%	(0,30%)
10	parent	0,11%	(0,75%)	5	dimanche	0,05%	(0,37%)	4	paris	0,04%	(0,30%)
10	frère	0,11%	(0,75%)	5	mari	0,05%	(0,37%)	4	table	0,04%	(0,30%)
9	années	0,10%	(0,67%)	5	bureau	0,05%	(0,37%)	4	reste	0,04%	(0,30%)
8	porte	0,09%	(0,60%)	5	envie	0,05%	(0,37%)	4	chambre	0,04%	(0,30%)
8	fille	0,09%	(0,60%)	5	façon	0,05%	(0,37%)	4	matin	0,04%	(0,30%)
8	nuit	0,09%	(0,60%)	5	mois	0,05%	(0,37%)	4	mot	0,04%	(0,30%)
7	classe	0,08%	(0,52%)	5	livre	0,05%	(0,37%)	4	ans	0,04%	(0,30%)
7	semaine	0,08%	(0,52%)	5	voix	0,05%	(0,37%)	4	monde	0,04%	(0,30%)
7	maman	0,08%	(0,52%)	5	moment	0,05%	(0,37%)				

Der Vergleich des absoluten Auftauchens der Elemente in den anderen Kontexten ergibt die folgende Tabelle. Die Prozentzahl, die direkt vor dem Substantiv steht, gibt den Anteil dieses Substantivs am gesamten Belegkontext an. Die drei folgenden Spalten geben zum Vergleich den Anteil bei den anderen Lesartenkontexten an. In der sechsten Spalte findet sich die Auswertung des jeweiligen Substantivs

im gesamten Untersuchungskorpus (Anzahl des Substantivs im Gesamtkorpus im Verhältnis zu allen *Graphien*). Die letzte Spalte gibt die Abweichung in Prozent der ersten zur sechsten Spalte an.

bei 3		bei 1	bei 2	bei 4	im Korpus	Korpus/Lesart:3
0,07%	vacances	0,04%			0,01%	11,12%
0,04%	copains	0,002%			0,01%	13,52%
0,05%	dimanche	0,01%			0,01%	18,16%
0,08%	classe	0,01%			0,02%	20,16%
0,29%	père	0,08%			0,06%	21,18%
0,11%	parent	0,06%			0,02%	21,34%
0,08%	semaine	0,01%		0,18%	0,02%	21,42%
0,21%	soir	0,07%			0,05%	21,89%
0,05%	mari	0,05%		0,09%	0,01%	22,91%
0,08%	maman	0,03%		0,09%	0,02%	24,17%
0,04%	papa	0,02%		0,27%	0,01%	24,40%
0,05%	bureau	0,01%		0,18%	0,01%	25,33%
0,04%	téléphone	0,01%		0,09%	0,01%	25,72%
0,07%	voiture	0,05%			0,02%	33,18%
0,07%	garçons	0,04%			0,02%	35,41%
0,29%	enfant	0,13%			0,10%	36,34%
0,08%	famille	0,11%		0,09%	0,03%	37,67%
0,13%	rue	0,09%			0,05%	38,96%
0,10%	années	0,05%			0,04%	39,48%
0,08%	lits	0,03%			0,03%	39,74%
0,05%	envie	0,01%		0,09%	0,02%	40,10%
0,20%	mère	0,10%		0,27%	0,08%	40,94%
0,16%	heure	0,002%			0,07%	42,42%
0,04%	chemin	0,03%			0,02%	43,43%
0,05%	façon	0,02%		0,09%	0,02%	43,71%
0,04%	travail	0,02%		0,09%	0,02%	46,17%
0,05%	mois	0,03%			0,03%	46,53%
0,22%	jour	0,14%		0,09%	0,11%	50,88%
0,05%	livre	0,02%			0,03%	52,99%
0,16%	vie	0,07%		0,09%	0,09%	53,18%
0,04%	silence	0,04%			0,03%	58,98%
0,07%	place	0,03%		0,09%	0,04%	59,50%
0,04%	paris	0,03%		0,35%	0,03%	63,71%
0,04%	table	0,04%		0,09%	0,03%	63,85%
0,04%	reste	0,01%		0,09%	0,03%	64,10%
0,09%	porte	0,12%			0,06%	65,52%
0,04%	chambre	0,10%			0,03%	67,87%
0,09%	fille	0,05%		0,09%	0,06%	68,41%
0,12%	femme	0,09%			0,08%	68,80%
0,09%	nuit	0,08%			0,06%	70,22%

0,04%	matin	0,03%		0,18%	0,03%	71,14%
0,05%	voix	0,03%			0,04%	79,65%
0,05%	moment	0,03%			0,05%	98,86%
0,05%	mort	0,04%			0,06%	113,69%
0,07%	tête	0,04%			0,08%	114,47%
0,04%	mot	0,04%	5,56%	0,09%	0,05%	119,13%
0,04%	ans	0,05%			0,06%	145,60%
0,07%	homme	0,11%		0,09%	0,10%	157,04%
0,08%	temps	0,06%		0,09%	0,12%	157,21%
0,04%	monde	0,03%			0,07%	169,29%
0,05%	main	0,06%		0,09%	0,09%	169,61%
0,05%	fois	0,06%		0,09%	0,09%	171,83%

chez-soi:2

Das Substantiv <chez-soi> wird bei *FWN* mit drei verschiedenen Lesarten geführt: [*chez-soi:*, *domicile:1*, *maison d'habitation:1*] (<„a physical structure (e.g., a house) that someone ist living in“>), [*chez-soi:2*, *maison:3*] (<„where you live“>) und [*chez-soi:3*, *logement:3*] (<„the act of lodging“>), die sich allerdings in einem französischen Wörterbuch (vgl. z.B. Legrain und Garnier 2000 und Rey und Rey-Debove 1993) so nicht wiederfinden. Am ehesten entspricht dem französischen Sprachgebrauch die Beziehung zu <maison:3>. Das Konzept der englischen Glosse, „the act of lodging“, ist im französischen Substantiv nicht enthalten, die erste hier beschriebene Lesart ist nur sehr schwer von der zweiten abzugrenzen. In dem betrachteten Subkorpus von *Frantext* finden sich insgesamt nur vier Belege, in dessen Kontext auch nur ein Substantiv zweimal (im selben Beleg) vorkommt (insgesamt 113 *Graphien*):

2 racine	1 rue	1 logement	1 école
1 travail	1 retour	1 jardin	1 dieu
1 traditions	1 prolongement	1 espèce	1 deuxième
1 terre	1 mémoire	1 espace	1 asile
1 secteur	1 loisirs	1 enfants	1 année

Hier entspricht dem einmaligen Auftauchen ein Prozentwert von 0,88%. Beim Vergleich dieser Liste mit der Liste von <maison:3> sind nur vier Substantive in beiden Listen vorhanden: <enfant>, <travail>, <rue> und <années>. Zu <rue> ist schon in der Auswertung zur ersten Lesart von <maison> ausgeführt worden, daß die Werte bei den einzelnen Lesarten zu nahe beieinander liegen. Bei <travail> ist die Frequenz in der vierten Lesart von <maison> sehr viel

höher. Zu <année> wurde in Abschnitt 8.4.2 gesagt, daß Zeitangaben insgesamt aus den Listen gestrichen werden. Daher wird hier nur das Substantiv <enfant> übernommen.

Ergebnis für die dritte Lesart

Intuitiv wären z.B. für diese Lesart die Substantive <famille>, <mère>,[14] <parent> oder <garçon> oder <maman> als charakteristisch angenommen worden. Allerdings zeigen die Werte aus der obigen Tabelle, daß die relative Häufigkeit in den Lesarten zu nahe beieinander liegt, um wirklich für eine der Lesarten aussagekräftig zu sein. Auch kommt das Substantiv <papa> in der vierten Lesart viel häufiger vor. Nach den Auswertungen analog zu den anderen Lesarten ergibt sich:

	bei 1	bei 2	bei 4	im Korpus
copain (m)	0,05			0,14
classe (m)	0,12			0,20
père (m)	0,30			0,21
enfant (m)	0,44			0,36
enfant (c)	0,44			0,09
lit (m)	0,41			0,40

Die erste Zeile sagt aus, daß die dritte Lesart von <maison> in der Kollokation mit <copain> zwanzigfach häufiger auftaucht (Kehrwert von 0.05) und insgesamt auf das gesamte Korpus gerechnet sechsfach. Bei dem Substantiv <classe> taucht die dritte Lesart etwa achtmal häufiger auf, auf das gesamte Korpus gerechtet fünffach.

8.4.5 Auswertung für [*maison:4*, *firme:1*]

maison:4

Auch hier veränderte sich die Liste der häufigsten Substantive durch die Zusammenführung von flektierten Formen. Herausgenommen wurden folgende Formen: Das Substantiv <disque> ist sehr frequent (als Teil von <maison de disques>), hier war die Abgrenzung zur Kategorie der Komposita nicht eindeutig. Das

[14] Dieses Substantiv fällt aus der Liste heraus, da es häufig in der Kombination <maison mère> in der vierten Lesart vorkommt.

Substantiv <mère> ist im Satzkontext dreifach vertreten, da in *Frantext* das Kompositum <maison mère>, das zu dieser Lesart gezählt wird, getrennt annotiert wird: `[maison S] [mère S]`. Die falsch getaggten Formen <a> und <paris> werden aus der Liste gestrichen. Die erste Zahl in der folgenden Tabelle gibt jeweils die absolute Anzahl an, die Zahl in Klammern den prozentualen Anteil an allen Substantiven des Satzkontexts:

```
3 production 0,27% (1,15%)  2 nom    0,18% (0,77%)  2 correction 0,18% (0,77%)
3 patron     0,27% (1,15%)  2 matin  0,18% (0,77%)  2 corps      0,18% (0,77%)
3 papa       0,27% (1,15%)  2 maison 0,18% (0,77%)  2 contrat    0,18% (0,77%)
3 mère       0,27% (1,15%)  2 guerre 0,18% (0,77%)  2 bureau     0,18% (0,77%)
3 directeur  0,27% (1,15%)  2 esprit 0,18% (0,77%)  2 arrêt      0,18% (0,77%)
2 semaine    0,18% (0,77%)  2 duc    0,18% (0,77%)
```

Zu <correction> und <arrêt> ist noch zu ergänzen, daß sie als Elemente der Komposita <maison de correction> und <maison d'arrêt> auftauchen, daher nicht als eigenständige Substantivkollokationen betrachtet werden können. Auch hier wird wieder ein Vergleich mit dem Auftauchen des Substantivs in den anderen Lesarten durchgeführt:

bei 4		bei 1	bei 2	bei 3	im Korpus	Korpus/Lesart:4
0,177%	correction	0,002%			0,001%	0,422%
0,266%	production	0,002%			0,001%	0,540%
0,177%	contrat				0,002%	0,889%
0,177%	duc	0,009%			0,002%	1,355%
0,266%	patron	0,005%		0,089%	0,005%	2,012%
0,266%	directeur	0,002%			0,005%	2,065%
0,177%	arrêt	0,002%			0,005%	2,727%
0,266%	papa	0,016%		0,355%	0,011%	4,024%
0,177%	bureau	0,014%		0,444%	0,014%	7,831%
0,177%	semaine	0,011%		0,621%	0,016%	9,273%
0,177%	esprit	0,014%		0,089%	0,021%	11,869%
0,177%	guerre	0,036%		0,266%	0,030%	16,981%
0,177%	matin	0,027%		0,177%	0,031%	17,597%
0,177%	nom	0,048%		0,089%	0,040%	22,506%
0,177%	corps	0,027%		0,089%	0,048%	26,923%
0,266%	mère	0,102%		1,597%	0,081%	30,378%

Die Zahlen in der letzten Spalte ergeben gegenüber der relativen Frequenz im Korpus gute Werte. Durch die Unterschiede bei den Werten der einzelnen Lesarten fallen einige Substantive weg: <mère>, <papa>, <bureau> und <semaine>, die bei der dritten Lesart auffälliger sind. Auch sind manche Werte zu

nahe beieinander, um aussagekräftige Ergebnisse zu bekommen. Das Substantiv <bureau> wäre intuitiv wohl eher als charakteristisch für diese Lesart ausgewählt worden. Die Werte bei der dritten Lesart zeigen aber, daß es im Kontext dieser Lesart ebenfalls häufig auftaucht und damit nicht sehr aussagekräftig für die vierte Lesart ist.

firme:1

Das Substantiv *firme* wird in *FWN* (wie im *Petit Robert*) nur unter einer Lesart geführt. Im Subkorpus von *Frantext* lassen sich 14 Belege finden, deren Satzkontext betrachtet wird (insgesamt 449 *Graphien* mit 88 Substantiven). Es finden sich nur wenige, die mehr als einmal vorkommen (<contrat> wurde übernommen, da es im Satzkontext von <maison:4> ebenfalls vorkommt). In der folgenden Tabelle wird das absolute und relative Auftauchen im Satzkontext von <firme> verglichen mit dem prozentualen Auftauchen im Korpus. Die letzte Spalte vergleicht die beiden Prozentzahlen:

	auf 449 Graphien	im Korpus	Verhältnis
3 vaccin	0,668%	0,00094	0,0014
2 sérum	0,445%	0,00011	0,0002
2 risques	0,445%	0,00781	0,0175
2 rachat	0,445%	0,00016	0,0004
2 producteur	0,445%	0,00086	0,0019
2 lutte	0,445%	0,00308	0,0069
2 journaux	0,445%	0,01348	0,0303
2 échelle	0,445%	0,00273	0,0061
2 compétition	0,445%	0,00083	0,0019
2 bill	0,445%	0,00328	0,0074
1 contrats	0,223%	0,00158	0,0071

Wegen der Unausgewogenheit des Korpus muß hier eingegriffen werden, da das Substantiv <vaccin> zwar hier häufig auftaucht, aber allein in einem Werk von Guibert 51mal (im Korpus insgesamt nur 60mal). Für das Substantiv <sérum> gelten ähnliche Zahlen. Daher werden hier die Ergebnisse zu stark verzerrt. Die anderen Substantive kommen jeweils zweimal im gleichen Satzkontext vor und das „Substantiv“ <bill> ist eigentlich ein Eigenname, der von *Frantext* nicht erkannt wurde. Damit bleibt von dieser Liste nur <contrat>, das ja schon in der Substantivliste von <maison:4> zu finden ist. Diese Fehler, die manuell korrigiert werden müssen, können durch ein ausgewogenes großes Korpus vermieden werden.

Ergebnis für die vierte Lesart

Damit ergibt sich die folgende Liste, in der wieder die Verhältnisse der Prozentzahlen der obigen Tabelle angegeben sind.

	bei 1	bei 2	bei 3	im Korpus
correction (m)	0,011			0,004
production (m)	0,008			0,005
contrat (m)				0,009
contrat (f)				0,0071
duc (m)	0,051			0,014
patron (m)	0,019		0,335	0,020
directeur (m)	0,008			0,021

8.5 Substantivlisten

Hier werden noch einmal als Ergebnis die vier Listen zusammengestellt. Diese Listen geben die auf der Grundlage von *Frantext* erstellten aussagekräftigen Kollokationen für die einzelnen *Synsets*, in denen das Substantiv *maison* enthalten ist, an. Diese Listen spiegeln daher die Daten des Korpus wieder, das durch seine Beschränkung hier auf die Kategorien *romans* und *essais* nicht repräsentativ ist. Bei einem ausgewogenen Korpus als Grundlage werden andere Listen entstehen.

- Als neue Verbindungen für die erste Lesart bieten sich folgende Substantive an: <étage>, <bois>, <chambre>, <fenêtre>, <mur>, <pièce>, <quartier> und <salle>.

- Die Substantive <mars>, <soleil>, <ciel>, <suite>, <milieu>, <astre>, <astrologie> und <ascendant> sind als Kollokation für die zweite Lesart ein guter Hinweis.

- Die Liste für die dritte Lesart enthält <copains>, <classe>, <père>, <enfant> und <lit>.

- Für die vierte Lesart haben sich <correction>, <production>, <contrat>, <duc>, <patron> und <directeur> als aussagekräftig herausgestellt.

8.6 Disambiguierung der Substantive

Die neuen Knoten werden anhand der Korpusbelege der jeweiligen Kollokation manuell disambiguiert (dieser manuelle Eingriff ist nur bei einem *FWN*-Lesarten annotierten Korpus zu vermeiden).

8.6.1 [*maison:1*, *habitation:1*]

Bei <étage> wird in den untersuchten Kollokationen nur die erste Lesart verwendet (<a room or set of rooms comprising a single level of a multi-level building; „what level is the office on?“>). Die zweite Lesart von <bois> im *Synset* [*bois:2*, *forêt:2*] wird in den Kollokationen verwendet und die vierte Lesart von <chambre> (<a room used primarily for sleeping>). Die einzige Lesart für <fenêtre> wird verwendet (<an opening in the wall of a building to admit light and air; usually closed by a casement containing transparent material and capable of being opened>). Für *pièce* wird die achte Lesart [*pièce:8*, *salle:4*] (<an area within a building enclosed by walls and floor and ceiling; „the rooms were very small but they had a nice view“>) ausgewählt. Damit erübrigt sich auch die Wahl für <salle>, da in den untersuchten Kollokationen Komposita wie <salle de séjour> verwendet werden, die sich als Unterknoten von [*salle:4*, *pièce:8*] im Netz befinden.

Die einzige Lesart von <quartier> (<a district of a town or city>) wird ausgewählt und die zweite Lesart von <mur>: (<a partition with a height and length greater than its thickness; used to divide or enclose or support>).

Das Substantiv <odeur> wird in den Korpusbelegen nicht durchgängig in einer eindeutigen Lesart des *FWN* verwendet, daher wird diese Kollokation nicht weiter verwertet.

8.6.2 [*maison:2*, *signe:1*]

Da die beiden Substantive <mars> und <soleil> neu in das Netz eingefügt wurden (zu <soleil> vgl. Abschnitt auf S. 103), ist die Zuweisung der Lesart hier klar: <mars:1> und <soleil:2>. Bei der Untersuchung der Kollokation <ciel> ist die Lesart <ciel:2> als Synonym zu <firmament:1> mit der (engl.) Glosse <the apparent sufrace of the imaginary sphere on which celestial bodies appear to be projected> zuzuweisen. Für <suite> wird die vierte Lesart gewählt: im *Synset*

sind <chapelet:1> und <rame:3> mit <a sequentially ordered set of things or events or ideas in which each successive member is related to the preceding: „a string of islands“; „train of mourners“; „a train of thought“>.

Die zweite Lesart mit der Glosse <natural objects visible in the sky> wird für den Knoten <astre> genommen, die einzige in *FWN* vorhandene Lesart für <astrologie>, die fünfte Lesart für <milieu> (<a point equidistant from the ends of a line or the extremities of a figure>), der Knoten <ascendant> wird als zweite Lesart neu eingehängt (siehe Abschnitt 9.6)

8.6.3 [*maison:3*, *chez-soi:2*]

Der Eintrag für <copain> im französischen Netz ist zu hinterfragen, da er zwei Lesarten gegenüberstellt: [*copain:1*, *camarade:2*, *ami:2*] <informal term (Australian or British) for a friend of the same sex> und [*copain:2*, *camarade:1*, *ami:4*] <a close friend who accompanies his buddies in their activities>. Da diese beiden Konzeptknoten unter demselben Mutterknoten ([*ami:1*]) eingehängt sind, ist hier wieder eine Unterscheidung übernommen worden, die sich im Englischen vornehmen läßt, aber im französischen Netz (besonders für die Verwendung „ma copine“ und „mon copain“ als Bezeichnung für den Partner) der Sprache nicht Rechnung trägt. Es wird daher keine der beiden Lesarten ausgewählt.

Die ersten beiden Lesarten von <classe> sind als direkte Unterknoten von [*attroupement:1*, *rassemblement:1*] eingehängt: <a body of students who graduate together: „the class of ‘97 “; „she was in my year at Hoehandle High“> und <a group of persons together in one place>. Ihnen entspricht die Verwendung in der Kollokationen mit <maison:3>. Die erste Lesart (<a male parent (also used as a term of address to your father); „his father was born in Atlanta“> wird dem neuen Knoten <père> zugewiesen und ebenfalls die erste Lesart für <enfant>: <a young person of either sex (between infancy and youth); „she writes books for children“; „they’re just kids“>. Für <lit> kommt in den Belegen nur die zweite Lesart in Frage: <a piece of furniture that provides a place to sleep; „he sat on the edge of the bed“; „the room had only a bed and chair“>.

8.6.4 [*maison:4*, *firme:1*]

Zuletzt erfolgt die Lesartenzuweisung für das vierte *Synset*. Für <correction> existiert nur ein Eintrag in *FWN*: <the act of offering an improvement to replace a mistake>, daher wird mit dem Knoten [*correction:1*] verknüpft. Für <production> ist die zweite oder dritte Lesart möglich: [*production:1*, *produit:1*] <an artifact that has been produced by someone or some process> und dessen Unterknoten [*production:3*, *rendement:3*] <things produced>. Die zweite Lesart ist in den Kollokationen die wahrscheinlichere.

Die beiden möglichen Lesarten für <contrat> sind als Schwesternknoten unter [*convention par écrit:1*] eingehängt: <a written agreement between two states or sovereigns> und <a legal document summarizing the agreement between parties>. In den Kollokationen wird die zweite Lesart verwendet, d.h. dieses Konzept wird mit [*maison:4*, *firme:1*] verknüpft. Der neue Knoten [*duc:1*] (nicht im Originalnetz vorhanden) wird als Lesart genommen und die neunte Lesart für <directeur> (ohne Glosse, mit *Synset* [*directeur:9*, *chef d'entreprise:1*]).

Die drei in Frage kommenden Lesarten für <patron> sind im Netz schon sehr eng miteinander verknüpft: Als Schwesternknoten sind <patron:3> (<a person who exercises control and makes decisions; „he is his own boss now“>) und <patron:8> (im *Synset* [*patron:8*, *employeur:2*] <a person or firm that employs workers>) unter dem Knoten [*chef:6*, *guide:5*, *leader:1*] eingehängt. Die zweite Lesart [*patron:2*, *employeur:1*] ist als Tochterknoten unter der dritten Lesart eingeknüpft. In den Kollokationen ist die zweite Lesart bevorzugt zuzuweisen.

Bei <arrêt> konnte in den Belegen keine eindeutige Zuweisung zu den sieben in *FWN* vorhandenen Lesarten vorgenommen werden, daher fällt dieser Knoten als Erweiterung aus.

8.7 Rangfolge

Die Substantivliste, die im vorangegangenen Abschnitt als charakteristisch für die erste Lesart zusammengestellt wurde, kann durch einen Vergleich der Werte in eine Rangreihenfolge gebracht werden. Somit kann unterschieden werden zwischen „sehr hoher Aussagekraft“ und „hoher Aussagekraft“.[15] Für jede einzelne Lesart

[15] Es können bei einer breiten und ausgewogenen Grundlage auch mehrere Kategoriengewichtungen eingeführt werden.

muß auf der Basis der erstellten Listen diese Zweiteilung vorgenommen werden. Wegen der unterschiedlichen Werte kann kein einzelner Schwellenwert benutzt werden, daher hat diese Einteilung wieder manuell zu erfolgen.

8.7.1 [*maison:1*, *habitation:1*]

Da der Knoten an das *Synset* [*maison:1*, *habitation:1*] angehängt wird, ist der Vergleich mit den anderen Lesarten ausschlaggebend. Die in dieser Liste vorhandenen Substantive kommen nur in der dritten Lesart vor. Das Substantiv <chambre> kommt in der Kollokation mit <maison> bei der dritten Lesart weniger als halb so oft (0,44mal) vor, bei der Kollokation mit <habitation> 10mal weniger häufig.

Ein großer Sprung ist bei den Werten der letzten drei Substantive in der Liste zu sehen. Dort taucht die Kollokation mit der Lesart aus diesem *Synset* mindestens achtmal so häufig auf (Kehrwert von 0,12). Bei <pièce> ist dies über 33mal der Fall und da <mur> in keiner der anderen Lesarten auftaucht und bei der ersten Lesart 1,7mal so häufig auftaucht, ist der Hinweis auf dieses *Synset* sehr stark. Daher läßt sich hier der Schnitt für die zweigewichtige Unterscheidung ziehen:

	maison:3	im Korpus
chambre (m/h)	0,44/0,10	0,30/0,07
étage (m)	0,37	0,15
salle (m)	0,28	0,45
bois (m)	0,24	0,24
fenêtre (m/h)	0,20/0,03	0,37/0,07
quartier (m/h)	0,12/0,05	0,25/0,05
pièce (m/h)	0/0,03	0,32/0,07
mur (m/h)		0,59/0,06

Damit werden in die Datenbank die Konzeptknoten mit <chambre:1>, <étage:1>, <salle:4>, <bois:2> und <fenêtre:1> mit der Gewichtung „hohe Aussagekraft“ und die Substantive <quartier:1>, <pièce:8> und <mur:2> mit der Gewichtung „sehr hohe Aussagekraft“ eingebunden.

Anschaulich kann das in der Abbildung 8.1 deutlich gemacht werden. Einfache Pfeile zeigen die ursprüngliche Netzstruktur, dicke Pfeile stellen die neuen Kanten

der Kategorie „sehr hohe Aussagekraft“ dar, mitteldicke die Kanten der Kategorie „hohe Aussagekraft“.

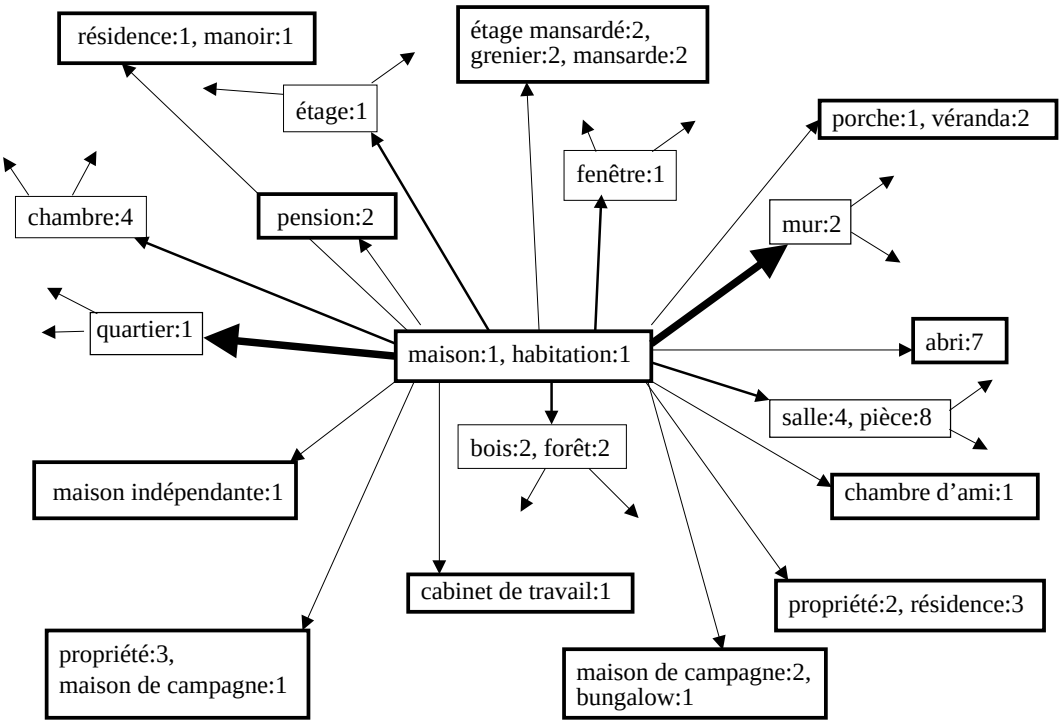


Abbildung 8.1: Einbindung von Korpusdaten ins Netz

8.7.2 [*maison:2, signe:1*]

Bei der Liste läßt sich bei dem Vergleich der Werte für die dritte Lesart und dem abschließenden Vergleich mit den Werten für die erste Lesart feststellen, daß sich <mars>, <milieu>, <ascendant>, <astrologie> und <astre> abheben, da sie in der einen Lesart nicht auftauchen und besonders geringe Werte in der anderen Lesart haben. Eine Hochrechnung (die allerdings durch die geringen Korpusdaten nicht gesichert ist) könnte für das Substantiv <mars> aussagen, daß es bei dieser Lesart über 714mal (Kehrwert von 0,0014) häufiger auftaucht als bei der ersten Lesart.

	bei 1	bei 3	im Korpus
soleil (m/s)	0,0222/0,0650	0,0670/0	0,01671/0,04893
suite (m)	0,0014	0,0670	0,01266
ciel (m/s)	0,0139/0,0407		0,01501/0,04397

mars (m/s)	0,0014/0,0041	0,00065/0,00190
milieu (m)	0,0014	0,01318
ascendant (m/s)	0,0014/0,0041	0,00039/0,00114
astrologie (s)		0,00014
astre (s)		0,00002

Damit lassen sich hier die Substantive <soleil>, <suite>, <explication>, <ciel> und <mot> in die Kategorie „hohe Aussagekraft“ und die Substantive <mars>, <milieu>, <ascendant>, <astrologie> und <astre> in die Kategorie „sehr hohe Aussagekraft“ einordnen.

8.7.3 [*maison:3*, *chez-soi:2*]

Hier ist eine unterschiedliche Gewichtung durch den Sprung zwischen den Substantiven <père> und <classe> gut zu sehen. Als Kollokation zu diesem *Synset* kommt <père> dreimal häufiger und das Substantiv <classe> mehr als achtmal häufiger vor als bei der Lesart [*maison:1*, *habitation:1*].

	bei 1	im Korpus
enfant (m/c)	0,44	0,36/0,09
lit (m)	0,41	0,40
père (m)	0,30	0,21
classe (m)	0,12	0,20

Für die Kategorie „sehr hohe Aussagekraft“ wird daher das Substantiv <classe> ausgesucht, für die Kategorie „hohe Aussagekraft“ <enfant>, <lit> und <père>.

8.7.4 [*maison:4*, *firme:1*]

Bei diesem *Synset* lassen sich die zwei Kategorien zuordnen durch das Auftauchen in zwei Lesarten und einem deutlich geringeren Wert beim ersten *Synset*:

	bei 1	bei 3	im Korpus
patron (m)	0,019	0,335	0,020
duc (m)	0,051		0,014
arrêt (m)	0,011		0,027
correction (m)	0,011		0,004
directeur (m)	0,008		0,021
production (m)	0,008		0,005
contrat (m/f)			0,009/0,007

Hier gehören daher <directeur>, <production> und <contrat> zur Kategorie „hohe Aussagekraft“ und <patron>, <duc>, <arrêt> und <correction> zur Kategorie „sehr hohe Aussagekraft“.

8.8 Ergebnis

Durch den manuellen Vergleich der Kollokationen für die Elemente jedes einzelnen *Synsets* hat sich für jede Lesart eine Liste von *Synsets* ergeben, die durch neue Verbindungen mit den jeweiligen Lesarten von <maison> verknüpft werden, um für ein Maß des Zusammenhangs einen weiteren Ausgangspunkt zu ergeben. Die Schwierigkeiten, die sich durch die fehlerhafte Annotation des Korpus *Frantext* ergaben, stellen allerdings – verbunden mit dem hohen manuellen Aufwand – für eine größere Studie und damit auch eine größere Erweiterung der Datenbank Probleme dar. Als Studie und Test können sie als erste Datengrundlage verwendet werden.

Mit der Verwendung eines größeren, gewichteten Korpus, das bessere Auswertungsmöglichkeiten bietet, können durch ein automatisiertes Verfahren einige der in diesem Abschnitt durchgeführten Untersuchungen schneller und für eine größere Datenmenge durchgeführt werden. Der manuelle Aufwand der Annotation wird sich in einem semantisch annotierten Korpus (idealerweise konform zur Lesartenunterscheidung von *FWN*, aber auch zu anderen mit der Möglichkeit der Zusammenführung der Lesarten) sehr reduzieren.

Allerdings existieren für die französische Sprache noch keine großen Korpora, die nach den Lesartenunterscheidungen von *FWN* semantisch annotiert sind.[16]

8.9 Ausbau des semantischen Netzes

Als nächster Schritt werden die in den vorangegangen Abschnitten ermittelten *Synsets* mit den Lesarten von <maison> in der Datenbank *FWN* verbunden. Die Abbildung 8.1 zeigt, wie die neuen Knoten mit gewichteten Kanten in das Netz bilateral eingeknüpft werden. Diese Knoten sind mit den Knoten der

[16] Für die englische Sprache gibt es Korpora die auszugsweise mit *WN*-Lesarten annotiert sind. Vgl. z.B. das *SemCor*-Projekt von Princeton auf der Basis des *English Brown Corpus*, in dem 200.000 der 700.000 *Graphien* mit ihrer jeweiligen *WN*-1.6-Lesart verknüpft sind oder die Annotationen eines Teils des *Wall Street Journal Treebank Corpus* von Wiebe et al. 1997.

jeweiligen Lesart identisch. Bei der Berechnung des Zusammenhangsmaßes wird allerdings beim Durchlaufen der neu erstellten Kante die Gewichtung, mit der diese Verknüpfung belegt ist, berücksichtigt.

Für die Einarbeitung in die Datenbank und die Auswertung muß das im speziellen Datenbankformat vorliegende Netz *FWN* in das Datenbankformat von *WN* umgewandelt werden (siehe Abschnitt 9.3). Danach müssen die zwei neuen Verknüpfungsarten „hohe Aussagekraft“ und „sehr hohe Aussagekraft“ zu den Verknüpfungen von *FWN* hinzugefügt werden und die in diesem Rahmen betrachteten vier *Synsets* von *maison* mit den neuen Knoten verknüpft werden.

Der genaue Vorgang des Einbindens wird im Anhang A.5.2 beschrieben und die vollständige Datei, die die neuen Knoten mit ihrem Ausgangslemma und der Gewichtung enthält, findet sich ebenfalls in diesem Abschnitt.

Kapitel 9

Programm und Datenbanken

Als Ausgangspunkt wird für die Berechnung der Zusammenhangsmaße ein schon existierendes Programm genommen: das von Pedersen entwickelte Perl-Programm `WordNet-Similarity`. Die semantischen Netze (*WN* und *FWN*) liegen als `ASCII`-Dateien vor.

9.1 `WordNet-Similarity`

Das Perl-Programm `WordNet-Similarity` wird in der Version 0.11 verwendet.[1] Dieses Programm greift auf die Datenbank von *WordNet 2.0* zu und berechnet die Abstandsmaße zweier Datenbankeinträge. Pakete, die von diesem Programm aufgerufen werden, stellen die verschiedenen Berechnungsmöglichkeiten von Zusammenhang zur Verfügung.[2]

- `WordNet::Similarity::edge`: Einfaches knotenzählendes Maß (invertiert) (vgl. 5.3)
- `WordNet::Similarity::hso`: Verknüpfungen durch Hyperonymie und Hyponymie (Hirst und St.Onge, vgl. 5.5)
- `WordNet::Similarity::lch`: Vergleich der Anzahl der Knoten auf dem Verbindungspfad mit der Tiefe der Taxonomie (Leacock und Chodorow, vgl. 5.6)

[1] Release vom 24.09.2004, `http://search.cpan.org/dist/WordNet-Similarity/` (13.01.2005). Ein Web-Interface von Jason Michelizzi und Ted Pedersen findet sich unter `http://www.d.umn.edu/~mich0212/cgi-bin/similarity/similarity.cgi` (13.01.2005).

[2] Siehe Kapitel 5.

- `WordNet::Similarity::lin`: Verrechnung der Gemeinsamkeiten und der trennenden Charakteristika der beiden *Synsets* (Lin, vgl. 5.10)

- `WordNet::Similarity::jcn`: Gewichtung durch den Informationsgehalt auf der Basis des einfachen knotenbasierten Ansatzes (Jiang und Conrath, vgl. 5.9)

- `WordNet::Similarity::res`: Informationsgehalt des gemeinsamen Ahnenknotens (Resnik, vgl. 5.8)

- `WordNet::Similarity::wup`: Verhältnis von Knotentiefe und Informationsgehalt (Wu und Palmer, vgl. 5.7)

- `WordNet::Similarity::random`: Ausgabe einer Zufallszahl (aus dem Intervall $[0, 1]$) als Vergleichswert zu den anderen „wirklichen“ Maßen

- `WordNet::Similarity::lesk`: Dieses Modul betrachtet die Glossen, die in *WordNet* den *Synsets* beigegeben werden. Da die englischen Glossen bei der Transformation der französischen Datenbank nicht übernommen wurden, wird dieses Modul im weiteren nicht berücksichtigt.

9.2 Neue Module

Zusätzlich wurde für ein einfaches Wahrscheinlichkeitsmaß ein Perl-Modul erstellt, das – wie in 5.2 beschrieben – die netzbasierte Wahrscheinlichkeit $p_N(k)$ (Verhältnis von Unterknoten von k im Vergleich zu der Anzahl der Knoten im gesamten Netz, vgl. 5.1.2) miteinbezieht (Perl-Modul `WordNet::Similarity::einfWahrsch`). Außerdem stellt ein neues Perl-Paket die Berechnung der Kantenanzahl auf dem Verbindungsweg zur Verfügung (vgl. 5.4, `WordNet::Similarity::edge2`). Die zugehörigen Quellcodes sind im Anhang unter A.3.1 und A.3.2 einzusehen.

Das Programm `similarity.pl` greift über das Paket `Query-Data` auf die Datenbank von *WordNet 2.0* zu. Die französischen Daten liegen aber im Format von *FWN* vor, das sich grundlegend unterscheidet. Daher wird zusätzlich zur Erweiterung der französischen Datenbank erstmalig die Transformation der *FWN*-Datenbank in das *WordNet*-Format vorgenommen.

9.3 Datenstruktur von *EWN*

In der Datenbankdatei `wn_fr.ewn` sind sämtliche Daten (22.745 Einträge, davon 17.826 Substantive und 4.919 Verben)[3] für den französischen Teil des *FWN* zusammengefaßt. Die Einträge der Datenbank sind aufsteigend numeriert und folgen außer der Trennung der Wortarten keiner bestimmten Reihenfolge. Ein beispielhafter Ausschnitt aus einem zusammengestellten Datenbankeintrag findet sich in Abbildung 9.1.

Es werden alle zu einem *Synset* gehörigen Informationen fortlaufend innerhalb der Hierarchie zusammengefaßt. Verweise werden durch die Angaben der Relationsart und des jeweiligen Lemma mit der Lesartennummer dargestellt. Eine interne Verwendung der durchlaufenden *Synset*-Nummer findet nicht statt.

Unter der Ebene `1 VARIANTS` werden die Elemente des *Synsets* mit ihrer jeweiligen Lesartennummer und (einmal pro *Synset*) der Angabe der externen Quelle angegeben. Unter der Angabe `1 INTERNAL_LINKS` werden die datenbankinternen Verknüpfungen aufgelistet: jeweils mit Angabe des angeknüpften *Synsets* (durch Angabe eines Elementes mit Lesartennummer) und evt. Ergänzungen.

Bei der lexikalischen Verknüpfung „antonym“ wird zusätzlich vermerkt, welches Element aus dem Ausgangssynset mit welchem Element aus dem Zielsynset verknüpft wird.

Die Verlinkung mit dem *ILI* wird durch die Angabe `1 EQ_LINKS` gegeben, das die *WN*-Offsets (Version 1.5) der *Synsets* enthält, mit denen das *EWN*-Synset verknüpft wird.

[3] Vgl. Catherin 1999, 3.

```
0 @4895@ WORD_MEANING                          @Nummer@ identifiziert Synset
   1 PART_OF_SPEECH ,,n“                       1. Ebene: POS-Information
   1 VARIANTS                                  1. Ebene: Synset-Mitglieder
      2 LITERAL ,,habitation“                   2. Ebene: Element des Synsets
         3 SENSE 1                               3. Ebene: dessen Lesartnummer
         3 EXTERNAL_INFO                         3. Ebene: Angaben zu Korpus
            4 SOURCE_ID 1
               5 TEXT_KEY ,,2728393-n“
      2 LITERAL ,,maison“                       2. Ebene: Element des Synsets
         3 SENSE 1                               3. Ebene: dessen Lesartnummer
   1 INTERNAL_LINKS                            1. Ebene: interne Verknüpfung
      2 RELATION ,,has_hyperonym“                 Hyperonymverknüpfung
         3 TARGET_CONCEPT                         zum Konzept in dem Substantiv
            4 PART_OF_SPEECH ,,n“                 <logement:1> enthalten ist
            4 LITERAL ,,logement“
               5 SENSE 1
      2 RELATION ,,has_hyponym“                   Hyponymverknüpfung
         3 TARGET_CONCEPT                         zum angegebenen Konzept
            4 PART_OF_SPEECH ,,n“
            4 LITERAL ,,pension“
               5 SENSE 2
⋮
      2 RELATION ,,antonym“                       Antonymverknüpfung
         3 TARGET_CONCEPT
            4 PART_OF_SPEECH ,,n“
            4 LITERAL ,,artefact“                 zum Konzept <artefact:1>
               5 SENSE 1
         3 FEATURES
            4 VARIANT_TO_VARIANT                  lexikalische Verknüpfung von
               5 SOURCE_VARIANT ,,objet naturel“  <objet naturel>
               5 TARGET_VARIANT ,,artefact“       mit <artefact>
⋮
      2 RELATION ,,has_mero_part“                 Meronymverknüpfung
         3 TARGET_CONCEPT                         zum angegebenen Konzept
            4 PART_OF_SPEECH ,,n“
            4 LITERAL ,,véranda“
               5 SENSE 2
⋮
   1 EQ_LINKS                                  1. Ebene: Links zum ILI
      2 EQ_RELATION ,,eq_synonym“
         3 TARGET_ILI
            4 PART_OF_SPEECH ,,n“
            4 WORDNET_OFFSET 2728393
```

Abbildung 9.1: Ausschnitt aus einem (imaginären) Datenbankeintrag.

9.4 Beschreibung der Datenbank *WordNet*

Die *WordNet*-Datenbank (in der Version 2.0) besteht aus mehreren Dateien.[4] Die verschiedenen Wortarten sind getrennt, aber auch die zu ihnen gehörige Information ist auf jeweils drei Dateien aufgeteilt.

Zu jeder Wortart (`pos`) findet sich in der Datei `index.pos` die alphabetische Liste der *WordNet*-Elemente der jeweiligen syntaktischen Kategorie. In der Datei `data.pos` werden in numerischer Folge die Offsets und weitere Angaben zu den vorhandenen Verlinkungen (siehe folgende Tabelle) aufgeführt.

!	Antonym	=	Attribute
@	Hypernym	+	Derivationally related form
~	Hyponym	;c	Domain of synset
#m	Member holonym	-c	Member of this domain
#s	Substance holonym	;r	Domain of synset
#p	Part holonym	-r	Member of this domain
%m	Member meronym	;u	Domain of synset
%s	Substance meronym	-u	Member of this domain
%p	Part meronym		

Eine genauere Beschreibung der beiden wichtigsten Dateien folgt in den nächsten Abschnitten. Zusätzlich gibt es für Ausnahmen in der Orthographie oder bei unregelmäßiger Flexion eine Auflistung der Formen in der Datei `pos.exc`. Für die Verben sind zusätzlich in den Dateien `*.vrb` Beispiele für die Verwendung der Verben gegeben.

9.4.1 `Index.pos`

In dieser Datei finden sich alphabetisch geordnet die zu einer Wortart gehörigen Elemente der Taxonomien. Nach einer Copyrightinformation finden sich jeweils zeilenweise, durch Leerzeichen getrennt, folgende Informationen:

```
activity n 6 5 ! @ ~ = ; 6 6 00389883 13208694 12679776 13702946
12755345 04452731
```

[4] Die Beschreibungen finden sich in den *Manuals*, die mit den Dateien als *manpages* zur Verfügung gestellt werden.

Die einzelnen Einträge lassen sich wie folgt aufschlüsseln:

Eintrag	**Beispiel**	**Beschreibung**
`lemma`	activity	Lemma: Kleinschreibung, ASCII, Leerzeichen durch _ ersetzt
`pos`	n	Wortart (n, v, a, r)[5]
`synset_cnt`	6	Anzahl der *Synsets*, in denen Lemma auftaucht
`p_cnt`	5	Anzahl der verschiedenenen Verknüpfungen (in allen *Synsets*)
`[ptr_symbol...]`	! @ ∼ = ;	Verknüpfungssymbole
`sense_cnt`	6	Entspricht `synset_cnt` (Kompatibilitätsgründen)
`tagsense_cnt`	6	Wieviele der *Senses* in Texten annotiert wurden
`[synset_offset...]`	00389883 13208694 12679776 13702946 12755345 04452731	Auflistung der `Byte-Offsets` des Lemma, 8bit Integer, wird für die Suchfunktion der Auswertungsprogramme verwendet

9.4.2 `Data.pos`

In der `data.pos`-Datei werden numerisch die einzelnen `Offsets` aufgeführt, denen Informationen über `lexicographerfiles` und Verknüpfungen folgen.

```
00389883 04 n 01 activity 0 077 @ 00026194 n 0000 ! 01001511 n 0101
∼ 00184351 n 0000 ∼ 00280016 n 0000 [...] | any specific activity;
„they avoided all recreational activity“
```

[5] Das Kürzel n steht für Substantive, v steht für Verben, a für Adjektive und r für Adverbien.

Eintrag	Beispiel	Beschreibung
synset_offset	00389883	Aktuelles Byte-Offset des Lemmaeintrags in der Datei, 8digit Integer
lex_filenum	04	2digit Integer, die auf die lexicographerfiles verweist, in denen das *Synset* enthalten ist
ss_type	n	Wortart (n, v, a, s, r)[6]
w_cnt	01	Anzahl der im *Synset* enthaltenen Elemente (Hexadezimalzahl)
word	activity	Lemma: Groß- und Kleinschreibung, ASCII, Leerzeichen durch _ ersetzt
lex_id	0	Hinweis auf lexicographerfiles, identifiziert Lesart eindeutig, 1-digit Hexadezimalzahl[7]
[word lex_id...]		evt. weitere Synsetelemente
p_cnt	077	3-digit Dezimalzahl, gibt Anzahl der Verknüpfungen von diesem *Synset* an[8]
pointer_symbol[9]	@	Zeichen für die Art der Verknüpfung
synset_offset	00026194	Byte-Offset des Lemma
pos	n	Wortart
source/target	0000	4-digit Heximalzahl, die angibt, welches Element des Ausgangssynsets mit welchem Element des Zielsynsets verknüpft wird (bei lexikalischer Verknüpfung $\neq$ 0000)
[frames...]		nur in verb.data: Liste von Verbframes
| gloss	| any specific activity; „they avoided all recreational activity“	durch „|“ gekennzeichnet beginnt ein Textstring, der eine Definition und Beispiele enthalten kann.

[6] Neben den vorher erwähnten Abkürzungen steht hier das Kürzel s für *adjective satellite*.
[7] Der default-Wert ist 0, d.h. das Lemma ist nicht in den lexicographerfiles enthalten.
[8] Die Angabe 000 würde auf keine Verknüpfungen hinweisen.
[9] Die folgenden vier Einträge werden je nach der Anzahl der Verknüpfungen wiederholt.

Lesartennummer

Innerhalb des *Synsets* sind die Elemente nach ihrer Frequenz angeordnet (das häufigste steht an erster Stelle), der Eintrag `tagsense_cnt` in der Datei `index.pos` gibt an, wieviele der Lesarten der Liste annotiert wurden.

Eine Lesart wird durch die drei Angaben `word`, `lex_id` und `lex_filenum` eindeutig identifiziert. In der Datei `index.sense` (siehe folgender Abschnitt) werden diese Angaben in `sense_key` codiert. Jedes *Synset* in der Datenbank wird durch die Angabe von `synset_offset` zusammen mit der Angabe der syntaktischen Kategorie eindeutig bezeichnet (die Zahl `synset_offset` ist nur innerhalb der jeweiligen Datei `data.pos` eindeutig).

9.4.3 Weitere Dateien

`sense.index`

Dieser Index bietet eine weitere Möglichkeit, die *Synsets* und Lesarten in der *WordNet*-Datenbank zu suchen. Die Einträge `lemma` und `lex_sense` werden durch `%` zum Eintrag `sense_key` zusammengekettet. Dieser Eintrag ist unabhängig von *WordNet*, denn in den unterschiedlichen Versionen von *WN* ist die Zuweisung der *sense_numbers* und der *synset_offsets* nicht einheitlich. Mit diesem *sense_key* kann das gesuchte *Synset* und die versionsabhängige *WordNet sense number* gefunden werden.

```
business%1:14:02:: 07485368  3  177
```

`pos.exc`

Diese Datei enthält in alphabetischer Reihenfolge die flektierten Formen und ihre Basisformen. Das erste Element einer Zeile ist eine flektierte Form gefolgt von einer oder mehreren möglichen Basisformen.

```
aardwolves aardwolf
agnomina agnomen
```

Die weiteren Dateien `cntlist` und `cntlist.rev` werden bei der Erstellung der *WN*-Datenbank benötigt. Für die vorliegende Arbeit sind sie weniger wichtig, da *FWN* durch eine Übersetzung der existierenden Netzstruktur erstellt wurde.

9.4.4 Das Werkzeug *wn*

Das Kommandozeilenprogramm `wn`[10] ermöglicht, direkt mit Daten im *WN*-Format zu arbeiten. Es lassen sich alle zu einem Eintrag gehörenden Angaben ausgeben (z.B. Liste der Synonyme im *Synset*, alle Verknüpfungen eines Typs, der gesamte Hyperonym- oder Hyponymbaum, etc.). Dieses Tool läßt sich leicht in kleinere Skripte einbauen und vereinfacht den Datenzugang.

9.5 Transformation der Datenbank

Für die Umwandlung der Datei im *EWN*-Format in das *WN*-Format wurde ein Perlprogramm (`ewn2wn.pl`[11]) erstellt. Aus der Ausgangsdatei werden die Informationen für die *WN*-Datenbankdateien geholt und im vorgegebenen Format in `data.noun` und `index.noun` abgelegt. Im vorliegenden Fall werden nur die Substantive aus der französischen Datenbank geholt. Als Ausgabedateien entsteht neben `data.noun` und `index.noun` noch eine Datei, die nur die Substantive der Datenbank enthält (in *EWN*-Format) und zwei Indexdateien (alphabetisch und numerisch). Da nicht mit flektierten Formen gearbeitet wird, wird eine *Default*-Datei `noun.exc` erstellt, die nur jeweils zwei unflektierte Formen aufeinander abbildet. Desweiteren wird eine Auswertungsdatei erstellt, mit der die Unterknotenwahrscheinlichkeit berechnet werden kann.[12]

Einige Informationen werden während des Programmverlaufes neu erstellt: Das `synset_offset` wird bei der Erstellung der Datei `data.noun` erstellt (ist jeweilig aktuelles `Byte-Offset` des Synseteintrags in dieser Datei). Andere Angaben (`wn_cnt` aus `data.noun` und `sense_cnt`, `synset_cnt` in `index.noun`) lassen sich einfach errechnen. Die Verknüpfungssymbole (`pointer_symbol` aus `data.noun` und `ptr_symbol` aus `index.noun`) werden soweit wie möglich auf die entsprechenden Symbole aus *WN* abgebildet.

[10] Ausführliche Dokumentation ist zu finden unter `http://www.cogsci.princeton.edu/~wn/doc.shtml` (13.01.2005).

[11] Für eine genauere Beschreibung des Programmablaufs siehe im Anhang unter A.2.

[12] Diese wird für die von Resnik, Jiang und Conrath und Lin verwendeten Zusammenhangsmaße benötigt (netzbasierte Wahrscheinlichkeitsberechnung aus Abschnitt 5.1.2). Für die Berechnung siehe den Abschnitt A.2.2 im Anhang.

Abbildung der Verknüpfungen

Ein Problem stellt der Übertrag der Relationen aus der *EWN*-Datenbank auf das *WN*-Format dar, da im *EWN* 28 Verknüpfungsarten verwendet werden, die nicht eineindeutig auf Verknüpfungen von *WN* abgebildet werden können.

Für die Verknüpfungen `has_holo_portion`/`has_mero_portion` (jeweils einmal) und `has_holonym`/`has_meronym` (jeweils 50mal vertreten) gibt es in *WN* keine eineindeutige Entsprechung. Daher müssen diese per Hand auf die in *WN* vorhandenen Teil-Ganzes-Verknüpfungen abgebildet werden. Insgesamt wurden die Links `has_holo_portion` und `has_holonym` auf `has_holo_part` und die Links `has_mero_portion` und `has_meronym` auf `has_mero_part` bis auf eine Ausnahme abgebildet (Verknüpfung von <cochon:1> und <lard:1>: mero_madeof).

Die jeweils zweimal auftretenden Verknüpfungen `involved`/`role`, `involved_agent`/ `role_agent` (4), `involved_instrument`/ `role_instrument` (10) und `involved_location`/ `role_location` (einmal) besitzen in *WN* ebenfalls keine Entsprechung und werden daher auf neue Symbole abgebildet, die von `Query-Data.pl` bei einer Erweiterung des Programmcodes ausgewertet werden könnten, aber vorläufig nicht beachtet werden.

Die durch die Verknüpfung mit dem *ILI* bestehende Verbindung zu *Synsets* in *WN* 1.5 (`eq_generalization`, `eq_metonym` und `eq_synonym`) geht bei dieser Transformation natürlich verloren, da in *WN* nur interne Links vorhanden sind. Auf diese Informationen kann für einen interlingualen Vergleich (englisch/französisch) über die *EWN*-Datenbank zugegriffen werden. Eine Zusammenfassung dieser Abbildung findet sich in Tabelle 9.1 auf Seite 98.

Glossen

Über den *ILI*-Index sind die französischen Datenbankeinträge mit englischen *Synsets* verknüpft. Daher könnte auch über diese Verbindung die jeweilige (englischsprachige) Glosse eingebunden werden. Allerdings wird die Glosse nicht von dem entwickelten Modul und auch nur von einem Modul von `similarity.pl` verwendet. Der Aufwand, die Nummern der *Offsets* der Version 1.5 (über Zwischenschritte) auf die jeweiligen Offsetnummern der Version 2.0 abzubilden erwies sich als zu groß, zudem ein Teil der im Rahmen des *EWN*-Projektes neu erstellten französischen Datenbankeinträge keine Verknüpfung zum *ILI* besitzen. Es wird daher in der Datei `data.noun` ein *Dummy*-Eintrag eingesetzt.

Verknüpfung in EWN	in WN	Symbol
has_hyperonym	@	@
has_hyponym	~	~
antonym	!	!
has_holo_madeof	#s	#s
has_holo_member	#m	#m
has_holo_part	#p	#p
has_holo_portion	wird zur Kategorie has_holo_portion	#p
has_holonym	Spezialisierung der Verknüpfung	#p | #m
has_mero_madeof	%s	%s
has_mero_member	%m	%m
has_mero_part	%p	%p
has_mero_portion	wird zur Kategorie has_mero_part	%p
has_meronym	Spezialisierung der Verknüpfung	%p | %m
involved	nicht vorhanden	§
involved_agent	nicht vorhanden	§a
involved_instrument	nicht vorhanden	§i
involved_location	nicht vorhanden	§l
role	nicht vorhanden	-§
role_agent	nicht vorhanden	-§a
role_instrument	nicht vorhanden	-§i
role_location	nicht vorhanden	-§l
eq_generalization	Verknüpfung mit ILI	
eq_metonym	Verknüpfung mit ILI	
eq_synonym	Verknüpfung mit ILI	
causes	nicht bei Substantiven	
is_caused_by	nicht bei Substantiven	
has_subevent	nicht bei Substantiven	
is_subevent_of	nicht bei Substantiven	

Tabelle 9.1: Übertragung der Relationen

Für den Vergleich mit den anderen Ähnlichkeitsmaßen wird deshalb das Maß `vector.pm` weggelassen, da dieses direkt mit den Elementen der Glossen arbeitet.

9.6 Verbesserung der Datenbank

Bei der Arbeit mit der vorliegenden Datenbank ergaben sich aus dem fehlerhaften Format der Strukturen mehrere Probleme. Da die Kohärenz der Datenbank aber die Grundlage für die vorliegende Arbeit darstellt, wurde in diesen Rahmen die Verbesserung als wichtiger Arbeitsteil mitaufgenommen. Die Fehler der Datenbank werden im folgenden dargestellt.

Freischwebende Knoten

Ein Fehler der zu vielen Problemen bei der Verarbeitung des Netzwerkes durch die für *WordNet* erstellten Programme führt, stellen die Knoten dar, die sich im Netz ohne Verbindungen zu anderen Knoten befinden:

```
00000074 00 n 01 zone_active 000 | GLOSSE
00000131 00 n 01 OEM 000 | GLOSSE
00002272 00 n 01 salmonella 000 | GLOSSE
00002370 00 n 01 caïman 000 | GLOSSE
00002666 00 n 01 puceron 000 | GLOSSE
00006171 00 n 02 épi voie_d'embranchement 000 | GLOSSE
00015299 00 n 01 crocus 000 | GLOSSE
00015312 00 n 01 haricot_vert 000 | GLOSSE
00015343 00 n 01 citronnier 000 | GLOSSE
00022775 00 n 01 bannière 000 | GLOSSE
00022818 00 n 01 compression 000 | GLOSSE
00022825 00 n 01 couplage 000 | GLOSSE
00022883 00 n 01 fragment 000 | GLOSSE
00022947 00 n 01 marquage 000 | GLOSSE
00022975 00 n 01 nom_de_fichier 000 | GLOSSE
00022982 00 n 01 oeM 000 | GLOSSE
```

Es zeigt sich (neben dem offensichtlichen Tippfehler der zwei Einträge für OEM als Abkürzung für: *original equipment manufacturer*, Hersteller des Originalerzeugnisses), daß diese Knoten ohne erkennbaren Grund im Netz „schweben“. Sie stammen aus den unterschiedlichen Netzbereichen und sind manchmal in anderer Lesart (z.B. <marquage:3> in erster und zweiter Lesart) fest im Netz verknüpft. Wie die unkorrekten Kopfknoten (siehe weiter unten) werden diese Knoten eingepflegt.

- Der Eintrag für <OEM:1> wird beibehalten. Der parallele Eintrag für <oeM> wird – nach der Übernahme der dort angegebenen Verknüpfung mit dem *ILI* – gelöscht. Dieses *Synset* wird unter <logiciel:1, software:1> eingehängt, da sich dort auch andere „unechte“ Hyponyme wie <WYSIWYG:1> befinden.
- Ebenfalls wird der Knoten <couplage:1> unter <logiciel:1, software:1> eingehängt.
- Als Meronym (genauer: mero-part) wird <zone active:1> unter <interface graphique:1> und als Hyponym unter <zone:3> eingebunden.
- Das Bakterium <salmonella:1> wird unter <bactérie:1> eingehängt.
- Das *Synset* <caïman:1> wird unter <reptile:1> eingepflegt.
- Als Hyperonym zu <puceron:2> wird <puceron:1> ausgesucht.
- Als Hyponym von <chemin de fer:1> wird das *Synset* <épi:3, voie d'embranchement:1> übernommen.
- Der Eintrag für <crocus:1> wird als Hyperonym zu <plante bulbeuse:1> eingearbeitet.
- Als Holonym für <haricot vert:1> wird <haricot vert:2> eingetragen. Außerdem wird <haricot vert:2> unter <plante cultivée:1> als Hyponym eingehängt.
- Das *Synset* <citronnier:1> wird unter <herbe:2, plante herbacée:1> eingehängt.
- Als Hyponym zu <marqueur:1> wird <bannière:4> eingepflegt.
- Ebenfalls wird <marquage:3> dort eingehängt.
- Das *Synset* <compression:3> wird als Unterbegriff von <contraction:4, compression:2> eingeordnet.
- Das *Synset* <fragment:3> wird unter <fragment:2, morceau:7> eingehängt.
- Als Teil von <fichier informatique:1, fichier:2> wird <nom de fichier:1> abgelegt, gleichzeitig auch als Hyperonym von <nom:1>.

Orthographische Fehler

Bei einem genaueren Durchgang der über 17.000 Graphien fanden sich viele Rechtschreibfehler, die durch Nachlässigkeit oder Probleme mit den Sonderzeichen entstanden.

Manchmal fanden sich einige Lesarten mit einer falschen Graphie neben der richtigen Schreibung für dasselbe Lemma, diese wurden somit im Transformationsprozeß nicht zusammengeführt. Eine besondere Fehlerquelle war die unterschiedliche Verwendung der Akzentzeichen. Das Substantiv <entraîneur> wird beispielsweise als Datenbankeintrag <entraîneur:1> im Synset zusammen mit <dompteur:2> geführt, allerdings als Hyperonym des Eintrags <entraineur:2>, wobei <entraineur:1> als weiterer Eintrag zusammen im *Synset* mit <tuteur:1> existiert. Daher wird der Eintrag <entraîneur:1> als <entraîneur:3> übernommen und der Orthographiefehler in den anderen Einträgen korrigiert.

Da die Schreibungen <pagaïe> und <pagaille> gleichberechtigt nebeneinanderstehen (in insgesamt drei Lesarten aber nur in jeweils zweien auftauchen), wird <pagaïe:3> zum Synset <*pagaille:1*, *désordre:1*> und <pagaille:3> zum Synset <*pagaïe:1*, *chaos:1*> hinzugefügt.

Zeicheninkohärenz

Ein weiterer Fehler, der die Umwandlung erschwert, ist die unterschiedliche Behandlung der Groß- und Kleinschreibung. Das Substantive *assainissement* wurde in Kleinschreibung mit der Lesart 2 in ein Synset geschrieben, die Lesart 1 wurde in Großschreibung *Assainissement* abgespeichert, wodurch eine völlig getrennte Verarbeitung beim automatischen Lesen der Datei veranlaßt wird. Dieser Fehler mußte bei vielen Einträgen manuell in der französischen Datenbank korrigiert werden.

Auch ist die unterschiedliche Verwendung von <-> in Lemmata (z.B. <tam tam> vs. <tam-tam>) eine Ursache für den Abbruch der Verarbeitungsprogramme. Diese beiden Konzepte sind als Schwesternknoten unter <tambour:2> eingehängt, daher wird der Datenbankeintrag mit dem Trennstrich gelöscht.

Die verarbeitenden Programme, die für die englische Sprache erstellt wurden, lesen „Nichtwortzeichen“, wie z.B. das bei der französischen Wortbildung häufig auftauchende Apostroph als Fehler der Datenbank. Daher mußte (um die Originalprogramme nicht zu verändern) bei den französischen Daten eingegriffen

werden: <allocation d'Aide publique> wurde durch <allocation d aide publique> ersetzt. Diese Veränderung muß bei der konkreten Verwendung berücksichtigt werden. Das Zeichen <œ> wird durchgängig durch <oe> ersetzt, da sich durch die Verwendung dieses Sonderzeichens in der Datenbank verschiedene Probleme ergeben.

Eine weitere Besonderheit der französischen Datenbank ist der Eintrag der genusflektierten Formen durch das Anfügen der femininen Form in Klammern (bei zélé(e), salarié(e), offensé(e), roturier(e), estropié(e), suppléant(e)). Diese Formen werden durch die unmarkierte Form ersetzt.

Bei zwei Formen werden alternative Schreibungen durch die ersetzt, die im *Petit Robert* zu finden sind: <veld(t)> wird durch <veld> ersetzt, <lande(s)> durch <landes> (getrennt von dem Eintrag <lande>) und <achar(d)> durch <achards>.

Die Verwendung der Pluralform im Konzept <vêtements:1> ist nicht klar. Wegen der Existenz von <vêtement:2> wird das Substantiv in den Singular gesetzt.

Bei manchen Einträgen (z.B. *acte d'accusation*) sind in der Datenbank zuviele Leerzeichen, die beim Transformationsprozeß umgedeutet und durch den Unterstrich (aus Kompabilitätsgründen) ersetzt werden. Dieser Fehler wird durch das Leerzeichen ␣ dargestellt.

Die Ansicht des Eintrages <a:> als Bezeichnung für das bei MS-DOS üblicherweise unter diesem Laufwerksbuchstaben eingehängte Diskettenlaufwerk in der Computerterminologie scheitert bei *Periscope* durch die Miteinbeziehung des Zeichens <:>. Es kann lediglich indirekt über sein Hyperonym (<partition:2>, wenn die Verknüpfung bekannt ist) im *Periscopeviewer* betrachtet werden. Bei der Datenbankumwandlung wurde das nicht berücksichtigt und der Eintrag <a:> wurde mit einer einzigen Lesart übernommen.

Fehlende Lesartennummern

Für den Eintrag <milieu> finden sich die Lesarten #2, #4, #5, #6, #7, #8, #9, #10, #12, #13. Beim Eintrag <Milieu> sind die Lesarten #1 und #3 zu finden, so daß nach einem Zusammenführen der Groß- und Kleinschreibung insgesamt 12 Lesarten vorliegen. Die Lesart #11 ist nicht in der Datenbank zu finden. Ob hier eine Lesart unterschlagen wurde oder der Zähler bei der Erstellung einen Sprung gemacht hat, ist nicht zu klären.

Für das *Synsetelement* <champignon> war z.B. die Lesartennummer #4 vergeben worden, während aber nur zwei Einträge insgesamt existieren. Dieser Fehler taucht insgesamt in über 20 Einträgen auf, so daß eine neue Verteilung der Lesartennummern stattfinden mußte, mit der Korrektur der jeweiligen Verknüpfungsknoten im Netz.

Weitere Ergänzungen

Interessant ist der einzige Eintrag zu <soleil:1> im semantischen Netzwerk. Er ist als Unterknoten zu <ami:1> eingetragen und mit der englischen Glosse <a person regarded very fondly; „the light of my life“> verknüpft. Diese im Englischen existierende Lesart ist im Französischen nicht vorhanden.

Die mit dem Himmelskörper im Zusammenhang stehenden Konzepte wie <lunettes de soleil>, <éclipse de soleil>, <rayon de soleil>, <lumière de soleil> und <coucher du soleil> sind im Netz eingebunden. Warum ausgerechnet der Himmelskörper als eigenes Konzept fehlt (während in *WN* 1.5 durchaus ein Eintrag zu <sun> (<any star around which a planetary system evolves>) existiert), ist nicht zu erklären. Daher wird ein neuer Knoten in das Netz eingefügt: <soleil:2> mit dem Hyperonym <astre:2> und dem Meronym <lumière du soleil:1>.[13]

Weitere solcher Ergänzungen sollten und konnten nicht vorgenommen werden, da dies mit automatischen Verfahren und französischen Datenbanken (vgl. die bei der Erstellung von *FWN* verwendete Datenbank *Dictionnaire Intégral*) in größerem Maßstab vorgenommen werden kann.

Einige Einträge sind sehr typisch für das britische Englisch. Sie wurden bei der Erstellung des französischen Teilnetzes mit übernommen, wobei ihre Existenzberechtigung anzuzweifeln ist: Als Beispiel ist der <solicitor:1> als *Synonym* zu <conseiller:7> mit der Glosse <a British lawyer who gives legal advice and prepares legal documents> wie auch <reel:1> (<a lively dance of Scottish highlanders; marked by circular moves and gliding steps>) und <reel:2> (<music composed for dancing a reel>) nicht in den verwendeten französischen Wörterbüchern zu finden.

[13] Bei der Verbindung mit *WN*-1.5 durch eine EQ-Synonym-Verknüpfung bietet sich das Konzept [*Sun:1*] an: <the star that is the source of light and heat for the planets in the solar system>. Die Schreibung mit Großbuchstaben in *WN*-1.5 läßt sich nicht erklären.

Einzelne Einträge, die im Englischen einfach lexikalisiert sind, werden im französischen Netz durch komplexe Lexien wiedergegeben. Als Übersetzung von [*memorizer:1*] wird der französische Konzeptknoten [*quel qu'un qui apprend par cœur:1*] eingetragen. Eine Verbesserung der Datenbank hinsichtlich dieser Probleme kann hier nicht vorgenommen werden.

Neue Knoten

Die Auswertungen aus dem Kapitel 8.5 wurden mit den im Netz existierenden Knoten verglichen. Einige der als aussagekräftig herausgestellten Graphien sind nicht als Konzeptknoten im Netz vorhanden: <mars>, <ascendant und <duc>. Diese wurden (jeweils in einer einzigen Lesart) in das Netz eingepflegt: <mars:1> als Hyperonym zu <planète:1>, <ascendant:2> als Hyponym zu <astre:2> und Meronym zu <zodiaque:1> und <duc:1> als Hyperonym zu <noble:2>.

Unmotivierte Kopfknoten

Einige Knoten des originalen Netzes wurden nur mit einer Meronymbeziehung in das Netz gehängt. Daraus resultiert wohl der Fehler, daß sie in den Publikationen zu *FWN*[14] als Kopfknoten behandelt werden. Beispiele werden dort nicht angegeben, allerdings die Begründung, daß nur *Base Concepts* oder deren direkte Hyperonyme als Kopfknoten in Frage kommen.

Da aber im französischen Netz diese Knoten ohne ersichtlichen Grund als Kopfknoten im Netz erscheinen, werden sie manuell im Netz verankert (es wird jeweils eine neue Hyperonym- und Hyponymverlinkung erstellt):

- Das *Synset* <mandarinier:1> wird analog zu anderen *Synsets* (wie <bananier:1>) unter <herbe:2, plante herbacée:1> eingehängt.
- Das *Synset* <artiodactyle:1> wird unter <ongulé:1> eingepflegt.
- Unter <marquage:1> wird <balise:3> und <lien hypertexte:3> abgelegt.
- Das *Synset* <navigateur:1> wird unter <programme utilitaire:1, utilitaire:1> eingehängt.
- Ebenfalls wird <lecteur de news:1> als Hyponym von <programme utilitaire:1, utilitaire:1> eingepflegt.

[14] Siehe besonders Catherin 1999, 5.

- Als weiterer Tochterknoten von <programme utilitaire:1, utilitaire:1> wird <lecteur multimédia:1> übernommen.
- Das *Synset* <constrictor:1> wird als Hyponym unter <reptile:1> eingehängt.
- Als neues Hyponym für <croyance:3> wird <doctrine:1> abgelegt.

Damit ergeben sich die folgenden 12 Kopfknoten für die Substantivhierarchie. Auf das *Synsetoffset* der verbesserten Datenbank (insgesamt 17.827 Knoten[15]) folgt die Tiefe der Hierarchie und die Lemmata des *Synsets*):

```
00000000 16 (imaginärer Kopfknoten) (20815 Unterknoten)
00000001 15 entité (10871 Unterknoten)
00006106 12 chose du psychisme (1176 Unterknoten)
00006211 12 abstraction (3198 Unterknoten)
00007581 9 emplacement, localisation (873 Unterknoten)
00007785 8 forme (141 Unterknoten)
00008347 10 état (776 Unterknoten)
00009050 7 événement (353 Unterknoten)
00009146 11 acte, action (2163 Unterknoten)
00009444 11 groupement, groupe (684 Unterknoten)
00009712 12 possession (251 Unterknoten)
00011156 10 phénomène (330 Unterknoten)
```

Die unterschiedliche Tiefe in den Subhierarchien spiegelt sich auch in der durchschnittlichen Tiefe wieder, die bei 6,5 liegt. Die Tiefe der Blattknoten ist der folgenden Übersicht zu entnehmen.

[15]Die Summe der Unterknotenzahl entspricht wegen der Möglichkeit der Doppelvererbung nicht der Anzahl der Knoten im Netz.

```
1. Ebene: 11
2. Ebene: 113
3. Ebene: 701
4. Ebene: 1794
5. Ebene: 2951
6. Ebene: 3540
7. Ebene: 3721
8. Ebene: 2429
9. Ebene: 1319
10. Ebene: 833
11. Ebene: 239
12. Ebene: 109
13. Ebene: 46
14. Ebene: 18
15. Ebene: 3
```

Die folgende Auflistung gibt einen Überblick über die in der entstandenen Datenbank vorhandenen Verknüpfungen (vgl. die ursprünglichen Werte in der Tabelle B.1 auf Seite 208).

```
18040 ,,has_hyperonym“
18040 ,,has_hyponym“
512 ,,antonym“
1121 ,,has_holo_part“
1121 ,,has_mero_part“
131 ,,has_holo_member“
131 ,,has_mero_member“
51 ,,has_holo_madeof“
51 ,,has_mero_madeof“
1 ,,has_holo_portion“
1 ,,has_mero_portion“
```

Die so entstandene Datenbank mit 17.828 Einträgen ist damit eine grundlegend korrigierte Fassung der originalen Datenbank, die während der Projektphase von *EWN* entstanden ist.

Kapitel 10

Zusammenhangsmaß

Bei der Betrachtung der in Kapitel 5 vorgestellten Zusammenhangsmaße fällt zuerst das intendierte Fehlen jeglicher textbasierter Information auf. Die statistischen Elemente, die von Resnik, Leacock-Chodorow und Lin verwendet werden, sind für die französische Sprache wegen des Fehlens entsprechend annotierter Korpora auf der Grundlage des Netzes berechnet.

Daher wird mit der im vorangegangenen Kapitel erarbeiteten korpusbasierten Erweiterung eine Erweiterung des semantischen Netzes vorgenommen, die durch ein Zusammenhangsmaß ausgewertet werden kann. Damit steht dieser Arbeitsansatz in der Tradition der *Bayesian Networks*.

10.1 *Bayesian Networks*

Mit *Bayesian Networks* werden basierend auf einem annotierten Korpus Informationen gesammelt, die Aussagen machen über Kollokationen und insbesondere über signifikant häufige Kollokationen im Kontext einer lexikalischen Einheit.

Aus einer Korpusauswertung heraus werden hier die einzelnen lexikalischen Einheiten durch gerichtete Kanten miteinander verbunden, die durch ihre „Gewichtung“ die bedingte Wahrscheinlichkeit wiedergeben, mit der der angeknüpfte Knoten nach dem Ursprungsknoten in einer Kollokation auftaucht.[1] Ein Überblick und eine Bewertung der einzelnen statistischen Methoden findet sich bei Gale

[1] Ausgehend von Quillian 1968 sind die Publikationen von Pearl grundlegend, besonders Pearl und Russel 2004 und im Kontext der Lesartendisambiguierung Eizirik et al. 1993, Hirst 1987 und die Publikationen von Yarowsky, besonders Yarowsky 1994a und Yarowsky 1994b.

et al. 1993. Mit diesen *Bayesian Networks* werden auf der Grundlage großer paralleler Datenmengen (der *Canadian Hansards*[2]) Disambiguierungsmodelle entwickelt, die in bis zu 92% der Fälle bei klar unterscheidbaren Lesarten die jeweils korrekte Lesart zuweisen. Andere Modelle greifen auf kleinere Datenmengen handannotierter Korpora zurück, erreichen aber ähnliche Wertungen.

Eine Brücke schlägt Wiebe (Wiebe et al. 1998) zum semantischen Netzwerk *WN*: Sie erstellt ein *Bayesian Network* auf der Grundlage von *WN* und verbindet es mit statistischen Elementen aus annotierten Korpora. Auf dieser Verbindung von manuell disambiguierten Korpusdaten und der Struktur von *WordNet* wird hier aufgebaut.

10.2 Korpusbasierte Netzerweiterung

Der grundlegende Unterschied des Ansatzes dieser Arbeit zum *Bayesian Network* ist, daß im *BN* der Ausgangsknoten durch mehrere gerichtete Kanten mit möglichen Kollokationen verbunden ist, so daß die Weiterleitung über die gewichteten Kanten zu den neu angeknüpften Knoten erfolgt. Im Gegensatz dazu wird bei dem Maß, das hier vorgestellt wird, von einem netzimmanenten Knoten eine neue gerichtete Kante ausgehen, die gewichtet wird durch die Häufigkeit ihrer Kollokation mit dem Zielknoten.

Durch die Suche nach dem kürzesten Verbindungspfad unter Einbeziehung dieser neuen Verbindungen werden daher gleichlange Verbindungspfade im ursprünglichen Netz in ihrer Wichtigkeit abgeschwächt. Auch werden damit Verknüpfungen zwischen zwei Konzeptknoten erstellt, die im ursprünglichen Netz nur durch sehr lange Verbindungspfade verbunden sind.

10.3 Mathematische Realisierung

Die Verwendung der Pfadlänge bietet sich bei der Form des semantischen Netzes als Wert für den semantischen Zusammenhang an. Die Miteinbeziehung der neuen Knoten wird durch eine unterschiedliche Gewichtung der Kanten beim Durchlaufen der Verbindung verwirklicht. Bei einer Kante, die frequent im

[2]Die Zusammenstellung der zweisprachig vorliegenden Debatten des kanadischen Parlamentes.

Kontext einer Lesart auftaucht, wird der Weg symbolisch verkürzt gegenüber einer normalen Netzkante.

Die Miteinbeziehung des Informationsgehaltes gewährleistet eine weitere Differenzierung des Maßes: Je weiter der Pfad in der Hierarchie nach oben steigt (zum gemeinsamen Ahnenknoten), desto kleiner sollte der Zusammenhangswert sein. Daher sollte wie im Maß von Lin das Verhältnis der Information, die in den Konzepten enthalten ist, miteinbezogen werden.

Außerdem enthält ein längerer Pfad in der Reihung der Konzepte unweigerlich Assoziationssprünge. Daher wird eine zusätzliche Schwächung des Wertes bei einem zu langen Pfad eingeführt.

Aus den vorangegangenen theoretischen Überlegungen ergibt sich die folgende Formel:

$$\left(32 - \sum_{\nu \in \min(\varphi)} \frac{g}{3} \right) * \frac{2 * \log (p_N(MK))}{\log (p_N(k_1)) + \log (p_N(k_2))} \tag{10.1}$$

Der Wert für den maximalen Abstand in der Taxonomie wird wie in den anderen Maßen auf 32 gesetzt, da die maximale Tiefe der Hierarchie 16 beträgt.[3] Für zwei Blattknoten dieser tiefsten Hierarchie, die nur über den Kopfknoten miteinander verbunden sind, ist der Abstandswert 32. Dem Problem der unterschiedlichen Dichte (die durchschnittliche Dichte liegt nicht bei 8, sondern bei 6,5, siehe Abschnitt 9.6) wird durch den nichtlinearen Verlauf Rechnung getragen.

Vom maximalen Abstand in der Taxonomie wird ein Wert abgezogen, der sich wie folgt berechnet: Es werden wie bei dem kantenzählenden Maß (vgl. Abschnitt 5.4) die Kanten auf dem minimalen Verbindungspfad gezählt. Es wird allerdings – abhängig von der jeweiligen Kante, die gezählt wird – eine Gewichtung vorgenommen. Dafür werden die neuen Kanten in zwei Klassen eingeteilt (vgl. Abschnitt 8.7):

$\mathcal{A}$ Diese Kanten weisen im Vergleich mit den anderen statistisch basierten Kanten einen sehr hohen Wert auf. Die *Synset*-Elemente der neuen Knoten tauchen mit dem Ausgangskonzept sehr häufig zusammen auf.

$\mathcal{B}$ Diese Kanten tauchen häufig mit dem Ausgangskonzept auf.

[3] Leider kann hier nicht auf die unterschiedliche Tiefe der Hierarchie Rücksicht genommen werden. Dazu müßte – je nach durchlaufener Subhierarchie – der maximale Abstand zweier Knoten dort während der Berechnung ermittelt werden.

EWN Diese Kante ist im ursprünglichen *EWN* vorhanden und wird daher normal mitgezählt. Der Abstand wird durch $g = 3$ nicht verkürzt.

Die Variable g nimmt je nach der Klassenzugehörigkeit Werte aus der Menge $(1, 2, 3)$ an:

$$g = \begin{cases} 1 & : \quad \nu \in \mathcal{A} \\ 2 & : \quad \nu \in \mathcal{B} \\ 3 & : \quad \nu \in EWN \end{cases}$$

Insgesamt verkürzt sich durch die verschiedenen Werte von g die Pfadlänge, d.h. der Abstand, den die beiden verglichenen Konzepte voneinander haben. Beispielsweise verkürzt sich die Länge eines acht Kanten umfassenden Pfades auf die Pfadlänge $|\varphi| = 7$, falls einmal über die Kante einer sehr häufigen Kollokation ($\nu \in \mathcal{A}$) gegangen wird und einmal über die einer häufigen Kollokation ($\nu \in \mathcal{B}$).

Zusätzlich wird noch eine Gewichtung vorgenommen, die den Informationsgehalt des kleinsten gemeinsamen Ahnenknotens ($kgAK$) und der beiden verglichenen Konzepte berücksichtigt (vgl. Maß von Lin, Abschnitt 5.10). Liegt der Ahnenknoten sehr niedrig in der Hierarchie und nah bei den Ausgangsknoten (der Pfad ist dann auch kürzer), ist der Bruch sehr nah bei 1 (≤ 1). Der Wert ist 1, wenn die Ausgangsknoten identisch sind (und damit $\log(\, p_N(k_1)\,) + \log(\, p_N(k_2)\,) = 2 * \log(\, p_N(MK)\,)$. Liegen die Ausgangsknoten tiefer in der Hierarchie und der Ahnenknoten in einer sehr viel höheren Ebene der Taxonomie, ist der Bruch sehr viel kleiner als 1. Wird der absolute Topknoten vom Verbindungspfad durchlaufen, ist der rechte Teil der Formel und damit der gesamte Ausdruck Null.

Damit wird insgesamt eine doppelte Verstärkung des Zusammenhangs vorgenommen, falls ein kurzer Verbindungspfad in einer niedrigen Ebene der Hierarchie vorliegt. Damit wird das Problem der zu langen Assoziationskette durch die Zuweisung sehr kleiner Werte bei einem langen Verbindungspfad gelöst.

10.3.1 Normalisierung des Maßes

Als letzter Schritt (der besseren Vergleichbarkeit mit den anderen Maßen wegen) wird dieses Maß durch die Division mit 32 normalisiert. Der Wertebereich ist das Intervall [0,1]. Der Wert 1 wird erreicht, wenn die Ausgangsknoten in demselben

Synset auftauchen, der Wert 0 wird vergeben, wenn der Kopfknoten durchlaufen wird. Es ergibt sich damit insgesamt:

$$\left(1 - \sum_{\nu \in \min(\varphi)} \frac{g}{96} \right) * \frac{2 * \log (p_N(MK))}{\log (p_N(k_1)) + \log (p_N(k_2).)} \tag{10.2}$$

Umgeformt als Abstandsmaß hat es die Eigenschaften einer Metrik, da es für Elemente desselben Konzeptknotens den Abstandswert 0 vergibt und die Symmetrieeigenschaft gegeben ist. Auch wird durch die Verwendung des jeweils kürzesten Verbindungspfades sichergestellt, daß die Dreiecksgleichung erfüllt ist, also ein Weg über einen dritten Konzeptknoten den Pfad nicht verkürzen kann.

Es wird jeweils beim Ausgangslemma und beim Ziellemma eine mögliche Verwendung der „Abkürzung“ getestet, die neu hinzugefügten Kanten werden also in beide Richtungen mit gleicher Wertung durchlaufen. Für feste Werte $g \in (1, 2, 3)$ ist die Funktion stetig, somit sind keine Sprünge im Funktionsgraphen vorhanden.

Die Berechnung wird durch ein Perl-Programm realisiert. Die genauere Erläuterung des erstellten Programms findet sich im Anhang im Abschnitt A.5.2.

Kapitel 11

Auswertung

Durch die korpusstatistische Erweiterung des Netzes, in der u.a. eine direkte Verknüpfung erstellt wird von Konzeptknoten wie [*maison:4*, *firme:1*] und [*employeur:1*, *patron:2*], die im originalen Netz nur über den imaginären Kopfknoten verbunden sind, werden Zusammenhänge erfaßt, die bei der Erstellung des Netzes (das auf semantischen Relationen beruht) nicht intendiert waren.

Allerdings ist der kontextuelle Zusammenhang durch die (wenn auch hier in kleinem Rahmen durchgeführten) Korpusuntersuchungen gestützt. So wird statt des Zusammenhangswertes 0 (wegen des Verbindungspfades über den subhierarchieübergreifenden imaginären Kopfknoten) ein höherer Wert vergeben. Dieser Zusammenhang kann für die Verwendung innerhalb eines Lesartendisambiguierungsalgorithmus verwendet werden, weil er die Vorteile des semantischen Netzwerkes mit den der *Bayesian Networks* verbindet.

Da zum Zeitpunkt der Erstellung der Kollokationen für ein Beispiellexem (vgl. Kapitel 8) noch nicht feststand, inwieweit die Ergebnisse des Projektes *ROMANSEVAL*[1] für eine Endauswertung und einen Vergleich des neu erstellten Maßes unter Einbeziehung der neuen korpusbasierten Konzeptknoten zur Verfügung stehen, mußte ohne Kenntnis der Vergleichsmöglichkeiten eine Lexie (bei der vorliegenden Auswertung *maison*) ausgewählt werden. Da das Projekt *ROMANSEVAL* schließlich nicht für die Auswertung verwendet werden konnte, kann dieser Teil der Arbeit (die Erweiterung des semantischen Netzes) im konkreten Disambiguierungsvorgang nicht ausgewertet werden.

[1] Vgl. auch Véronis und Ide 1998.

Daher wird in einem ersten Abschnitt beispielhaft gezeigt, welche Werte das neue Zusammenhangsmaß bei der Berechnung des Zusammenhangs zur Verfügung stellt. Eine detailliertere Auswertung kann nur auf einer größeren Datengrundlage innerhalb eines Vergleichs mit anderen Lesartendisambiguierungsmodellen durchgeführt werden. Wegen des Fehlens eines französischen Korpus mit einer *WN*-kompatiblen semantischen Annotation ist dies zurzeit nicht möglich.

In Abschnitt 11.2 wird dann auf der Grundlage von Rubenstein und Goodenough 1965 ein Vergleich des entwickelten Maßes mit den anderen Maßen vorgenommen, der sich ausschließlich auf die ursprüngliche Netzstruktur bezieht. Damit kann festgestellt werden, ob das neue Maß innerhalb der alten Struktur den guten Resultaten der anderen Maße entspricht und wie gut es mit den vergebenen Zusammenhangswerten der menschlichen Sprecher korreliert.

11.1 Vergleich der Werte für das neue Maß

In einem der Beispiele wird der Zusammenhang durch das neue Maß nicht verstärkt, da die originale Netzstruktur einen kürzeren Pfad vorgibt. Das zweite Beispiel verkürzt durch die neuen Knoten den Pfad und verstärkt so den Zusammenhangswert, und im dritten Beispiel wird gezeigt, daß durch die Verknüpfung von Subhierarchien, die einen sehr hochangesiedelten gemeinsamen Ahnenknoten haben, sich durch das neue Zusammenhangsmaß Verbindungen ergeben, die durch die Korpusuntersuchungen belegt sind, sich aber wegen der fehlenden Vergleichsdaten (ähnlich den Untersuchungen für englische Lexiepaare von Rubenstein und Goodenough) noch nicht überprüfen lassen. Eine Einbindung in ein konkretes Disambiguierungsprogramm (wie im Projekt *ROMANSEVAL*) könnte für diese Kategorie von neuen Zusammenhängen interessante Ergebnisse liefern.

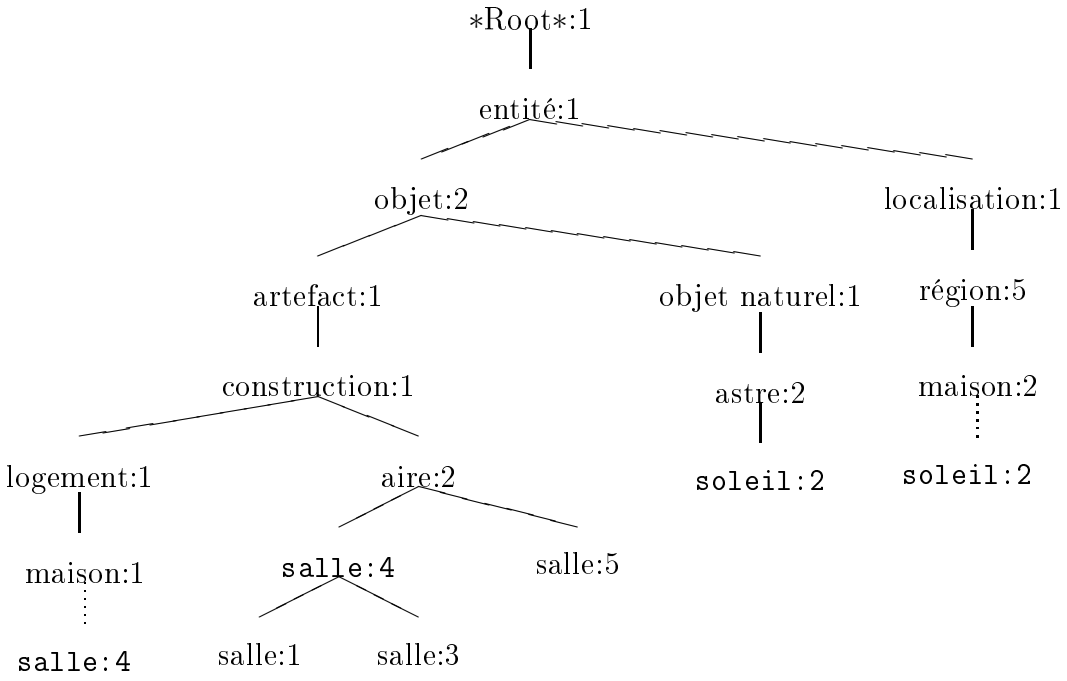
11.1.1 Zusammenhang von *salle:4* und *soleil:2*

Hier werden die folgenden beiden Konzeptknoten verglichen:
[*salle:4*, *pièce:8*:] <an area within a building enclosed by walls and floor and ceiling; „the rooms were very small but they had a nice view“>
[*soleil:2*] ist als neuer Knoten hinzugefügt worden, daher kann hier keine Glosse angegeben werden.[2]

[2]Siehe Abschnitt 9.6.

In diesem Fall ist der durch das neue Maß ausgewertete Verbindungspfad nicht kürzer als der netzimmanente Pfad. Bei der Einbeziehung der neuen Knoten ist der Pfad zehn Kanten lang, wobei die beiden neuen Kanten jeweils mit dem Wert 1 gewichtet werden. Damit verkürzt sich der Weg um den Wert $0.\bar{6}$.

Bei der ausschließlichen Verwendung der netzimmanenten Verbindungen werden sieben Kanten durchlaufen, der gemeinsame Ahnenknoten liegt eine Ebene tiefer [*objet:2*] als derjenige des anderen Pfades [*entité:1*]. Damit hat der erste auch einen höheren Informationsgehalt (1,4561 gegenüber 0,8242). Bei der Auswertung durch das neue Maß werden die verschiedenen Pfadverläufe verglichen und in diesem Fall zugunsten des netzimmanenten entschieden.



Auffällig sind hier die großen Unterschiede bei den Werten der unterschiedlichen Maße. Das Kantenmaß verteilt einen hohen Wert (auch bedingt durch die geringe Tiefe der jeweiligen Konzeptknoten: [*soleil:2*] ist ein Blattknoten mit der Hierarchietiefe 6, [*salle:4*, *pièce:8*:] liegt in der siebten Ebene, die imaginäre Ebene eingerechnet). Die vier Maße, die die Verbindungslänge unberücksichtigt lassen und den unterschiedlichen Informationsgehalt betrachten, vergeben Werte im Intervall [0.3, 0.5], Resniks Wert ist ähnlich, wegen des Kurvenverlaufs ist dieser Wert nicht zu normalisieren.

Durch den langen Verbindungspfad vergeben die anderen Maße sehr niedrige Werte. Hier wäre die Einbindung in einen Lesartenalgorithmus und der Vergleich der Ergebnisse interessant.[3]

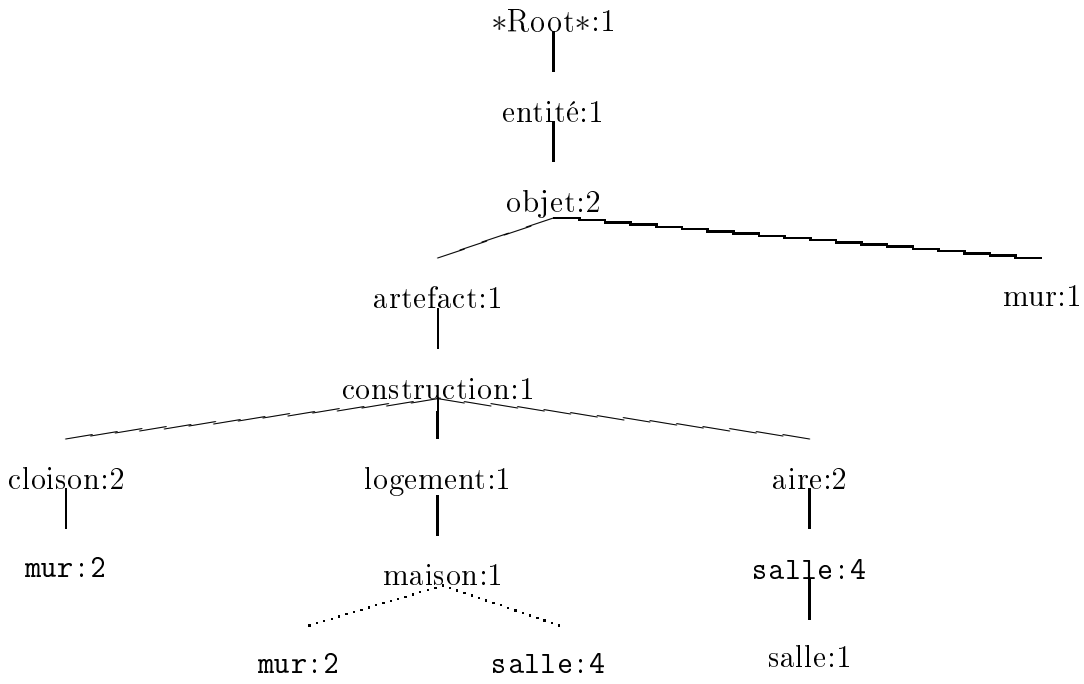
Knoten	**Kanten**	**HSO**	**LCH**	**RES**	**JCN**	**LIN**	**WUP**	**H**
0,1250	0,7812	0,0000	0,4000	1,4561	0,3448	0,1818	0,4615	0,1364

11.1.2 Zusammenhang von *salle:4* und *mur:2*

In diesem Beispiel sind verschiedene Verbindungspfade möglich. Ohne die Einbeziehung der neuen Knoten ist der Verbindungsweg vier Kanten lang und geht über den gemeinsamen Ahnenknoten [*construction:1*]. Bei der Einbeziehung der neuen Kantenverbindung geht der Pfad nur über zwei Kanten und der kleinste gemeinsame Ahnenknoten ist [*maison:1*]. Dabei sind die Kanten auch jeweils gewichtet: Die Verbindung mit [*mur:1*] wird mit $g = 2$, die Verbindung mit [*salle:1*] mit $g = 1$ berechnet, damit geht die Verbindung nur über einen Pfad mit der verkürzten Länge $|\varphi| = 1$.

Das neue Maß wird bei der Berechnung die Pfadlänge vergleichen und sich zugunsten des kürzeren Pfades mit dem niedrigeren Ahnenknoten entscheiden. Es wird dadurch ein sehr hoher Zusammenhangswert zugeordnet.
[*mur:2*:] <a partition with a height and length greater than its thickness; used to divide or enclose or support>
[*salle:4*, *pièce:8*:] <an area within a building enclosed by walls and floor and ceiling; „the rooms were very small but they had a nice view“>

[3] Die jeweiligen Abkürzungen stehen für die vorgestellten Maße: *Knoten* und *Kanten* beziehen sich auf die einfachen knoten- bzw. kantenzählenden Maße, *HSO* steh für das Maß von Hirst und St.-Onge, *LCH* steht für Leacock und Chodorow, *RES* steht für das Maß von Resnik, *JCN* für Jiang und Conrath, *LIN* für das Maß von Lin, *WUP* für Wu und Palmer und *H* für das in dieser Arbeit vorgestellte Zusammenhangsmaß.

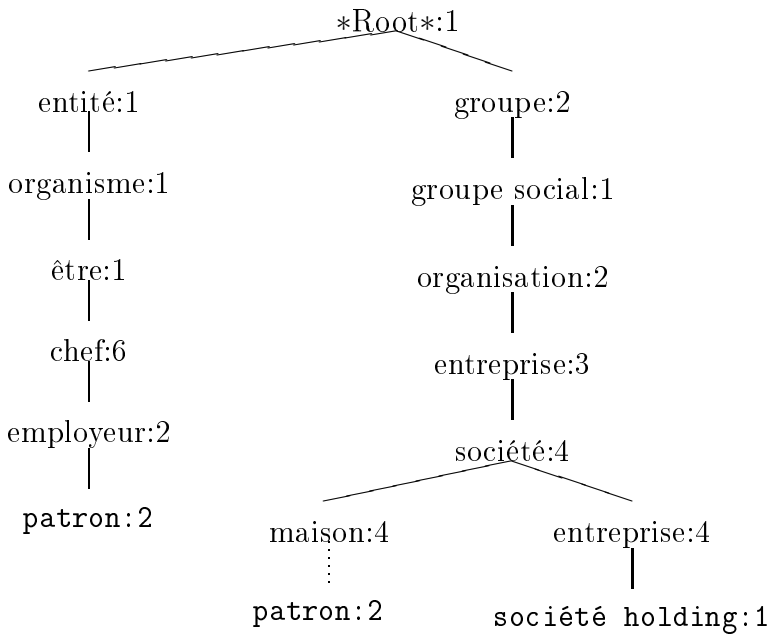


Die anderen Maße, die den Verbindungspfad unberücksichtigt lassen, vergeben Werte im Intervall [0.5,0.7], auch Resnik vergibt einen verhältnismäßig hohen mittleren Wert. Ein Vergleich mit menschlichen Sprechern und die Einbindung in einen Lesartenalgorithmus könnten die Werte in ihrer Anwendungsperformanz beurteilen.

Knoten	**Kanten**	**HSO**	**LCH**	**RES**	**JCN**	**LIN**	**WUP**	**H**
0,2000	0,8750	0,2500	0,5356	3,9590	0,6847	0,5566	0,7143	0,8750

11.1.3 Zusammenhang von *patron:2* und *société holding:1*

In diesem Beispiel wird ein Zusammenhang, der – durch die Konstruktion des semantischen Netzes bedingt – netzimmanent nicht groß ist, durch die Hinzufügung der statistisch begründeten neuen Verbindung erhöht.
[*patron:2*:] <a person responsible for hiring workers; „the boss hired three more men for the new job“>
[*société holding:1*:] <a company with controlling shares in other companies>
Durch die Abbildung des Maßes von Jiang und Conrath in das ungefähre Intervall [0,1] kommt hier ein negativer Zusammenhangswert zustande, der mit dem Wert



0 gleichgesetzt wird. Das Durchlaufen des Kopfknotens und der lange Verbindungspfad führt für Hirst-St.Onge, Lin und Resnik zum Zusammenhangswert 0. Wu-Palmer und das einfache Knotenmaß vergeben ebenfalls Werte nahe an 0. Der höhere Wert von Leacock-Chodorow ist durch den relativ flachen Kurvenverlauf zu erklären, der für Verbindungspfade der Länge $|\varphi| < 16$ einen Zusammenhangswert > 0.2 vergibt.

Lediglich das Kantenmaß vergibt hier im Vergleich zu den anderen Maßen einen hohen Wert, da die Verbindungslänge (13 Kanten) in Relation zur maximalen Verbindungslänge (32) im gesamten Netz gesetzt wird.

Knoten	**Kanten**	**HSO**	**LCH**	**RES**	**JCN**	**LIN**	**WUP**	**H**
0,0714	0,5938	0,0000	0,2385	0,0000	-0,0118	0,0000	0,1333	0,5419

Besonders für dieses Vergleichspaar wäre eine Überprüfung durch eine Testreihe mit menschlichen Sprechern interessant, da diese Verknüpfung, die ein Element der Subhierarchie, die von [*organisme:1*, *vie:11*, *être:2*, *forme de vie:1*] dominiert wird, mit einem Element einer anderen Subhierarchie ([*groupe social:1*]) in den Vergleichen von Rubenstein und Goodenough 1965 (siehe folgender Abschnitt) nicht vorkommt.

11.2 Vergleich mit menschlichen Sprechern

Ein direkter Vergleich mit Werten, die von französischen Muttersprachlern stammen, ist wegen des Fehlens dieser Daten für die französische Sprache nicht möglich. Für die englische Sprache wurden von Rubenstein und Goodenough und Miller und Charles Vergleichswerte im Intervall [0,4] von menschlichen Sprechern gesammelt, die als Vergleichswert von Budanitsky und Hirst verwendet wurden.[4]

Um die Performanz des neu entwickelten Maßes zu bestimmen, sollen nun diese Vergleichswerte für die französische Sprache übernommen werden. Ein Vergleich mit den Werten, die Budanitsky angibt,[5] ist nur über viele Umwege möglich und kann daher insgesamt nur einen ungefähren Anhaltspunkt geben. Die von Rubenstein und Goodenough erarbeiteten Werte für 65 Substantive werden in der Auswahl von Miller und Charles auf die Möglichkeit der Übertragung in die französische Sprache überprüft.

Dort, wo die Maße von Lin, Leacock-Chodorow und Jiang-Conrath in ihrer Berechnung auf dieselben Konzeptknoten zurückgreifen (Rubenstein und Goodenough arbeiteten mit nicht disambiguierten Lexien, wegen der unterschiedlichen Übertragungen kann dies leider nicht übernommen werden), wird nach der Entsprechung im französischen Teilnetz gesucht.

Der stark kritisierte Punkt, daß das französische Teilnetz durch die Übersetzung des Netzes *WN*-1.5 entstand, birgt hier den Vorteil, daß die Werte der menschlichen Annotatoren annähernd übernommen werden können. Einen wirklichen Performanztest auf grundständig französischer Basis ersetzt dieser Test auf keinen Fall. Es folgt die Liste der von Budanitsky getesteten Paare:

[4] Vgl. Rubenstein und Goodenough 1965 und Miller und Charles 1991. Vgl. auch Abschnitt 5.12. Diese Werte der Zusammenhangsmaße sind wegen der veränderten Netzstruktur in *WN*-2.0 nicht direkt zu übernehmen.

[5] Vgl. Budanitsky 1999, 37.

car - automobile	bird - cock	coast - hull
gem - jewel	bird - crane	forest - graveyard
journey - voyage	tool - implement	shore - woodland
boy - lad	brother - monk	monk - slave
coast - shore	crane - implement	coast - forest
asylum - madhouse	lad - brother	lad - wizard
magician - wizard	journey - car	chord - smile
midday - noon	monk - oracle	glass - magician
furnace - stove	cemetery - woodland	noon - string
food - fruit	food - rooster	rooster - voyage

11.2.1 Erste Auswertung

Von den oben aufgeführten 30 Paaren, die Budanitsky vergleicht, werden 11 von allen Zusammenhangsmaßen während der Berechnung auf eindeutige *WN*-2.0 Lesarten zurückgeführt. Dabei wird von den Zusammenhangsmaßen jeweils der bestmögliche Wert innerhalb der möglichen Kombinationen berechnet. Die für diesen Wert verwendeten Lesarten lassen sich über den *InterLingual-Index* eindeutig französischen Lesarten zuordnen.

Die Zahl vor der Klammer ist die Lesart in *WN*-2.0, die bei der Berechnung der Werte mit `similarity.pl` zugrunde liegt. Die Zahl in der Klammer ist die abweichende Lesart in *WN*-1.5, die bei der Übertragung in das französische Netz verwendet wurde, da sich größtenteils die Lesarten, aber nicht der Inhalt des *Synsets* und die Glosse im Übergang von der Version 1.5 nach 2.0 verändert haben.

Drei Paare befinden sich im englischen Netz – wie auch im französischen – in demselben *Synset*:

car:1(2) - automobile:1(1): <*automobile:1*, *auto:1*, *voiture:2*> (4-wheeled; usually propelled by an internal combustion engine; „he needed a car to get to work“)

gem:5(1) - jewel:1(1): <*bijou:1*, *pierre précieuse:1*> (a precious or semiprecious stone used in jewelry)

magician:2(2) - wizard:2(2): <*sorcier:1*, *magicien:2*>

Die restlichen Paare werden auf die jeweils durch den EQ-Link in *EWN* verknüpften Konzeptknoten im französischen Teilnetz abgebildet.

asylum:2(1): <*asile de fous:1*> (a hospital for mentally imcompetent or unbalanced person

madhouse:1(1): <*asile de fous:2*> (pejorative terms for an insane asylum)

coast:1(1): <*côte:5*, *littoral:1*> (keine Glosse)

shore:1(1): <*littoral:2*, *bord:13*, *côte:4*> (keine Glosse)

implement:1(1): <*outil:1*, *instrument:2*> (a piece of equipment or tool used to effect an end)

tool:1(2): <*instrument:4*, *outil:4*> (an implement used in the ractice of a vocation)

boy:1(3): <*garçon:2*> (a young male person; „the baby was a boy“; „she made the boy brush his teeth every night“)

lad:2(1): <*gars:1*> (a male child, a familiar term of address to a boy)

rooster:1(1): <*coq:1*> (adult male chicken)

food:1(1): <*nourriture:4*, *aliment:4*, *nutriment:1*, *élément nutritif:1*> (any substance that can be metabolized by an organism to give enery and build tissue)

woodland:1, timberland:1, timber:4, forest:2: <*forêt:1*> (land that is covered with trees and shrubs)

shore:1(1): <*littoral:2*, *bord:13*, *côte:4*> (keine Glosse)

monk:1(1): <*moine:1*> (a male religious living in a cloister and devoting himself to contemplation and prayer and work)

oracle:1(3): <*prophète:1*, *prophétesse:1*> (an authoritative person who divines the future)

cemetery:1: <*cimetière:2*> (a tract of land used for burials)

Damit werden für die folgenden Paare die Werte ermittelt, die im französischen Netz durch die Zusammenhangsmaße von Jiang-Conrath, Lin und Leacock-Chodorow sowie dem neu erstellten Maß zur Verfügung gestellt werden. Daneben werden die Werte gestellt, die von den menschlichen Sprechern dem jeweils englischen Pendant (mit den oben erwähnten Vorbehalten) gegeben wurden.

```
voiture:2-automobile:1      côte:5-littoral:2       forêt:1-littoral:2
bijou:1-pierre précieuse:1  outil:1-instrument:4    moine:1-prophète:1
sorcier:1-magicien:2        garçon:2-gars:1         cimetière:2-forêt:1
coq:1-nourriture:1          asile de fous:1-asile de fous:2
```

Da diese Knoten nicht durch die Erweiterung auf der Grundlage der korpusstatistischen Auswertung neue Verbindung erhalten haben, kann die Auswirkung dieser Komponente nicht untersucht werden. Daher wird für das neu erstellte Maß die Variable g durchweg den Wert 3 annehmen.

In der folgenden Tabelle sind die Werte für die Zusammenhangsmaße zusammengestellt. Der Vergleichswert, der von Miller-Charles bei der menschlichen Annotation berechnet wurde, wird zur besseren Vergleichbarkeit normalisiert, da der Ursprungswert im Intervall [0,4] liegt.[6]

Graphien	**MC**	**HSO**	**LCH**	**JCN**	**LIN**	**H**
voiture:2 automobile:1	0,9800	1,0000	1,0000	1,0000	1,0000	1,0000
bijou:1 pierre précieuse:1	0,9600	1,0000	1,0000	1,0000	1,0000	1,0000
garçon:2 gars:1	0,9400	0,2500	0,8000	0,9104	0,9029	0,8746
côte:5 littoral:2	0,9250	0,2500	0,8000	0,9451	0,9426	0,9131
asile de fous:1 asile de fous:2	0,9025	1,0000	0,8000	0,9653	0,9645	0,9344
sorcier:1 magicien:2	0,8750	1,0000	1,0000	1,0000	1,0000	1,0000
outil:1 instrument:4	0,7375	0,2500	0,8000	0,9528	0,9109	0,8825
moine:1 prophète:1	0,2750	0,0000	0,4000	0,2790	0,2621	0,2048
cimetière:2 forêt:1	0,2375	0,0000	0,3356	0,0228	0,0000	0,0000
coq:1 nourriture:1	0,2225	0,0000	0,2385	0,1679	0,0901	0,0535
forêt:1 littoral:2	0,1575	0,1250	0,4830	0,2234	0,1579	0,1332

[6]Realisiert wurde die Auswertung mit dem Perlprogramm `endvergleich.pl`, siehe Anhang A.5. Dort finden sich auch die weiteren Änderungen an den Paketen, die für die bessere Vergleichbarkeit (Abbildung ins Intervall [0,1]) vorgenommen wurden.

Diese Auswertung wird übersichtlicher, wenn die Korrelation der einzelnen Zusammenhangsmaße mit den Werten der menschlichen Annotierer berechnet wird.

Zusammenhangsmaße	**Abweichung**
Jiang/Conrath	0,9225
Lin	0,9220
H	0,9145
Leacock/Chodorow	0,8928
Hirst/St.Onge	0,7361

Wie schon gut zu sehen ist, sind die beiden Maße von Hirst-St.Onge und Leacock-Chodorow auch in dieser Netzstruktur im Vergleich zu Lin und Jiang-Conrath nicht so gut.[7] Das Maß von Jiang-Conrath bekommt auch hier den geringsten Abweichungswert, wobei bei dieser Untersuchung die Differenz zur Abweichung von Lin nicht so deutlich ist.

Das im Zusammenhang mit dieser Arbeit erstellte Maß erhält damit auf der ursprünglichen Netzstruktur insgesamt einen Korrelationskoeffizenten, der sich durchaus mit denen der beiden besten Maße vergleichen läßt.

11.2.2 Wertung über alle Substantive

Werden alle von Miller und Charles ausgewerteten Substantive betrachtet, ergibt sich insgesamt ein niedrigeres Korrelationsniveau. Obwohl diese Übertragbarkeit – wie oben erwähnt – nicht ohne Probleme ist, ergibt sich hier die gleiche Reihenfolge der Korrelationskoeffizienten, d.h. in der Annäherung an das menschliche Empfinden von Zusammenhang.

Da sich bei fünf Paaren keine Übersetzung in das französische Netz durchführen läßt, werden insgesamt nur 25 Paare betrachtet. Die Werte von Miller-Charles werden hier unverändert übernommen und für die Berechnung der Korrelation wird eine Normalisierung durchgeführt.[8]

[7] Die untersuchten Elemente von Miller-Charles, bzw. Rubenstein-Goodenough wurden übernommen, so daß sich in der Bewertung der Zusammenhangsmaße, d.h. der Reihenfolge im Vergleich mit den englischen Lexien keine Veränderung zu erwarten war. Vgl. die Ergebnisse von Budanitsky in Budanitsky 1999, 40.

[8] Durchgeführt wurde die Untersuchung mit dem leicht veränderten Perlprogramm `endvergleich.pl`, ausgehend von der Datei `FranzLexeme.dat`.

MC	Prob	Knoten	Kanten	H/SO	Res	Wup	J/C	Lch	Lin	H
voiture:2 automobile:1										
3,9200	0,9992	1,0000	1,0000	1,0000	7,0736	1,0000	1,0000	1,0000	1,0000	1,0000
bijou:1 pierre précieuse:1										
3,8400	1,0000	1,0000	1,0000	1,0000	10,1181	1,0000	1,0000	1,0000	1,0000	1,0000
croisière:1 voyage:2										
3,8400	0,9992	0,5000	0,9688	0,2500	7,1223	0,9333	0,8502	0,8000	0,8262	0,8004
garçon:2 gars:1										
3,7600	0,9998	0,5000	0,9688	0,2500	8,3263	0,9231	0,9104	0,8000	0,9029	0,8746
côte:5 littoral:2										
3,7000	0,9999	0,5000	0,9688	0,2500	9,0195	0,9231	0,9451	0,8000	0,9426	0,9131
asile de fous:1 asile de fous:2										
3,6100	0,9999	0,5000	0,9688	1,0000	9,4249	0,9474	0,9653	0,8000	0,9645	0,9344
sorcier:1 magicien:2										
3,0500	0,9996	1,0000	1,0000	1,0000	7,8155	1,0000	1,0000	1,0000	1,0000	1,0000
calorifère:1 cuisinière:3										
3,1100	0,8492	0,1250	0,7812	0,0000	1,8918	0,5333	0,2578	0,4000	0,2031	0,1587
nourriture:1 fruit:2										
3,0800	0,7669	0,1111	0,7500	0,0000	1,4561	0,4286	0,4449	0,3660	0,2078	0,1559
oiseau:1 coq:1										
3,0500	0,9946	0,2000	0,8750	0,2500	5,2277	0,7778	0,7555	0,5356	0,6813	0,5962
outil:1 instrument:4										
2,9500	0,9920	0,5000	0,9688	0,2500	4,8298	0,9231	0,9528	0,8000	0,9109	0,8825
frère:1 moine:1										
2,8200	0,9228	0,1250	0,7812	0,0000	2,5611	0,5333	0,2790	0,4000	0,2621	0,2048
mec:2 frère:1										
1,6600	0,9228	0,1429	0,8125	0,0000	2,5611	0,5714	0,3136	0,4385	0,2717	0,2208
croisière:1 voiture:2										
1,1600	0,0000	0,0625	0,5312	0,0000	0,0000	0,1176	0,1404	0,2000	0,0000	0,0000
moine:1 prophète:1										
1,1000	0,9228	0,1250	0,7812	0,0000	2,5611	0,5333	0,2790	0,4000	0,2621	0,2048
cimetière:2 forêt:1										
0,9500	0,0000	0,1000	0,7188	0,0000	0,0000	0,1818	0,0228	0,3356	0,0000	0,0000
nourriture:1 coq:1										
0,8900	0,5614	0,0714	0,5938	0,0000	0,8242	0,2353	0,1679	0,2385	0,0901	0,0535

MC	Prob	Knoten	Kanten	H/SO	Res	Wup	J/C	Lch	Lin	H
forêt:1 cimetière:2										
0,8400	0,0000	0,1000	0,7188	0,0000	0,0000	0,1818	0,0228	0,3356	0,0000	0,0000
littoral:2 forêt:1										
0,6300	0,7669	0,1667	0,8438	0,1250	1,4561	0,5455	0,2234	0,4830	0,1579	0,1332
moine:1 esclave:6										
0,5500	0,9228	0,2000	0,8750	0,1875	2,5611	0,6667	0,3136	0,5356	0,2717	0,2378
littoral:1 forêt:1										
0,4200	0,7669	0,1429	0,8125	0,0000	1,4561	0,5000	0,1685	0,4385	0,1490	0,1211
mec:2 crack:2										
0,4200	0,9228	0,2000	0,8750	0,1875	2,5611	0,6667	0,2790	0,5356	0,2621	0,2293
verre:6 prestidigitateur:2										
0,1100	0,5614	0,0909	0,6875	0,0000	0,8242	0,2857	0,1602	0,3081	0,0894	0,0614
coq:1 voyage:2										
0,0800	0,0000	0,0588	0,5000	0,0000	0,0000	0,1111	0,1380	0,1825	0,0000	0,0000
midi:1 cordon:2										
0,0800	0,0000	0,0667	0,5625	0,0000	0,0000	0,1250	-0,0118	0,2186	0,0000	0,0000

Auch hier wird in der Übersicht der Korrelationskoeffizienten deutlich (allerdings auf niedrigerem Niveau, das auch die Methodik hinterfragen läßt), daß Jiang-Conrath, Lin und das hier vorgestellte Maß in der ursprünglichen Netzstruktur am besten mit den menschlichen Werten korrelieren.

Zusammenhangsmaße	Korrelation
Jiang-Conrath	0.6831
Lin	0.6504
H	0.6254
Wu-Palmer	0.6038
Leacock-Chodorow	0.5629
Knoten	0.4438
Einf. Wahrscheinlichkeit	0.3612
Hirst-St. Onge	0.2991
Kanten	0.2746
Resnik	-6.1656

Kapitel 12

Zusammenfassung und Ausblick

12.1 Zusammenfassung

Das geplante Ziel der Erweiterung des assoziativen semantischen Netzes *FrenchWordNet* zur Entwicklung eines korpusgestützten Zusammenhangsmaßes konnte in einer Studie für eine Beispiellexie durchgeführt werden. Dabei stellte sich häufig das Problem des notwendigen manuellen Eingriffs. Allgemein gültige Schwellenwerte konnten nicht verwendet werden, auch mußten bei den verschiedenen Lesarten unterschiedliche Kriterien für die Zusammenstellung der aussagekräftigen Kollokationen angewandt werden. Das entwickelte Zusammenhangsmaß wurde für diese Lexie mit den anderen Zusammenhangsmaßen verglichen.

Da ein Vergleich mit den anderen Maßen unter den Bedingungen des *ROMANSEVAL*-Projekts nicht durchgeführt werden konnte, mußte der Test des Zusammenhangsmaßes innerhalb des Lesartendisambiguierungalgorithmus entfallen. Somit konnte die Annahme, die Lesartendisambiguierung mit dieser Erweiterung des semantischen Netzes durch die korpusbasierte Komponente zu verbessern, nicht überprüft werden.

Da dadurch auch die entsprechenden Daten für einen Vergleich mit französischen Muttersprachlern fehlten, konnten die Ergebnisse nur ohne die Einbeziehung der neuen korpusbasierten Komponente mit den anderen Werten verglichen werden. Es wurde ein direkter Vergleich dieser Zusammenhangsmodelle, die die Assoziationen des Menschen nachahmen, mit Daten aus früheren englischen Untersuchungen (Rubenstein und Goodenough 1965) durchgeführt. Dort ergab sich ein Korrelationswert zu den Vergleichswerten menschlicher Annotierer, der

den beiden Zusammenhangsmaßen, die in anderen Studien als beste Zusammenhangsmaße getestet wurden, sehr nahe kommt. Damit ist das entwickelte Maß eine gute Alternative zu den anderen Maßen.

Ein Schwerpunkt der technischen Arbeit lag auf der Bearbeitung des französischen Teilnetzes *FrenchWordNet*. Die für die erstmalig durchgeführte Transformation in das *WN*-Format benötigte Grundlage wurde wegen technischer und lexikographischer Fehler verbessert, so daß hier erstmals eine gründliche Überprüfung dieser Datenbank vorgenommen wurde.

12.2 Ausblick

Auf der Basis von semantisch annotierten Korpora, die mit den Lesartenunterscheidungen von *FrenchWordNet* verglichen werden können (allerdings nach einer gründlichen Revision der sehr an das englische *WN*-1.5 angelehnte Differenzierung), könnte in einem größeren Rahmen eine Verdichtung des Netzes durch korpusbasierte Verknüpfungen stattfinden. Die Arbeiten mit englischen Korpora (siehe Abschnitt 8.8) geben hier ein Beispiel, wie die vorhandenen Strukturen von *WN* sinnvoll mit Korpora kombiniert werden können. Allerdings ist noch kein solches französisches Korpus vorhanden.

Die Automatisierung des Verfahrens ist wegen des häufigen manuellen Eingriffs (bedingt besonders durch die Bedingungen des verwendeten Korpus *Frantext*) nur bei einigen Schritten möglich. Auf der Grundlage eines großen gewichteten Korpus könnten mit Einschränkung weitere Automatisierungen vorgenommen werden, um einheitliche Schwellenwerte (unabhängig von der betrachteten Lesart und der Anzahl der Belege) zu erhalten.

Anhang A

Programme

Hier sind die einzelnen – im Lauf der Arbeit erstellten Programme zusammengefaßt und erläutert. In einem ersten Abschnitt (A.1) finden sich alle Programme, die für die Bereitstellung des Untersuchungskorpus aus dem französischen Originalkorpus verwendet wurden. Im zweiten Teil (A.2) befinden sich die Programme für die Datenbanktransformation und im letzten (A.3 bis A.5) sind die für die Auswertung der zusammengestellten Daten erstellten Programme aufgeführt inklusive der verwendeten Ausgangsdateien.

A.1 Vom Korpus zum Untersuchungskorpus

Hier wird dargestellt, wie aus dem online verfügbaren französischen Korpus *Frantext* ein Subkorpus erstellt wird, das die Basis für die weitere Untersuchung bildet.

Dabei werden die Programme (Shellskripte und Perlprogramme) erläutert, die für die Bearbeitung des Korpus erstellt werden und die entstandenen Formate (`.mvf` und `.son`) werden beschrieben. Am Ende dieser Korpusauswertung stehen die Dateien, die für die statistische Auswertung des Korpus verwendet werden.

A.1.1 Formatierung

Das Untersuchungskorpus (ein Ausschnitt aus dem *Frantext*-Korpus) umfaßt 87 hauptsächlich literarische Texte aus den Jahren 1980 bis 2000 (6.404.598 *Graphien*)[1] Bei der vorliegenden Untersuchung wurde in diesem Korpus nach

[1] Vgl. Beschreibung des *Frantext*-Korpus, Abschnitt 6.4.

Kookkurrenzen des Substantivs *maison* in einem Satz gesucht: es fanden sich 2.361 Belege.[2]

Die Ausgabe der Korpusbelege durch *Frantext* erfolgt in Hunderter-Schritten im Format `.html`, mit der Möglichkeit, sich die von *Frantext* annotierten Kategorien anzeigen zu lassen (siehe Abbildung A.1). Die weitere Arbeit mit den POS-Angaben wurde durch eine teilweise unkorrekte Annotation erschwert: Es fanden sich nicht-annotierte Elemente und kryptische Annotationen, die die automatische Weiterbearbeitung durch die Programme immer wieder unvollständige Ergebnisse liefern ließen. Hier mußten mehrfach fehlerhafte Stellen manuell verbessert werden.

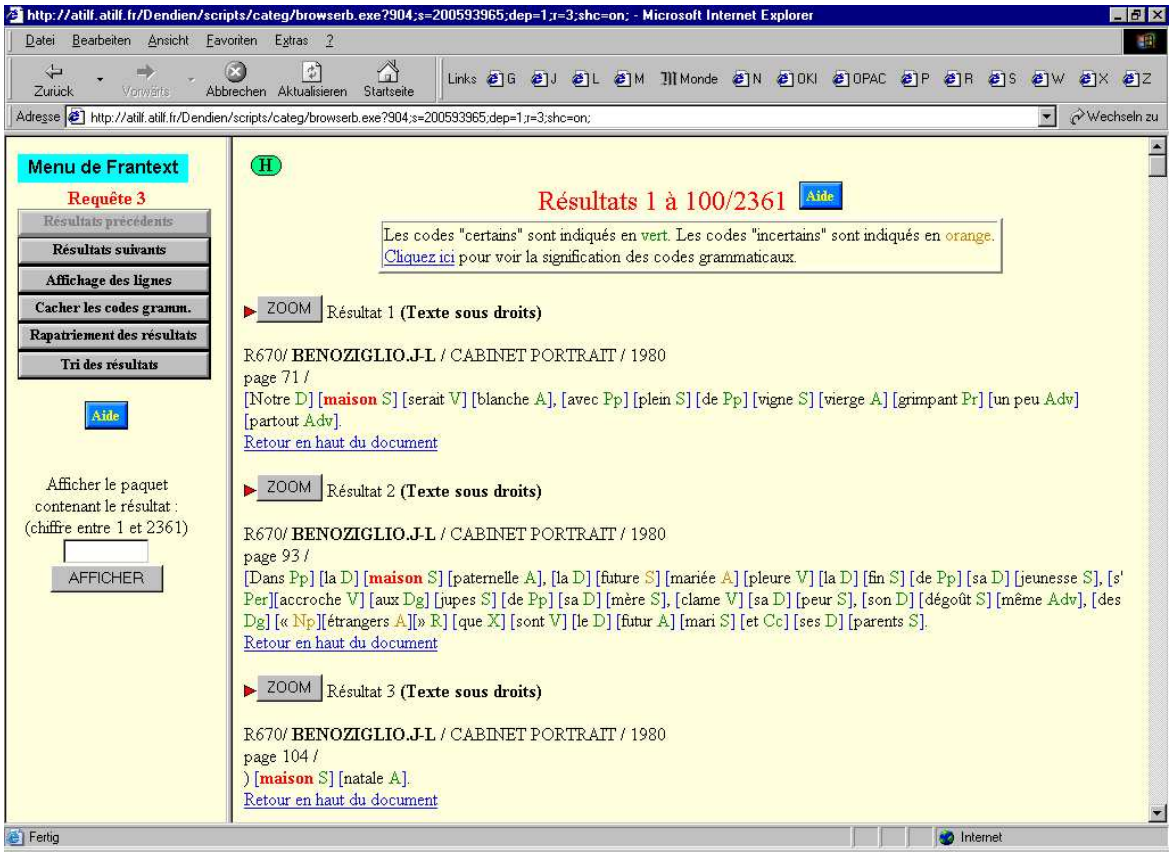


Abbildung A.1: Korpusbelege für das Substantiv *maison* in *Frantext*.

[2]In dieser Untersuchung wurde nur der Singular recherchiert, da *Frantext* bei der Zusammenstellung der Belegstellen für den Singular und den Plural von <maison> aus dem Untersuchungskorpus überfordert war. Es finden sich 409 Belege für *maisons* ohne gleichzeitiges Auftauchen des Singulars im gleichen Satz. In 21 Belegen finden sich Singular und Plural, in diesem Fall wurde nur der Singular berücksichtigt.

Da eine direkte Abspeicherung der Korpusbelege mit den syntaktischen Kategorien von *Frantext* nicht vorgesehen ist, werden die einzelnen `.html`-Seiten per Hand abgespeichert und zu einer großen `.txt`-Datei zusammengefaßt. Dabei befinden sich neben den Korpusbelegen die zusätzlichen (überflüssigen) Angaben der `.html`-Seiten (siehe Abbildung A.2).

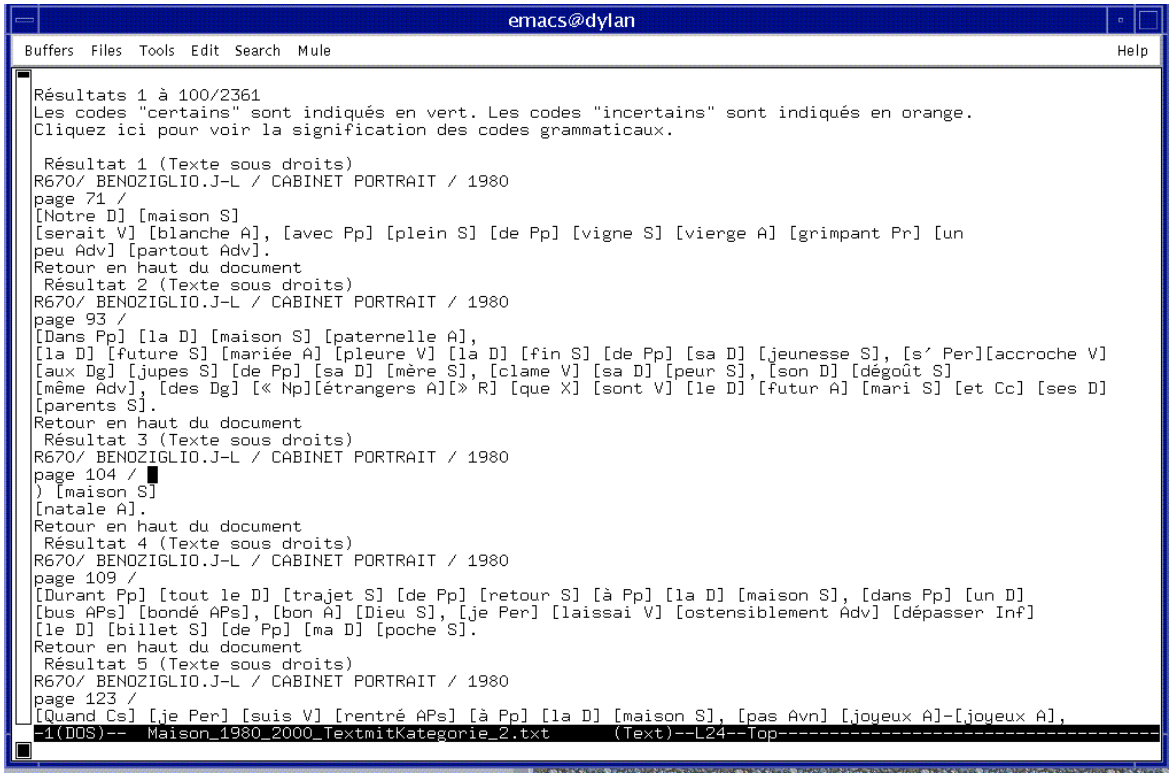


Abbildung A.2: Unveränderte `.txt`-Datei in *Emacs*.

Nach einer kurzen Vorformatierung wird die entstandene `.txt`-Datei mit den Perl-Programmen *Fran2mvf.pl* (Abschnitt A.1.3) und *Fran2son.pl* (Abschnitt A.1.4) in das `.mvf`-Format und das `.son`-Format überführt. Diese beiden Formate bilden die Grundlage für die weitere Untersuchung.

Das `.mvf`-Format bietet ein Format an, das leicht von Perl- und Shellskripten verarbeitet wird. Die Ausgangsdatei `maison_1980_2000Kategorie.txt` wird durch das Perl-Programm *Fran2mvf.pl* in eine Datei umformatiert, bei der pro Zeile eine *Graphie* mit der – von *Frantext* bereitgestellten – POS-Annotation steht.

Ein Ausschnitt von maison_1980_2000Kategorie.txt:

```
Résultat 2360 (Texte sous droits)
S361/ RAMBAUD.P / LA BATAILLE / 1997
page 274 / CHAPITRE VII, Après l'hécatombe
[Périgord Np] [lui Per] [apporta V] [son D] [aide S], [puisqu' Cs]
[il Per] [était V] [revenu APs] [prendre Inf] [ses D] [quartiers S]
[dans Pp] [la D] [maison S] [rose A], [avec Pp] [son D] [gros A]
[valet S] [et Cc] [sa D] [giberne S] [en Pp] [vermeil S] [qui P]
[contenait V] [un D] [nécessaire S] [de Pp] [toilette S], [du Dg]
[gratte S]-[langue S] [aux Dg] [fards S].
Retour en haut du document
```

Nach der Umformung entsteht die Datei maison_1980_2000Kategoriemvf.txt. Es folgt ein Ausschnitt (für eine bessere Übersicht verteilt auf mehrere Spalten):

Périgord	Np	ses	D	et	Cc	,	Pon
lui	Per	quartiers	S	sa	D	du	Dg
apporta	V	dans	Pp	giberne	S	gratte	S
son	D	la	D	en	Pp	-	Pon
aide	S	maison	S	vermeil	S	langue	S
,	Pon	rose	A	qui	P	aux	Dg
puisqu'	Cs	,	Pon	contenait	V	fards	S
il	Per	avec	Pp	un	D	.	Pon
était	V	son	D	nécessaire	S		
revenu	APs	gros	A	de	Pp		
prendre	Inf	valet	S	toilette	S		

Das .son-Format ist eine (für den menschlichen Annotierer) gut lesbare Formatierung, die aus der Datei maison_1980_2000Kategorie.txt für die manuelle Annotation erstellt wird. Ein Ausschnitt aus maison_1980_2000Kategorie.son:

```
2360 - S361 - RAMBAUD.P - LA BATAILLE - 1997 - page 274
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
[Périgord Np] [lui Per] [apporta V] [son D] [aide S], [puisqu' Cs]
[il Per] [était V] [revenu APs] [prendre Inf] [ses D] [quartiers S]
[dans Pp] [la D] [maison S] [rose A], [avec Pp] [son D] [gros A] [valet S]
[et Cc] [sa D] [giberne S] [en Pp] [vermeil S] [qui P] [contenait V]
[un D] [nécessaire S] [de Pp] [toilette S], [du Dg] [gratte S]-[langue S]
[aux Dg] [fards S].
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
```

A.1.2 Annotation

Nach der manuellen Annotation der `*.son`-Datei wird mit dem Perl-Programm *Ann2einzeln.pl* (Abschnitt A.1.5) die Datei in die – den Lesartenkategorien entsprechenden – Dateien aufgespalten. In der `.log`-Datei ist die Anzahl der Belege erfaßt:

```
sensetag '1': 1.562 Belege
sensetag '2': 2 Belege
sensetag '3': 360 Belege
sensetag '4': 49 Belege
sensetag '!!!': 237 Belege
sensetag '===': 133 Belege
sensetag '???': 18 Belege
Summe Belege: 2.361
```

Als Zwischenschritt werden die entstandenen Dateien durch das Perlprogramm *senserename.pl* umbenannt: z.B. `maisonBedeutung1.txt`, `maisonunklar.txt` oder `maisonKompositaBedeutung1.txt`.

Mit dem Perlprogramm *sense2Nomen* (Abschnitt A.1.7) werden die Substantive aus diesen Dateien herausgesucht und mit dem Shellskript *Auswertungson.sh* (Abschnitt A.1.8) gezählt und sortiert. Die folgenden Ausschnitte aus den Dateien geben die häufigsten Substantive in der Umgebung von *maison* wieder – nach Lesarten sortiert.

Ausschnitt aus `maison1_NomenErgebnis.txt`

```
1.562 §1§maison
113 maison
51 porte
46 famille
45 mère
42 jour
41 jardin
41 bois
```

Ausschnitt aus `maison2_NomenErgebnis.txt`

```
maison2_NomenErgebnis.txt
2 §2§maison
2 mars
1 suite
1 soleil
1 scorpion
```

Ausschnitt aus `maison3_NomenErgebnis.txt`

```
360 §3§maison
23 père
17 soir
17 mère
15 vie
15 enfants
13 jour
13 heures
```

Ausschnitt aus `maison4_NomenErgebnis.txt`

```
49 §4§maison
21 disques
4 paris
3 production
3 patron
3 papa
3 mère
```

Ausschnitt aus `maisonunklar_NomenErgebnis.txt`

```
18 §???§maison
6 nord
4 bois
2 verger
2 treilles
2 mousse
2 heures
2 acacias
1 tempête
```

Ausschnitt aus `maisonEigenname_NomenErgebnis.txt`

```
133 §===§maison
25 enfant
15 maison
15 enfants
13 mère
9 parents
8 maman
8 fille
7 jour
7 crèche
```

Ausschnitt aus `maisonKomposita_NomenErgebnis.txt`

```
237 §!!!§maison
19 maîtresse
18 maître
17 campagne
12 enfants
12 édition
11 maison
11 jour
```

Mit diesen Dateien kann eine statistische Auswertung vorgenommen werden. Die unterschiedlichen Wahrscheinlichkeiten, ein bestimmtes Substantiv in einem kleinen Kontextfenster um *maison* in einer der möglichen Lesarten zu finden, können Aufschluß geben auf die Zuweisung von Lesarten bei noch nicht disambiguierten Vorkommen von *maison*.

Es folgen jetzt die Quellcodes der einzelnen Shell-Skripte und Perl-Programme. Die Beispieldateien sind im vorangegangenen Abschnitt vorgestellt worden und Anmerkungen zu den Programmen finden sich im Quellcode.

A.1.3 `Fran2mvf.pl`

Die mit den von *Frantext* zur Verfügung gestellten Kategorien annotierte Datei wird in das `.mvf`-Format konvertiert, bei der sämtliche Annotationen übernommen werden, die Belege zum Korpus allerdings gelöscht werden. Ausgangsdatei ist die leicht veränderte Frantextdatei, die manuell aus *Frantext* kopiert wurde.

Das Programm *Fran2mvf.pl*:

```
#!/usr/local/bin/perl
$Datei = join("", <>);
while($Datei =~ s/(\[.*?)Retour en haut du document//s)
{
    push(@saetze,$1);                    # Beleg wird in Array gepusht
}
foreach(@saetze)
{
    while(s/^\[(.*?) ([A-Z][A-Za-z]*)\]([^\[]*)//)
    {
        $Pon=$3;                         # wird umbenannt (Ponctuation)
        $wort=$1;                        # umbenannt
        $wortart=$2;                     # umbenannt
        $wort=~s/^\*//;                  # Sternchen am Anfang entfernt
        print "$wort\t$wortart\n";       # Ausgabe: Wort \Tab  Wortart
        if(defined $Pon)                 # $3 def: eigene Zeile
        {
            $Pon=~s/\s//g;               # Nichtsatzzeichen entfernt
            unless($Pon eq "")
            {
                print "$Pon\tPon\n";
            }
        }
    }
}
```

A.1.4 `Fran2son.pl`

Aus der aus *Frantext* geholten Datei wird das `.son`-Format generiert, in dem die Belege in ein einheitliches Format gebracht werden, das für die Annotation verwendet wird.

Das Programm *Fran2son.pl*:

```
#!/usr/local/bin/perl -w
my %originalsatz;
```

```
my $i = 0;
my $datei = join("", <>);
while ( $datei =~ s/
    ^.*?R.sultat \s (\d+) [^\n]* \n+   # Resultatsnummer
    ([^\n]*?) \/                      # z.B. R361
    ([^\n\/]*?) \/ ([^\n\/]*?) \/     # Autor, Titel, getrennt
    ([^\n]*?) \n                      # Jahr, Newline
    page \s (\d+) \s \/ [^\n]*? \n    # Seite, Kapitel, Newline
    (.*?) \n Retour\ en\ haut\ du\ document # Satz
    //sx )                           # x, ueber mehrere Zeilen
{
 print "$1 - $2 - $3 - $4 - $5 - page $6\n";
 print "&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&\n";
 print "$7\n";
 print "&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&\n\n";
 $i++;
}
print $i, "\n";
exit;
foreach (sort {$a cmp $b} (keys (%originalsatz))) # Hash sortieren
{
  print "$_\t";
  print $#{ $originalsatz{$_} } + 1, "\t";
  print join(" ", @{ $originalsatz{$_} } );
  print "\n\n";
}
exit;
```

A.1.5 Ann2einzeln.pl

Die annotierte Datei wird durch das folgende Programm in einzelne Dateien aufgesplittet: die einzelnen Lesarten, eine Datei jeweils für die Eigennamen, die Komposita und die unklar gebliebenen Belege.

```
#!/usr/local/bin/perl -w
my %belege;
my $beleg;
```

```
my $i = 0;
my $datei = join("", <>);
while($datei =~ s/(^\d+[^\n]+\n[\&]*\n[^\&]*[\&]*\n\n)//sx)
{
  push(@belege,$1);                # nach Muster suchen, in Hash
}
foreach $beleg (@belege)
{
  $i++;                                          # Zaehler erhoehen
  if ($beleg§ =~ /(\d§+|!!!|\?\?\?|===)/) # Annotation suchen
  {
     my $sensetag = $1;
     push (@{$belege{$sensetag}}, $beleg);# in Array
  }
  else
  {
     print STDERR "Achtung, kein Sensetag in $i! \n";
  }
}
print "Ende bei $i. \n";                 # Zaehler als Test
open( LOG, "> sensetag_log.txt" )
  or die("Kann Log nicht oeffnen.");
my $summe_belege = 0;
foreach $sensetag (sort keys %belege) # in Dateien einlesen
{
    $summe_belege += $#{ $belege{$sensetag}} + 1;
    print LOG "sensetag '$sensetag': ",
              $#{$belege{$sensetag}} + 1," Belege\n";
    open( OUT, "> sensetag_$sensetag.txt")
      or die("Kann Ausgabe nicht oeffnen.");
    print OUT @{ $belege{$sensetag} };
    close OUT;
}
print LOG "Summe Belege:  $summe_belege\n";
close LOG;
```

A.1.6 Senserename.pl

Dieses Perlprogramm benennt die aus *Ann2einzeln* entstandenen Dateien um.

```
#!/usr/local/bin/perl
print "Wie sollen die Dateien umbenannt werden?\n";
$name =<STDIN>;
chomp($name);
rename("sensetag_1.txt", "${name}Bedeutung1.txt");
rename("sensetag_2.txt", "${name}Bedeutung2.txt");
rename("sensetag_3.txt", "${name}Bedeutung3.txt");
rename("sensetag_4.txt", "${name}Bedeutung4.txt");
rename("sensetag_!!!.txt", "${name}Komposita.txt");
rename("sensetag_===.txt", "${name}Eigenname.txt");
rename("sensetag_???.txt", "${name}unklar.txt");
rename("sensetag_log.txt", "${name}log.txt");
```

A.1.7 Sense2Nomen.pl

Aus der mit*Ann2einzeln.pl* und *senserename.pl* entstandenen Datei werden alle Substantive extrahiert und als Datei zusammengefaßt.

```
#!/usr/local/bin/perl -w
my $datei = join("", <>);
my $i = 1;
while ($datei =~ s/\[(\S*?)\sS\]//sx)
{ print $i, "\t", $1, "\n";
  $i++;
}
```

A.1.8 Auswertungson.sh

Sortiert und zählt die Substantive aus den einzelnen Substantivlisten der annotierten Dateien und schreibt sie getrennt in eine Datei `*Ergebnis.txt`.

```
if [ "$1" = "" ]; then     # $1 untersuchtes Wort
  echo "no arg"
  exit
fi
```

```
grep "  S" ${1}.txt |      # annotierte S herauslesen
tee ${1}Substantive.txt |
cut -d "         " -f1 |   # Woerter extrahieren
tr 'A-Z' 'a-z' |           # sortieren
sort|
uniq -c |
sort -nr > ${1}Ergebniszahl.txt
```

A.2 Datenbanktransformation

Das Perl-Programm, das erstellt wurde, um das Datenbankformat von *EWN* in das *WN*-Format zu überführen, wird durch die Kommentare im Quellcode kommentiert. Der Vorgang und die einzelnen Probleme werden im Kapitel 9.5 genauer dargestellt.

A.2.1 `Ewn2wn.pl`

```
#!/usr/bin/perl
use Time::Local;
use POSIX qw(strftime);
##########################
#DIVERSE DEKLARATIONEN:#
##########################
my $day;
my $month;
my $year;
$tm = localtime;
($sec, $min, $stunde, $day, $month, $year) = (localtime)[0..5];
my $dummy = 0;
my %index;                                  # "$lem*$sen" => "$zahl"
 # POINTERVERKNUEPFUNGEN
my %ptr_hash = ( "antonym"                  => "!",
                 "has_holo_madeof"          => "#s",
                 "has_holo_member"          => "#m",
                 "has_holo_part"            => "#p",
                 "has_holo_portion"         => "#p",
                 "has_hyperonym"            => "@",
                 "has_hyponym"              => "~",
                 "has_mero_madeof"          => "%s",
                 "has_mero_member"          => "%m",
                 "has_mero_part"            => "%p",
                 "has_mero_portion"         => "%p",
                 "involved"                 => "§",
                 "involved_agent"           => "§a",
                 "involved_instrument"      => "§i",
                 "involved_location"        => "§l",
                 "role"                     => "§-",
                 "role_agent"               => "§-a",
                 "role_instrument"          => "§-i",
                 "role_location"            => "§-l");
my $variants_schalter = 0;
```

```
35 my $internallinks_schalter = 0;
   my $eqlinks_schalter = 0;
   my $literal_schalter = 0;
   my $antonym_schalter = 0;
   my $reversed_schalter = 0;
40 my $eqsyn_schalter = 0;
   my $eqgen_schalter = 0;
   my $eqmet_schalter = 0;
   my $externalinfo_schalter = 0;
   my $features_schalter = 0;
45 my $targetconcept_schalter = 0;
   my $targetili_schalter = 0;
   my $sourceid_schalter = 0;
   my $variant_schalter = 0;
   my @alles;                          # ENTHAELT DATENBANKEINTRAEGE
50 my $k = 0;                          # $k ZAEHLT DIE EINTRAEGE
   my $z = 0;                          # ZAEHLT GEFUNDENEN MATCH
   my $lauf = 0;                       # VERGLEICHSVARIABLE
   my $zahl = 0;
   my %ebene_1 = ("LEMMATA"   => "",   # ARRAY MIT LEMMATA
55                "POS"       => "",   # POS ("n")
                  "LINKS"     => "",   # HASH MIT LINKS
                  "EQ_LINKS" => "",    # HASH MIT EQLINKS
                  "TEXT_KEY" => "");   # TEXTKEY
   my $pos = 0;
60 my $synset_offset = 0;               # 8DIGIT INT
   my @lemmata = 0;                    # ELEMENTE DES SYNSETS
   my $variantenzaehler = 0;           # ZAEHLT DIE VARIANTEN
   my %lemma = ("LEM"      => "",      # VARIANTENHASH
                "SEN"      => "");     # LEM: LEMMA, SEN: SENSE
65 my $lem = 0;                         # LEMMA-LITERAL
   my $sen = 0;                        # SEN (BEDEUTUNGSNUMMER)
   my %links = ();                     # HASH MIT LINKS
   my $ptr_lemma = 0;                  # VERKNUEPFTES LEMMA
   my $ptr_lemma_schalter = 0;
70 my $linkkey = 0;
   my $textkey = 0;
   my $sourcevariant = 0;              # SOURCE-VARIANTE
   my $source_offset = 0;              # OFFSET SOURCEVARIANTE
   my $targetvariant = 0;              # TARGET-VARIANTE
75 my $target_offset = 0;               # OFFSET TARGETVARIANTE
   my $symbolliteral = 0;              # FORM DER VERKNUEPFUNG
   my $symbolzeichen = 0;              # ZEICHEN DER VERKNUEPFUNG
   my $symbol_schalter = 0;
   my $linksense = 0;                  # BEDEUTUNG DES VERK. LEMMA
```

```
my %eqlinks = ("SYNONYM" => "",      # HASH FUER EQ-LINKS
               "GENERALIZATION" => "",
               "METONYM" => "");
my $eqsynonym_offset = 0;            # OFFSET DES EQ-LINKS
open(LOG, ">Hauptprogrammlog.txt")
    or die ("Kann Logdatei nicht oeffnen.");
printf LOG ("Transformation vom %02d.%02d.%04d,
                          %02d:%02d:%02d Uhr.\n",
            $day, $month+1, $year+1900, $stunde, $min, $sec);

#####################################
#BEGINN: ERSTELLUNG SUBSTANTIV.EWN#
#####################################
print STDOUT "Einlesen der Daten.\n";
my $datei = join("", <>);
open(OUT, "> Substantive.ewn")
    or die("Kann Ausgabe Substantive.ewn nicht oeffnen.");
print STDOUT "Aus der Datenbank werden die
              Substantive herausgeholt.\n";
print STDOUT "Das kann dauern...\n";
while ($datei =~ s/(^0 \@\d+\@.*?1 PART_OF_SPEECH
                    \"n\".*?\n)\n//s)
{
    print OUT $1;
}
close OUT;
print STDOUT "Datei Substantive.ewn wurde erstellt.\n";
#####################################
#ENDE: ERSTELLUNG SUBSTANTIV.EWN#
#####################################

###################
#ERSTELLUNG INDEX#
###################
print STDOUT "Einlesen fuer internen Index.\n";
open(IN, "<Substantive.ewn")
    or die ("Kann Substantive.ewn nicht wieder oeffnen.\n");
while (<IN>)
{
    if ($_ =~ /^(0)\s\@(\d+)\@\sWORD_MEANING/)
    {
        $k++;
        if($1 >= $lauf)
        {
            $lauf = $1;
```

```
125         }
            else
            {
                $zahl = 0;
                %ebene_1 = ();
130             $synset_offset = 0;
                $variantenzaehler = 0;
                $lauf = $1;
            }
            $zahl = $2;                                     # OFFSETNUMMER
135         $synset_offset = offsetausgabe($zahl);     # 8DIGIT INT
        }
    ################################EBENE 1: ELEMENTE DES SYNSETS
        elsif ($_ =~/\s*?(1) VARIANTS/)
        {
140         if($1 >= $lauf)
            {
                $lauf = $1;
            }
            else
145         {
                @lemmata = ();
                $lauf = $1;
            }
            $variants_schalter = 1;
150     }
    ################################EBENE 2: LITERAL
        elsif ($_ =~/\s*?(2) LITERAL \"(.*?)\"/
               && $variants_schalter == 1)
        {
155         if($1 >= $lauf)
            {
                $lauf = $1;
            }
            else
160         {
                $lem = 0;
                %lemma = ();
                $lauf = $1;
            }
165         $lem = $2;                                  # LITERAL IN $lem
            $alles[$zahl]{'LEMMATA'}[$variantenzaehler]{"LEM"}
                                                            = $lem;
            $literal_schalter = 1;
        }
```

```
################################EBENE 3: SENSE
    elsif ($_ =~/\s*?(3) SENSE (\d+)/
           && $literal_schalter == 1)
    {
        if($1 >= $lauf)
        {
            $lauf = $1;
        }
        else
        {
            $sen = 0;
            $lauf = $1;
        }
        $sen = twodigit($2);                    # BEDEUTUNGSNUMMER
        $alles[$zahl]{'LEMMATA'}[$variantenzaehler]{"SEN"}
                                                    = $sen;
        $variantenzaehler++;
        if (defined ($index{"$lem\*$sen"}))# EINTRAG INDEX-HASH
        {
        }
        else
        {
            $index{"$lem\*$sen"} = $synset_offset;
        }
    }
}
print LOG "Index erfolgreich in Speicher eingelesen:\n";
print LOG "Es wurden " .$k." Datenbankeintraege gelesen.\n";
#########################
#ENDE ERSTELLUNG INDEX#
#########################
print STDOUT "Einlesen fuer internen Index fertig.\n";
print STDOUT "Erstellung: Index_numerisch und
                          Index_alphabetisch.\n";
################
# INDEXAUSGABE#
################
open (NUMERISCH, ">Index_numerisch.ewn")
    or die("Kann Datei Index_numerisch.ewn nicht oeffnen.");
open (ALPHABETISCH, ">Index_alphabetisch.ewn")
    or die("Kann Datei Index-alphabetisch.ewn nicht oeffnen.");
$a = 0;
$b = 0;
foreach (sort {$index{$a} cmp $index{$b}} (keys(%index)))
{ if (defined keys(%index))
```

```
215   {
          print NUMERISCH $_. ": ". $index{$_}. "\n";
      }
      else
      {
220   }
  }
  print STDOUT "Index_numerisch.ewn erfolgreich erstellt.\n";
  print LOG "Index_numerisch.ewn erfolgreich erstellt.\n";
  close NUMERISCH;
225 foreach (sort {$a cmp $b} (keys(%index)))
  { if (defined keys(%index))
    {
          print ALPHABETISCH $_. ": ". $index{$_}. "\n";
    }
230 else
    {
    }
  }
  print STDOUT "Index_alphabetisch.ewn erfolgreich erstellt.\n";
235 print LOG "Index_alphabetisch.ewn erfolgreich erstellt.\n";
  close ALPHABETISCH;
  close IN;
  ############################
  #ENDE: INDEXERSTELLUNG#
240 ############################

  ############################
  #BEGINN: HAUPTPROGRAMM#
  ############################
245 print STDOUT "Beginn der Auswertung.\n";
  $k = 0;
  open(IN, "<Substantive.ewn")
      or die ("Kann Substantive.ewn nicht wieder oeffnen.\n");
  while (<IN>)
250 {
      if ($_ =~ /^(0)\s\@(\d+)\@\sWORD_MEANING/)
      {
          $z++;
          $k++;
255       if($1 >= $lauf)
          {
              $lauf = $1;
          }
          else
```

```
260         {
                $zahl = 0;
                %ebene_1 = ();
                $pos = 0;
                $synset_offset = 0;
265             $variantenzaehler = 0;
                $eqsyn_schalter = 0;
                $eqgen_schalter = 0;
                $eqmet_schalter = 0;
                $lauf = $1;
270         }
            $zahl = $2;                                     # OFFSETNUMMER
            $synset_offset = offsetausgabe($zahl);  # ALS 8DIGIT
        }
    ##################################EBENE 1: PART OF SPEECH
275     elsif ($_ =~ /\s*?(1) PART_OF_SPEECH \"(\w)\"/)
        {
            $z++;
            if($1 >= $lauf)
            {
280             $variants_schalter = 0;
                $internallinks_schalter = 0;
                $eqlinks_schalter = 0;
                $lauf = $1;
            }
285         else
            {
                $lauf = $1;
            }
            $pos = $2;
290         $alles[$zahl]{'POS'} = "$pos";
        }
    ##################################EBENE 1: ELEMENTE DES SYNSETS
        elsif ($_ =~/\s*?(1) VARIANTS/)
        {
295         $z++;
            if($1 >= $lauf)
            {
                $internallinks_schalter = 0;
                $eqlinks_schalter = 0;
300             $lauf = $1;
            }
            else
            {
                @lemmata = ();
```

```
305             $lauf = $1;
            }
            $variants_schalter = 1;
        }
    ################################EBENE 1: INTERNAL_LINKS
310     elsif ($_ =~/\s*?(1) INTERNAL_LINKS/)
        {
            $z++;
            if($1 >=$lauf)
            {
315             $eqlinks_schalter = 0;
                $variants_schalter = 0;
                $lauf = $1;
            }
            else
320         {
                %links = ();
                $lauf = $1;
            }
            $internallinks_schalter = 1;
325     }
    ################################EBENE 1: EQ-RELATIONEN
        elsif ($_ =~/\s*?(1) EQ_LINKS/)
        {
            $z++;
330         if($1 >= $lauf)
            {
                $eqlinks_schalter = 0;
                $variants_schalter = 0;
                $internallinks_schalter = 0;
335             $lauf = $1;
            }
            else
            {
                $eqsyn_schalter = 0;
340             $eqgen_schalter = 0;
                $eqmet_schalter = 0;
                %eqlinks = ();
                $eqsynonym_offset = 0;
                $lauf = $1;
345         }
            $eqlinks_schalter = 1;
        }
    ################################EBENE 2: LITERAL
        elsif ($_ =~/\s*?(2) LITERAL \"(.*?)\"/
```

```
350             && $variants_schalter == 1)
    {
        $z++;
        if($1 >= $lauf)
        {
355             $lauf = $1;
        }
        else
        {
            $lem = 0;
            %lemma = ();
360         $lauf = $1;
        }
        $lem = $2;                                  # LITERAL IS $lem
        $alles[$zahl]{'LEMMATA'}[$variantenzaehler]{"LEM"}
365                                                         = $lem;
        $literal_schalter = 1;
    }
################################EBENE 2: RELATION
    elsif ($_ =~/\s*?(2) RELATION \"(.*?)\"/
370         && $internallinks_schalter == 1)
    {
        $z++;
        if($1 >=$lauf)
        {
375         $lauf = $1;
        }
        else
        {
            $lauf = $1;
380     }
        $symbolliteral = $2;
        # SYMBOL ZUWEISEN
        $symbolzeichen = $ptr_hash{"$symbolliteral"};
        $symbol_schalter = 1;
385     if ($symbolliteral =~ /antonym/)            # LEX. LINK
        {
            $antonym_schalter = 1;
        }
        elsif ($symbolliteral =~ /has_meronym/      ||
390            $symbolliteral =~ /has_mero_member/  ||
               $symbolliteral =~ /has_holo_madeof/  ||
               $symbolliteral =~ /has_mero_part/    ||
               $symbolliteral =~ /has_holo_part/     )
        {
```

```
395             $reversed_schalter = 1;
            }
            else
            {
            }
400         if ($symbolliteral =~ /has_meronym/ ||
                $symbolliteral =~ /has_holonym/)
            {
                print (LOG "Fehler: es wurde in der Datenbank "
                       . $symbolliteral." gefunden.");
405             print (LOG " Dafuer existiert keine
                            Entsprechung in WN.\n");
            }
            else
            {
            }
410     }
    ################################EBENE 2: EQ-SYNONYM
        elsif ($_ =~/\s*?(2) EQ_RELATION "eq_synonym"/
               && $eqlinks_schalter == 1)
415     {
            $z++;
            if($1 >= $lauf)
            {
                $lauf = $1;
420         }
            else
            {
                $targetili_schalter = 0;
                $lauf = $1;
425         }
            $eqsyn_schalter = 1;
            $eqlinks_schalter = 0;
        }
    ################################EBENE 2: EQ_GENERALIZATION
430     elsif ($_ =~/\s*?(2) EQ_RELATION "eq_generalization"/ )
        {
            $z++;
            if($1 >= $lauf)
            {
435             $lauf = $1;
            }
            else
            {
                $targetili_schalter = 0;
```

```
440             $lauf = $1;
            }
            $eqgen_schalter = 1;
            $eqmet_schalter = 0;
        }
445 ##################################EBENE 2: EQ_METONYM
        elsif ($_ =~/\s*?(2) EQ_RELATION "eq_metonym"/ )
        {
            $z++;
            if($1 >= $lauf)
450         {
                $lauf = $1;
            }
            else
            {
455             $lauf = $1;
            }
            $eqmet_schalter = 1;
            $eqgen_schalter = 0;
        }
460 ##################################EBENE 3: SENSE
    elsif ($_ =~/\s*?(3) SENSE (\d+)/
            && $literal_schalter == 1)
    {
        $z++;
465     if($1 >= $lauf)
        {
            $lauf = $1;
        }
        else
470     {
            $sen = 0;
            $lauf = $1;
        }
        $sen = twodigit($2);                    # BEDEUTUNGSNUMMER
475     $alles[$zahl]{'LEMMATA'}[$variantenzaehler]{"SEN"}
                                                    = $sen;
        $variantenzaehler++;
    }
  ######################################EBENE 3: EXTERNAL_INFO
480 elsif ($_ =~/\s*?(3) EXTERNAL_INFO/
            && $literal_schalter == 1)
    {
        $z++;
        if($1 >=$lauf)
```

```
485         {
                $lauf = $1;
            }
            else
            {
490             $lauf = $1;
            }
            $externalinfo_schalter = 1;
        }
    ########################################EBENE 3: FEATURES
495     elsif ($_ =~/\s*?(3) FEATURES/)
        {
            $z++;
            if($1 >=$lauf)
            {
500             $lauf = $1;
            }
            else
            {
                $lauf = $1;
505         }
            $features_schalter = 1;
        }
    ###################################EBENE 3: TARGET_CONCEPT
        elsif ($_ =~/\s*?(3) TARGET_CONCEPT/
510             && $symbol_schalter == 1)
        {
            $z++;
            if($1 >=$lauf)
            {
515             $lauf = $1;
            }
            else
            {
                $lauf = $1;
520         }
            $symbol_schalter = 0;
            $targetili_schalter = 0;
            $targetconcept_schalter = 1;
        }
525 ###################################EBENE 3: TARGET_ILI
        elsif ($_ =~/\s*?(3) TARGET_ILI/)
        {
            $z++;
            if($1 >= $lauf)
```

```
530         {
                $lauf = $1;
            }
            else
            {
535             $lauf = $1;
            }
            $targetili_schalter = 0;
            $targetili_schalter = 1;
        }
540 ##################################EBENE 4: SOURCE_ID
        elsif ($_ =~/\s*?(4) SOURCE_ID 1/
                && $externalinfo_schalter == 1)
        {
            $z++;
545         if($1 >=$lauf)
            {
                $lauf = $1;
            }
            else
550         {
                $lauf = $1;
            }
            $externalinfo_schalter = 0;
            $sourceid_schalter = 1;
555     }
    ##################################EBENE 4: PART_OF_SPEECH
        elsif ($_ =~/\s*?(4) PART_OF_SPEECH \"n\"/
                && $targetconcept_schalter == 1)
        {
560         $z++;
            if($1 >=$lauf)
            {
                $lauf = $1;
            }
565         else
            {
                $lauf = $1;
            }
        }
570 ##################################EBENE 4: LITERAL
        elsif ($_ =~/\s*?(4) LITERAL \"(.*?)\"/
                && $targetconcept_schalter == 1)
        {
            $z++;
```

```
575         if($1 >=$lauf)
            {
                $lauf = $1;
            }
            else
580         {
                $ptr_lemma = 0;
                $ptr_lemma_schalter = O;
                $lauf = $1;
            }
585         $ptr_lemma = $2;                                    # LEMMA DES LINKS
            $ptr_lemma_schalter = 1;
        }
    ##################################### EBENE 4: PART_OF_SPEECH
        elsif ($_ =~/\s*?(4) PART_OF_SPEECH \"n\"/
590             && $targetili_schalter == 1)
        {
            $z++;
            if($1 >=$lauf)
            {
595             $lauf = $1;
            }
            else
            {
                $lauf = $1;
600         }
        }
    ##################################### EBENE 4: WORDNET_OFFSET
        elsif ($_ =~/\s*?(4) WORDNET_OFFSET (\d+)/
            && $targetili_schalter == 1
605         && $eqsyn_schalter == 1)
        {
            $z++;
            if($1 >= $lauf)
            {
610             $lauf = $1;
            }
            else
            {
                $lauf = $1;
615             $eqsynonym_offset = 0;
            }
            # WN_OFFSET ALS EQ-LINK
            $alles[$zahl]{'EQ_LINKS'}{'SYNONYM'} = $2;
            $eqsynonym_offset = $2;
```

```
620         $eqsyn_schalter = 0;
    }
    ##################################EBENE 4: ADD_ON
    elsif ($_ =~/\s*?(4) ADD_ON_ID (\d+)/
            && $targetili_schalter == 1 )
625 {
        $z++;
        if($1 >= $lauf)
        {
            $lauf = $1;
630         }
        else
        {
            $lauf = $1;
        }
635
        if ($eqgen_schalter == 1)
        {
            $alles[$zahl]->{'EQ_LINKS'}->{'GENERALIZATION'}
            = $eqsynonym_offset+$2;          # WN_OFFSET ALS EQ-LINK
640         $eqgen_schalter = 0;
        }
        elsif ($eqmet_schalter == 1)
        {
            $alles[$zahl]->{'EQ_LINKS'}->{'METONYM'}
645         = $eqsynonym_offset+$2;          # WN_OFFSET ALS EQ-LINK
            $eqmet_schalter = 0;
        }
    }
    ##################################EBENE 4: VARIANT_TO_VARIANT
650 elsif ($_ =~/\s*?(4) VARIANT_TO_VARIANT/
            && $antonym_schalter == 1
            && $features_schalter == 1)
    {
        $z++;
655     if($1 >=$lauf)
        {
            $lauf = $1;
        }
        else
660     {
            $variant_schalter = 0;
            $lauf = $1;
        }
        $antonym_schalter = 0;
```

```
665         $features_schalter = 0;
            $variant_schalter = 1;
        }
        ##################################EBENE 4: REVERSED
        elsif ($_ =~/\s*?(4) REVERSED/ && $reversed_schalter == 1
670             && $features_schalter == 1)
        {
            $z++;
            if($1 >=$lauf)
            {
675             $lauf = $1;
            }
            else
            {
                $lauf = $1;
680         }
            $alles[$zahl]{'LINKS'}{"$linkkey"} .= "\*\*\*REVERSED";
        }
        ##################################EBENE 5: TEXT_KEY
        elsif ($_ =~/\s*?(5) TEXT_KEY \"(.*?)\"/
685             && $sourceid_schalter == 1)
        {
            $z++;
            if($1 >=$lauf)
            {
690             $lauf = $1;
            }
            else
            {
                $textkey = 0;
695             $lauf = $1;
            }
            $textkey = $2;
            $alles[$zahl]{'TEXT_KEY'} = $textkey;
            $sourceid_schalter = 0;
700         $literal_schalter = 0;
        }
        ##################################EBENE 5: SENSE
        elsif ($_ =~/\s*?(5) SENSE (\d+)/
            && $ptr_lemma_schalter == 1)
705     {
            $z++;
            if($1 >=$lauf)
            {
                $lauf = $1;
```

```
        }
        else
        {
            $lauf = $1;
        }
        $linksense = twodigit($2);                  # BEDEUTUNGSNUMMER
        $linkkey = $ptr_lemma."\*".$linksense;      # LEMMA#SENSE
        # IN %LINKS
        $alles[$zahl]{'LINKS'}{"$linkkey"} = $symbolzeichen;
        $ptr_lemma_schalter = 0;
        $symbolliteral = 0;
        $symbolzeichen = 0;
        $symbol_schalter = 0;
    }
    ##################################EBENE 5: SOURCE_VARIANT
    elsif ($_ =~/\s*?(5) SOURCE_VARIANT \"(.*?)\"/
            && $variant_schalter == 1)
    {
        $z++;
        if($1 >=$lauf)
        {
            $lauf = $1;
        }
        else
        {
            $sourcevariant = 0;
            $lauf = $1;
        }
        $sourcevariant = $2;
    }
    ##################################EBENE 5: TARGET_VARIANT
    elsif ($_ =~/\s*?(5) TARGET_VARIANT \"(.*?)\"/
            && $variant_schalter == 1)
    {
        $z++;
        if($1 >=$lauf)
        {
            $lauf = $1;
        }
        else
        {
            $targetvariant = 0;
            $lauf = $1;
        }
        $targetvariant = $2;
```

```
        $source_offset = offsetausgabe($zahl);# OFFSETS SPEICHERN
        $target_offset = $index{$linkkey};
        $alles[$zahl]{'LINKS'}{$linkkey} .=
                              "\*\*\*$sourcevariant\*$source_offset
                              \*\*$targetvariant\*$target_offset";
        $linkkey = 0;
        $linksense = 0;
        $variant_schalter = 0;
    }
      elsif ($_ =~ /^\s$/)
      {
        $z++;
        print LOG "Zeilenvorschub in Zeile ". $z. "!\n";
    }
      else
      { $z++;
        print LOG "Der Eintrag in Zeile " . $z." passt nicht!\n";
        print LOG "Die Datenbank wurde nicht
                   erfolgreich eingelesen!\n";
        exit;
    }
  }
  print STDOUT "Ende der Auswertung.\n";
  print LOG "Datenbank erfolgreich zur Auswertung
             in Speicher eingelesen:\n";
  print LOG "Es wurden " .$z." Zeilen eingelesen.\n";
  ######################
  #ENDE DES EINLESENS#
  ######################

  ########################################
  #BEGINN: AUSGABE IN INDEX UND DATA#
  ########################################
  my $dummy_wnoffset = 0;
  my $dummy_tagsensecnt =  0;
  my $dummy_lexid = "0";
  my $dummy_twodigits = "00";
  my $dummy_fourdigits = "0000";
  my $dummy_pcnt = 0;
  my $dummy_ptrsymbol = 0;
  my $dummie;
  my @symbolarray = ("0","0","0","0","0","0","0","0","0");
  my @allesymbole = ("0","0","0","0","0","0","0","0","0");
  my $t = 0;                    # ZAEHLVARIABLE
  my %indexalphabetisch;
```

```
my $lex = 0;                    # EINTRAG IM SYNSET
my $bed = 0;                    # BEDEUTUNGSNUMMER
my $syn_offset = 0;             # OFFSETNUMMER
my $l = 0;                      # LAUFVARIABLE
my $j;
my $cle = 0;
my $r = 0;
my @symbolliste = ("","","","","","","","","","");
my $sicher;
my $sicher2;
my $key = 0;
my $y = 0;

#######################################
#BEGINN: VORBEREITUNG FUER INDEX#
#######################################
open (INDEX_ALPHABETISCH, "<Index_alphabetisch.ewn")
    or die("Kann Index_alphabetisch.ewn nicht oeffnen.");
while (<INDEX_ALPHABETISCH>)
{
    $l++;
    if ($_ =~ /^(.*?)\*(\w+?): (\w*?)$/)
    {
        $lex = $1;                      # ALS KEY
        $bed = $2;                      # POSITION IM ARRAY + 1
        $syn_offset = $3;               # WIRD AUF ARRAY GEPUSHT
        push(@{$indexalphabetisch{"$lex"}},$3);
    }
    else
    {
        print STDOUT "Fehler beim Lesen von Index_alphabetisch
                      in Zeile: ". $l."\n";
    }
}
print LOG "Es wurden aus Index_alphabetisch.ewn "
                        . $l." Zeilen gelesen.\n";
print STDOUT "Einlesen in hash fertig\n";

######ERSTELLUNG VON NOUN.EXC
print STDOUT "Erstellung von noun.exc.\n";
open (INDEX_ALPHABETISCH, "<Index_alphabetisch.ewn")
    or die("Kann Index_alphabetisch.ewn nicht oeffnen.");
open (OUT, ">noun.exc")
    or die("Kann noun.exc nicht zum Schreiben oeffnen.");
while (<INDEX_ALPHABETISCH>)
```

```
{
    $l++;
    if ($_ =~ /^(.*?)\*(\w+?): (\w*?)$/)
    {
        $lex = $1;
        print OUT $lex. " ". $lex. "\n";
    }
    else
    {
        print LOG "Fehler beim Lesen von Index_alphabetisch
                   in Zeile: ". $l."\n";
    }
}
close OUT;
print STDOUT "Erledigt.\n";

##############################################
# BEGINN DER TESTAUSGABE VON DATA.NOUN#
##############################################
open(DUMMY, "> data.dummy")
    or die("Kann Ausgabe data.dummy nicht oeffnen.");
print STDOUT "Testausgabe von data.dummy.\n";
$t = 0;
for ($j=1; $j<=$#alles; $j++)
{
    if (defined($alles[$j]))
    {
        $t++;
        print DUMMY offsetausgabe($j)." "; # 8 DIGIT
        print DUMMY $dummy_twodigits. " "; # DUMMY (2DIGIT INT)
        print DUMMY $alles[$j]{'POS'}. " ";# $POS (IMMER N)
        $lemmata_laenge = $#{$alles[$j]{'LEMMATA'}}+1;
        print DUMMY twodigit($lemmata_laenge)." ";

        for ($a = 0; $a <=$#{$alles[$j]{'LEMMATA'}}; $a++)
        {
            if (defined $alles[$j]{'LEMMATA'}[$a]{'LEM'})
            {
                print DUMMY leerzeichen($alles[$j]
                                {'LEMMATA'}[$a]{'LEM'})." ";
                print DUMMY $dummy_lexid. " ";
            }
            else
            {
            }
```

```
890         }
            $cle = 0;
            $r = 0;
            foreach $cle (keys (%{$alles[$j]{'LINKS'}}))
            {
895             $r++;
            }
            print DUMMY threedigit($r). " ";
            if (defined %{$alles[$j]{'LINKS'}})
            {
900             # LINKLISTE
                foreach $key (keys (%{$alles[$j]{'LINKS'}}))
                {
                    if ($alles[$j]{'LINKS'}{$key} eq "@")
                    {
905                     $symbolliste[0]
                            .= "@ 00000000 n $dummy_fourdigits ";
                    }
                    if ($alles[$j]{'LINKS'}{$key} =~
                       /^(\!)\*\*\*(.*?)\*(\d*?)\*\*(.*?)\*(.*?)$/)
910                 {
                      $sicher = $1;
                      $alles[$j]{'LINKS'}{$key}{'SOURCELEX'} = $2;
                      $alles[$j]{'LINKS'}{$key}{'SOURCEOFFSET'}=$3;
                      $alles[$j]{'LINKS'}{$key}{'TARGETLEX'}=$4;
915                   $alles[$j]{'LINKS'}{$key}{'TARGETOFFSET'}=$5;
                      $sicher2 = $sicher." 00000000 n ";
                      for($a=0;$a<=$#{$alles[$j]{'LEMMATA'}};$a++)
                        {
                           if ($alles[$j]{'LEMMATA'}[$a]{'LEM'} =~
920                       /$alles[$j]{'LINKS'}{$key}{'SOURCELEX'}/)
                            {
                                my $b = 0;
                                $b = $a+1;        # BEDEUTUNGSNUMMER
                                my $c = twodigit($b);
925                             $sicher2 .= "$c";
                            }
                            else
                            {
                            }
930                     }
                      my $schluessel = 0;
                      foreach $schluessel (keys(%index))
                      {
                            if ($index{$schluessel} =~
```

```
935                 $alles[$j]{'LINKS'}{$key}{'TARGETOFFSET'})
                        {
                            if  ($schluessel =~
                                 /^$alles[$j]{'LINKS'}{$key}
                                 {'TARGETLEX'}\*(\d*?)$/)
940                         {
                                my $d = scalar($1);
                                $sicher2 .= "$d ";
                            }
                            else
945                         {
                            }
                        }
                    }
                    $symbolliste[1] .= "$sicher2";
950                 $sicher2 = 0;
                }
                if ($alles[$j]{'LINKS'}{$key} eq "~")
                {
                    $symbolliste[2] .= "~ 00000000 n "
955                     .$dummy_fourdigits." ";
                }
                if ($alles[$j]{'LINKS'}{$key} eq "#m")
                {
                    $symbolliste[3] .= "#m 00000000 n "
960                     .$dummy_fourdigits." ";
                }
                if ($alles[$j]{'LINKS'}{$key} eq "#s")
                {
                    $symbolliste[4] .= "#s 00000000 n "
965                     .$dummy_fourdigits." ";
                }
                if ($alles[$j]{'LINKS'}{$key} eq "#p")
                {
                    $symbolliste[5] .= "#p 00000000 n "
970                     .$dummy_fourdigits." ";
                }
                if ($alles[$j]{'LINKS'}{$key} eq "%m")
                {
                    $symbolliste[6] .= "%m 00000000 n "
975                     .$dummy_fourdigits." ";
                }
                if ($alles[$j]{'LINKS'}{$key} eq "%s")
                {
                    $symbolliste[7] .= "%s 00000000 n "
```

```
                        .$dummy_fourdigits." ";
                }
                if ($alles[$j]{'LINKS'}{$key} eq "%p")
                {
                    $symbolliste[8] .= "%p 00000000 n "
                        .$dummy_fourdigits." ";
                }
            }
            $y = 0;
            for ($y=0; $y<=$#symbolliste; $y++)
            {
                print DUMMY $symbolliste[$y];
            }
            @symbolliste = ("","","","","","","","","");
        }
        else
        {
        }
        print DUMMY "\| GLOSSE";
        print DUMMY "\n";
    }
}
close DUMMY;
print STDOUT "Data.dummy wurde erstellt.\n";
##############################
#ENDE: TESTAUSGABE IN DATA#
##############################
#AUSLESEN VON DATA.DUMMY, UM NEUE OFFSETS ZU BESTIMMEN:
open(IN, "<data.dummy")
    or die ("Kann data.dummy nicht wieder oeffnen.");
open(OUT, ">alt.neu")
    or die ("Kann alt.neu nicht oeffnen");
my $sicher = 1;
my $position;
my %offset;                         # Form: (alt => neu)
my $nummer;
while (<IN>)
{
    if ($_ =~ /^(\d\d\d\d\d\d\d)\s*.?| GLOSSE\n$/)
    {
        $nummer = $1;
        $offset{$nummer} = $sicher;
        $sicher = tell(IN);
    }
    else
```

```
1025    {
            print "else";
        }
    }
    foreach $key (keys %offset)
1030 {
        print OUT $key. ": " . $offset{$key}."\n";
    }
    close IN;
    close OUT;
1035 #GRUNDLAGE FUER STATISTISCHE AUSWERTUNG
    open(KREUZ, ">IndexKreuz.ewn")
        or die("Kann Datei IndexKreuz.ewn nicht oeffnen.");
    foreach (sort {$a cmp $b} (keys(%index)))
    { if (defined keys(%index))
1040   {
           my $num = $index{$_};
           s/^(.*?)\*(\d*?)$/$1\#n\#$2/;
           my $ind = leerzeichen($_);
           print KREUZ lowercase($ind). ": ". $offset{$num}. "\n";
1045   }
       else
       {
       }
    }
1050 print STDOUT "IndexKreuz.ewn erfolgreich erstellt.\n";
    print LOG "IndexKreuz.ewn erfolgreich erstellt.\n";
    close KREUZ;

    ############################
1055 #BEGINN: AUSGABE IN INDEX#
    ############################
    open(INDEX, ">index.noun")
        or die("Kann Ausgabe index.noun nicht oeffnen.");
    print STDOUT "Beginn der Erstellung von index.noun.\n";
1060 $a = 0;
    $b = 0;
    my $s = 0;              # POSITION IM ARRAY
    my $v = 0;              # ZAEHLVARIABLE
    my $u = 0;
1065 my $w = 0;              # ARRAYLAENGE
    my $x = 0;              # SUMME ALLER $w
    my $wechsel = 0;
    $cle = 0;           # ZAEHLVARIABLE
    $key = 0;
```

```
foreach $key (sort{$indexalphabetisch{$a}<=>
                  $indexalphabetisch{$b}}(keys %indexalphabetisch))
{
    $t++;
    $wechsel = leerzeichen($key);
    print INDEX lowercase($wechsel)." ";    # AUSGABE DES LEMMA
    print INDEX "n ";                       # POS
    # ANZAHL DER LEMMATA
    print INDEX @{$indexalphabetisch{$key}}. " ";
    for ($j=0; $j<=$#{$indexalphabetisch{$key}}; $j++)
    {
        $s = offseteingabe($indexalphabetisch{$key}[$j]);
        # LINKS ERMITTLEN UND AUSGEBEN
        if (defined $alles[$s]{'LINKS'})
        {
            foreach $cle (keys (%{$alles[$s]{'LINKS'}}))
            {
                if ($alles[$s]{'LINKS'}{$cle} =~ /^([^\*]*?)$/)
                {
                    $zeichen = $1;
                }
                elsif ($alles[$s]{'LINKS'}{$cle} =~
                        /([^\*]*?)[\*]*?REVERSED$/)
                {
                    $zeichen = $1;
                }
                elsif ($alles[$s]{'LINKS'}{$cle} =~
                    /^(!)\*\*\*(.*?)\*(\d*?)\*\*(.*?)\*(.*?)$/)
                {
                    $zeichen = $1;
                }
                else
                {
                }
                if ($zeichen eq "@")
                {
                    $symbolarray[0] = "@";
                }
                if ($zeichen eq "!")
                {
                    $symbolarray[1] = "!";
                }
                if ($zeichen eq "~")
                {
                    $symbolarray[2] = "~";
```

```
1115                    }
                        if ($zeichen eq "#m")
                        {
                            $symbolarray[3] = "#m";
                        }
1120                    if ($zeichen eq "#s")
                        {
                            $symbolarray[4] = "#s";
                        }
                        if ($zeichen eq "#p")
1125                    {
                            $symbolarray[5] = "#p";
                        }
                        if ($zeichen eq  "%m")
                        {
1130                        $symbolarray[6] = "%m";
                        }
                        if ($zeichen eq "%s")
                        {
                            $symbolarray[7] = "%s";
1135                    }
                        if ($zeichen eq "%p")
                        {
                            $symbolarray[8] = "%p";
                        }
1140                    else
                        {
                        }
                }
            }
1145        for ($u=0; $u<=$#symbolarray; $u++)
            {
                if ($symbolarray[$u] ne "0")
                {
                    $allesymbole[$u] = $symbolarray[$u];
1150            }
                else
                {
                }
            }
1155        @symbolarray = ("0","0","0","0","0","0","0","0","0");
        }
        for ($v=0; $v<=$#allesymbole; $v++)
        {
            if ($allesymbole[$v] ne "0")
```

```
1160            {
                    $w++;
                }
                else
                {
1165            }
        }
        print INDEX $w. " ";                    # SUMME DER SYMBOLE
        $w = 0;
        for ($u=0; $u<=$#allesymbole; $u++) # AUSGABE SYMBOLLINKS
1170    {
            if ($allesymbole[$u] ne "0")
            {
                print INDEX $allesymbole[$u]. " ";
            }
1175        else
            {
            }
        }
        @allesymbole = ("0","0","0","0","0","0","0","0","0");
1180    # ANZAHL DER LEMMATA
        print INDEX @{$indexalphabetisch{$key}}. " ";
        print INDEX $dummy_tagsensecnt. " ";    # TAGSENSE (DUMMY)
        my $q = @{$indexalphabetisch{$key}};
        for ($i = 0; $i<$q; $i++)
1185    {
     print INDEX
     offsetausgabe($offset{"${$indexalphabetisch{$key}}[$i]"})." ";
        }
        print INDEX "\n";
1190 }
    close INDEX;
    print STDOUT "Index.noun wurde erfolgreich erstellt.\n";
    print LOG "In Index.noun befinden sich ". $t. " Eintraege.\n";
    ############################
1195 ##ENDE: AUSGABE IN INDEX#
    ############################

    ##########################################
    ####BEGINN DER ERSTELLUNG VON DATA.NOUN#
1200 ##########################################
    open(DATA, "> data.noun") or
        die("Kann Ausgabe data.noun nicht oeffnen.");
    print STDOUT "Beginn der Erstellung von data.noun.\n";
    $t = 0;
```

```
$dummie = 0;
$j = 0;
my $zwischen = 0;
for ($j=1; $j<=$#alles; $j++)
{
    if (defined($alles[$j]))
    {
        $t++;
        $dummie = offsetausgabe($j);            # 8DIGIT INTEGER
        print DATA offsetausgabe($offset{"$dummie"}). " ";
        print DATA $dummy_twodigits. " ";       # DUMMY
        print DATA $alles[$j]{'POS'}. " ";      # $POS
        # LEMMATA-ANZAHL
        $lemmata_laenge = $#{$alles[$j]{'LEMMATA'}}+1;
        print DATA twodigit($lemmata_laenge)." ";
        for ($a = 0; $a <=$#{$alles[$j]{'LEMMATA'}}; $a++)
        {
            if (defined $alles[$j]{'LEMMATA'}[$a]{'LEM'})
            # LEMMA ANGEBEN
            {
                $zwischen =
                 leerzeichen($alles[$j]{'LEMMATA'}[$a]{'LEM'});
                print DATA lowercase($zwischen)." ";
                print DATA $dummy_lexid. " "; # DUMMY LEXID
            }
            else
            {
            }
        }
        foreach $cle (keys (%{$alles[$j]{'LINKS'}}))
        {
            $r++;
        }
        print DATA threedigit($r). " ";         # ZAHL DER LINKS
        if (defined %{$alles[$j]{'LINKS'}})
        {
            @symbolliste = ("","","","","","","","","");
            $sicher = 0;
            $sicher2 = 0;
            $key = 0;
            # LINKLISTE
            foreach $key (keys (%{$alles[$j]{'LINKS'}}))
            {
                if ($alles[$j]{'LINKS'}{$key} eq "@")
                {
```

```
1250                $symbolliste[0] .= "@ "
                        .offsetausgabe($offset{"$index{$key}"})
                            ." n $dummy_fourdigits ";
                }
                if ($alles[$j]{'LINKS'}{$key} =~ /
1255                ^(\!)\*\*\*(.*?)\*(\d*?)\*\*(.*?)\*(.*?)$/)
                {
                  $sicher = $1;
                  $alles[$j]{'LINKS'}{$key}{'SOURCELEX'}= $2;
                  $alles[$j]{'LINKS'}{$key}{'SOURCEOFFSET'}=$3;
1260              $alles[$j]{'LINKS'}{$key}{'TARGETLEX'} = $4;
                  $alles[$j]{'LINKS'}{$key}{'TARGETOFFSET'}=$5;
                  $sicher2 = $sicher." "
                      .offsetausgabe($offset{"$index{$key}"})
                          ." n ";
1265              for($a=0;$a<=$#{$alles[$j]{'LEMMATA'}}; $a++)
                    {
                      if ($alles[$j]{'LEMMATA'}[$a]{'LEM'} =~
                     /$alles[$j]{'LINKS'}{$key}{'SOURCELEX'}/)
                        {
1270                        my $b = 0;
                            $b = $a+1;
                            my $c = twodigit($b);
                            # BEDSNUMMER DES LINKS
                            $sicher2 .= "$c";
1275                    }
                        else
                        {
                        }
                    }
1280                # BEDNUMMER VON TARGETLEX IN %INDEX
                    $schluessel = 0;
                    foreach $schluessel (keys(%index))
                    {
                        if ($index{$schluessel} =~
1285                  $alles[$j]{'LINKS'}{$key}{'TARGETOFFSET'})
                            {
                                if  ($schluessel =~
/^$alles[$j]{'LINKS'}{$key}{'TARGETLEX'}\*(\d*?)$/)
                                {
1290                                my $d = scalar($1);
                                    #BEDNUMMER
                                    $sicher2 .= "$d ";
                                }
                                else
```

```
                                {
                                }
                            }
                        }
                    $symbolliste[1] .= "$sicher2";
                    $sicher2 = 0;
                }
                if ($alles[$j]{'LINKS'}{$key} eq "~")
                {
                    $symbolliste[2] .= "~ "
                        .offsetausgabe($offset{"$index{$key}"})
                            ." n ".$dummy_fourdigits." ";
                }
                if ($alles[$j]{'LINKS'}{$key} eq "#m")
                {
                    $symbolliste[3] .= "#m "
                        .offsetausgabe($offset{"$index{$key}"})
                            ." n ".$dummy_fourdigits." ";
                }
                if ($alles[$j]{'LINKS'}{$key} eq "#s")
                {
                    $symbolliste[4] .= "#s "
                        .offsetausgabe($offset{"$index{$key}"})
                            ." n ".$dummy_fourdigits." ";
                }
                if ($alles[$j]{'LINKS'}{$key} eq "#p")
                {
                    $symbolliste[5] .= "#p "
                        .offsetausgabe($offset{"$index{$key}"})
                            ." n ".$dummy_fourdigits." ";
                }
                if ($alles[$j]{'LINKS'}{$key} eq "%m")
                {
                    $symbolliste[6] .= "%m "
                        .offsetausgabe($offset{"$index{$key}"})
                            ." n ".$dummy_fourdigits." ";
                }
                if ($alles[$j]{'LINKS'}{$key} eq "%s")
                {
                    $symbolliste[7] .= "%s "
                        .offsetausgabe($offset{"$index{$key}"})
                            ." n ".$dummy_fourdigits." ";
                }
                if ($alles[$j]{'LINKS'}{$key} eq "%p")
                {
```

```
1340                        $symbolliste[8] .= "%p "
                                .offsetausgabe($offset{"$index{$key}"})
                                    ." n ".$dummy_fourdigits." ";
                        }
                    }
                    for ($y=0; $y<=$#symbolliste; $y++)
                    {
                        print DATA $symbolliste[$y]; # AUSGABE SYMBOLE
                    }
                    @symbolliste = ("","","","","","","","","");
1350            }
                else
                {
                }
                print DATA "\| glosse";
                print DATA "\n";
        }
    }
    close DATA;
    print STDOUT "Data.noun wurde erfolgreich erstellt.\n";
1360 print LOG "In Data.noun befinden sich ". $t. " Eintraege.\n";
    ###################################
    #ENDE: AUSGABE IN INDEX UND DATA#
    ###################################
    close LOG;
1365 exit;
    ######################
    #ENDE: HAUPTPROGRAMM#
    ######################

1370 ###############
    # SUBROUTINEN#
    ###############
    sub offsetausgabe                          # RUECKGABE VON 8-DIGIT
    {
1375    my($nummer) = @_;
        if ($nummer < 10)
        {
            $nummer = "0000000".$nummer;
        }
1380    elsif ($nummer < 100)
        {
            $nummer = "000000".$nummer;
        }
        elsif ($nummer < 1000)
```

```
1385        {
                $nummer = "00000".$nummer;
            }
            elsif ($nummer < 10000)
            {
1390            $nummer = "0000".$nummer;
            }
            elsif ($nummer < 100000)
            {
                $nummer = "000".$nummer;
1395        }
            elsif ($nummer < 1000000)
            {
                $nummer = "00".$nummer;
            }
1400        elsif ($nummer < 10000000)
            {
                $nummer = "0".$nummer;
            }
            else
1405        {
            }
            return $nummer;
        }

1410    sub offseteingabe       # EINGABE VON 8DIGIT, GIBT ZAHL ZURUECK
        {
            my($nummer) = @_;
            if ($nummer =~ /^0000000(\d)$/)
            {
1415            $nummer = $1;
            }
            elsif ($nummer =~ /^000000(\d\d)$/)
            {
                $nummer = $1;
1420        }
            elsif ($nummer =~ /^00000(\d\d\d)$/)
            {
                $nummer = $1;
            }
1425        elsif ($nummer =~ /^0000(\d\d\d\d)$/)
            {
                $nummer = $1;
            }
            elsif ($nummer =~ /^000(\d\d\d\d\d)$/)
```

```
1430        {
                $nummer = $1;
            }
            elsif ($nummer =~ /^00(\d\d\d\d\d\d)$/)
            {
1435            $nummer = $1;
            }
            elsif ($nummer =~ /^0(\d\d\d\d\d\d)$/)
            {
                $nummer = $1;
1440        }
            else
            {
            }
            return $nummer;
1445    }

        sub twodigit
        {
            my($nummer) = @_;
1450        if ($nummer < 10)
            {
                $nummer = "0".$nummer;
            }
            else
1455        {
            }
            return $nummer;
        }

1460    sub threedigit
        {
            my($nummer) = @_;
            if ($nummer < 10)
            {
1465            $nummer = "00".$nummer;
            }
            elsif ($nummer < 100)
            {
                $nummer = "0".$nummer;
1470        }
            else
            {
            }
            return $nummer;
```

```
}

sub leerzeichen
{
    ($_) = join(" ",@_);
    s/\s/\_/g;
    return $_;
}

sub lowercase
{
    ($_) = join(" ",@_);
    s/([^\W0-9_])/\l$1/g;
    return $_;
}
####################
# ENDE SUBROUTINEN#
####################
```

A.2.2 `Unterknoten.pl`

Um für die in drei Zusammenhangsmaßen benötigte Netzwahrscheinlichkeit zu berechnen wird das Programm `Unterknoten.pl` verwendet:

```
#! /usr/bin/perl
use WordNet::QueryData;
my $wn = WordNet::QueryData->new;
my $dings;
my $z = 0;
open (IN, "<IndexKreuz.ewn")
    or die ("Kann IndexKreuz nicht oeffnen");
open (OUT, ">ic-semcor.dat")
    or die ("Kann ic-semcor.dat nicht zum Schreiben oeffnen");
open (LOG, ">IndexKreuz.log")
    or die ("Kann IndexKreuz.log nicht oeffnen");
print LOG "IndexKreuz geoeffnet.\n";

my $kreuze = join("", <IN>);
print OUT "wnver::2.0\n";
while ($kreuze =~ s/(.*?): (\d*?)\n//)
{
    $z = 0;
```

```
    holetiefe($1);
    print OUT $2 . "n " . $z;
    if ($2 =~ /^1$/
        || $2 =~ /^5971$/
        || $2 =~ /^6076$/
        || $2 =~ /^7446$/
        || $2 =~ /^7650$/
        || $2 =~ /^8212$/
        || $2 =~ /^8915$/
        || $2 =~ /^9011$/
        || $2 =~ /^9309$/
        || $2 =~ /^9577$/
        || $2 =~ /^11021$/)
    {
        print OUT " ROOT\n";
    }
    else
    {
        print OUT "\n";
    }
}
close IN;
print STDOUT "ic-semcor erstellt.\n";
print LOG "ic-semcor erfolgreich erstellt.\n";
close OUT;
close LOG;

sub holetiefe
{
    my ($lemmaeintrag) = @_;
    $z++;
    my @hyponyme = $wn->querySense("$lemmaeintrag", "hypo");
    if ($#hyponyme  == -1)
    {
    }
    else
    {
        for (my $i=0; $i<=$#hyponyme; $i++)
        {
            holetiefe($hyponyme[$i]);
        }
    }
    return;
}
```

```
sub offseteingabe
{
   my($nummer) = @_;
    if ($nummer =~ /^0000000(\d)$/)
     {
       $nummer = $1;
     }
     elsif ($nummer =~ /^000000(\d\d)$/)
     {
       $nummer = $1;
     }
     elsif ($nummer =~ /^00000(\d\d\d)$/)
     {
       $nummer = $1;
     }
     elsif ($nummer =~ /^0000(\d\d\d\d)$/)
     {
       $nummer = $1;
     }
     elsif ($nummer =~ /^000(\d\d\d\d\d)$/)
     {
       $nummer = $1;
     }
     elsif ($nummer =~ /^00(\d\d\d\d\d\d)$/)
     {
       $nummer = $1;
     }
     elsif ($nummer =~ /^0(\d\d\d\d\d\d\d)$/)
     {
       $nummer = $1;
     }
     else
     {
     }
     return $nummer;
  }
```

Dieses Programm erstellt die Datei `ic-semcor.dat`, die wegen des Aufrufs durch das Perlpaket `ICFinder.pm` den Namen der Datei erhält, die auf der Basis des *SemCor*-Projektes entstand (vgl. Abschnitt 8.8).

A.2.3 Verwendete Dateien

Dieses Programm greift auf die Dateien `synsetdepth-2.0.dat` und `treedepth-2.0.dat` zu, die von dem durch die Datenbank *WordNet* bereitgestellten Programm `wnDepth.pl` auf der Grundlage der neuen Datenbank berechnet wurden. Es folgt ein Ausschnitt aus der für die verbesserte Datenbank erstellte Datei `synsetdepth-2.0.dat`:

```
wnver::2.0
n 00000001 1:00000001
n 00000219 2:00000001
n 00000544 3:00000001
⋮
n 00003553 4:00007581
n 00003688 10:00000001
n 00003744 11:00006211 12:00000001
n 00003806 3:00000001
```

Damit ist z.B. für das *Synset* mit dem Offset 00003553 der Kopfknoten das Synset mit dem Offset 00000001. Der Wert für die Tiefe dieses Knotens ist 3. Bei mehreren möglichen Kopfknoten werden alle angegeben (geordnet nach der Tiefe): mit dem Tiefenwert 11 ist 00003744 unter 00006211 eingehängt, mit dem Tiefenwert 12 unter 00000001.

Als nächstes folgt ein Ausschnitt aus der Datei `treedepth-2.0.dat`. Von `wnDepth.pl` werden ebenfalls die Hierarchietiefen der Verbtaxonomie berechnet, die aus Kompabilitätsgründen die englischen Daten enthält. Daher werden diese Angaben hier unterschlagen.

```
wnver::2.0
n 00009444 11
n 00007581 9
n 00009050 7
n 00009712 12
n 00011156 10
n 00008347 10
n 00006106 12
n 00006211 12
n 00000001 15
n 00007785 8
n 00009146 11
n 00000000 16
```

Unter einem (imaginären) Kopfknoten wird die Hierarchie zusammengefaßt (dieser bekommt das Offset 00000000 und steht an der Spitze einer Hierarchie mit 16 Ebenen. Die anderen Teilhierarchien sind unterschiedlich tief: das Offset 00009050 steht als oberster Knoten über 7 Ebenen, das Offset 00000001 über einer Hierarchie mit 15 Ebenen (vgl. dazu auch die Veränderungen, die an der Datenbank vorgenommen wurden, S. 105ff.).

Es folgt ein Ausschnitt aus der mit dem Programm `Unterknoten.pl` erstellten Datei `ic-semcor.dat`:

```
wnver::2.0
⋮
147653n 1
6211n 3198 ROOT
723279n 18
⋮
935554n 296
936668n 391
5131n 789
247202n 1
935069n 5
```

Das *Synset* mit der Offsetnummer 147653 (als Substantiv gekennzeichnet durch „n“) besitzt den Wert 1 (da es keinen Unterknoten hat und der Knoten selbst mitgezählt wird). Der Knoten mit der Offsetnummer 6211 ist ein Kopfknoten und besitzt (sich selbst mitgezählt) 3.198 Unterknoten.

A.3 Zusammenhangsmaße

A.3.1 Das kantenzählende Perlmodul

Als erstes kleineres Modulpaket für das Perlprogramm `similarity.pl` wurde ein vorhandenes Programmpaket leicht abgeändert (vgl. Abschnitt 9.2). Aus dem Paket `edge.pm` wird aus dem knotenzählenden Maß ein kantenzählendes Maß `edge2.pm`, wobei die maximale Tiefe der Hierarchie berücksichtigt wird. Aus Platzgründen wurde aus dem Quellcode neben den Kommentaren auch der einleitende Hinweis auf die *GNU General Public License* herausgenommen. Die eigenen Änderungen am Quellcode werden gesondert markiert.

```
package WordNet::Similarity::edge2;
use strict;
use Exporter;
use WordNet::Similarity::edge;
use vars qw($VERSION @ISA @EXPORT @EXPORT_OK %EXPORT_TAGS);
@ISA = qw(Exporter);
%EXPORT_TAGS = ();
@EXPORT_OK = ();
@EXPORT = ();
$VERSION = '0.06';
my $maxdist = 16; # AENDERUNG: MAX. TIEFE IN EWN
sub new
{
    my $className;
    my $self = {};
    my $wn;
    $className = shift;      # The name of my class.
    $self->{'errorString'} = "";
    $self->{'error'} = 0;
    $wn = shift;      # The WordNet::QueryData object.
    $self->{'wn'} = $wn;
    if(!$wn)
    {
        $self->{'errorString'}
            .= "\nError (WordNet::Similarity::edge->new()) - ";
        $self->{'errorString'}
            .= "A WordNet::QueryData object is required.";
        $self->{'error'} = 2;
    }
    bless($self, $className);
```

```
    $self->_initialize(shift) if($self->{'error'} < 2);
    $self->{'traceString'} = "";
    return $self;
}

sub _initialize # Initialization WordNet::Similarity::edge object
{
    my $self;
    my $paramFile;
    my $infoContentFile;
    my $wn;
    $self = shift;
    $wn = $self->{'wn'};
    $paramFile = shift;
    $self->{"n"} = 1;
    $self->{"v"} = 1;
    $self->{'doCache'} = 1;
    $self->{'simCache'} = ();      # Initialize the cache stuff.
    $self->{'traceCache'} = ();
    $self->{'cacheQ'} = ();
    $self->{'maxCacheSize'} = 1000;
    $self->{'trace'} = 0;      # Initialize tracing.
    $self->{'traceString'} = "" if($self->{'trace'});
    if(defined $paramFile)
    {
        my $modname;
        if(open(PARAM, $paramFile))
        {
            $modname = <PARAM>;
            $modname =~ s/[\r\f\n]//g;
            $modname =~ s/\s+//g;
            if($modname =~ /^WordNet::Similarity::edge/)
            {
                while(<PARAM>)
                {
                    s/[\r\f\n]//g;
                    s/\#.*//;
                    s/\s+//g;
                    if(/^trace::(.*)/)
                    {
                      my $tmp = $1;
                      $self->{'trace'} = 1;
                      $self->{'trace'} = $tmp if($tmp =~ /^[012]$/);
                    }
                    elsif(/^cache::(.*)/)
```

```
                    {
                      my $tmp = $1;
                      $self->{'doCache'} = 1;
                      $self->{'doCache'} = $tmp if($tmp =~ /^[01]$/);
                    }
                    elsif(m/^(?:max)?CacheSize::(.*)/i)
                    {
                      my $mcs = $1;
                      $self->{'maxCacheSize'} = 1000;
                      $self->{'maxCacheSize'} = $mcs
                            if(defined ($mcs) && $mcs =~ m/^\d+$/);
                      $self->{'maxCacheSize'} = 0
                          if($self->{'maxCacheSize'} < 0);
                    }
                    elsif($_ ne "")
                    {
                        s/::.*//;
                        $self->{'errorString'}
                        .= "\nWarning
                     (WordNet::Similarity::edge->_initialize()) - ";
                        $self->{'errorString'}
                        .= "Unrecognized parameter '$_'. Ignoring.";
                        $self->{'error'} = 1;
                    }
                }
            }
            else
            {
                $self->{'errorString'}
                    .= "\nError
                      (WordNet::Similarity::edge->_initialize()) - ";
                $self->{'errorString'}
                .= "$paramFile does not appear to be a config file.";
                $self->{'error'} = 2;
                return;
            }
            close(PARAM);
        }
        else
        {
            $self->{'errorString'} .= "\nError
                      (WordNet::Similarity::edge->_initialize()) - ";
            $self->{'errorString'}
                        .= "Unable to open config file $paramFile.";
            $self->{'error'} = 2;
```

```
            return;
        }
    }
}

sub getRelatedness
{
    my $self = shift;
    my $wps1 = shift;
    my $wps2 = shift;
    my $wn = $self->{'wn'};
    my $pos;
    my $pos1;
    my $pos2;
    my $offset;
    my $lOffset;
    my $rOffset;
    my $lTree;
    my $rTree;
    my $lCount;
    my $rCount;
    my $leastCommonSubsumer;
    my $LCSOffset;
    my $minDist;
    my $score;
    my @lTrees;
    my @rTrees;
    if(!$wn)
    {
        $self->{'errorString'} .= "\nError
                (WordNet::Similarity::edge->getRelatedness()) - ";
        $self->{'errorString'}
                    .= "A WordNet::QueryData object is required.";
        $self->{'error'} = 2;
        return undef;
    }
    $self->{'traceString'} = "" if($self->{'trace'});
    if(!$wps1 || !$wps2)
    {
        $self->{'errorString'} .= "\nWarning
                    (WordNet::Similarity::edge->getRelatedness())
                    - Undefined input values.";
        $self->{'error'} =
            ($self->{'error'} < 1) ? 1 : $self->{'error'};
        return undef;
```

```
    }
    if($wps1 =~ /^\S+\#([nvar])\#\d+$/)
    {
        $pos1 = $1;
    }
    else
    {
        $self->{'errorString'} .= "\nWarning
                        (WordNet::Similarity::edge->getRelatedness()) - ";
        $self->{'errorString'} .= "Input not in word\#pos\#sense format.";
        $self->{'error'} = ($self->{'error'} < 1) ? 1 : $self->{'error'};
        return undef;
    }
    if($wps2 =~ /^\S+\#([nvar])\#\d+$/)
    {
        $pos2 = $1;
    }
    else
    {
        $self->{'errorString'} .= "\nWarning
                    (WordNet::Similarity::edge->getRelatedness()) - ";
        $self->{'errorString'}
                        .= "Input not in word\#pos\#sense format.";
        $self->{'error'} =
                       ($self->{'error'} < 1) ? 1 : $self->{'error'};
        return undef;
    }
    if($pos1 ne $pos2)      # Relatedness 0 across parts of speech.
    {
        $self->{'traceString'} =
            "Relatedness 0 across parts of speech ($wps1, $wps2).\n"
                if($self->{'trace'});
        return 0;
    }
    $pos = $pos1;
    if($pos !~ /[nv]/)      # Relatedness only for nouns and verbs.
    {
        $self->{'traceString'} =
         "Only verbs and nouns have hypernym trees ($wps1, $wps2).\n"
             if($self->{'trace'});
        return 0;
    }
    if($self->{'doCache'} &&
       defined $self->{'simCache'}->{"${wps1}::$wps2"})
    {
```

```
        if(defined $self->{'traceCache'}->{"${wps1}::$wps2"})
        {
            $self->{'traceString'}
            = $self->{'traceCache'}->{"${wps1}::$wps2"}
            if($self->{'trace'});
        }
        return $self->{'simCache'}->{"${wps1}::$wps2"};
    }
    $lOffset = $wn->offset($wps1);
    $rOffset = $wn->offset($wps2);
    if(!$lOffset || !$rOffset)
    {
        $self->{'errorString'} .= "\nWarning
                (WordNet::Similarity::edge->getRelatedness()) - ";
        $self->{'errorString'}
                        .= "Input senses not found in WordNet.";
        $self->{'error'} =
            ($self->{'error'} < 1) ? 1 : $self->{'error'};
        return undef;
    }
    @lTrees = &getHypernymTrees($self->{'wn'}, $lOffset, $pos);
    foreach $lTree (@lTrees)
    {
        push(@{$lTree}, $lOffset);
    }
    @rTrees = &getHypernymTrees($self->{'wn'}, $rOffset, $pos);
    foreach $rTree (@rTrees)
    {
        push(@{$rTree}, $rOffset);
    }
    if($self->{'trace'})
    {
        $self->{'traceString'} = "";
        foreach $lTree (@lTrees)
        {
            $self->{'traceString'} .= "HyperTree: ";
            $self->_printSet($pos, @{$lTree});
            $self->{'traceString'} .= "\n";
        }
        foreach $rTree (@rTrees)
        {
            $self->{'traceString'} .= "HyperTree: ";
            $self->_printSet($pos, @{$rTree});
            $self->{'traceString'} .= "\n";
        }
```

```
    }
    $minDist = 100; # Find the smallest path in these trees.
    foreach $lTree (@lTrees)
    {
260     foreach $rTree (@rTrees)
        {
            $leastCommonSubsumer = &getLCSfromTrees($lTree, $rTree);
            $lCount = 0;
            foreach $offset (reverse @{$lTree})
265         {
                $lCount++;
                last if($offset == $leastCommonSubsumer);
            }
            $rCount = 0;
270         foreach $offset (reverse @{$rTree})
            {
                $rCount++;
                last if($offset == $leastCommonSubsumer);
            }
275         if($rCount + $lCount - 1 < $minDist)
            {
                $minDist = $lCount + $rCount - 1;
                $LCSOffset = $leastCommonSubsumer;
            }
280     }
    }
    if($self->{'trace'})
    {
        $self->{'traceString'} .= "LCS: ";
285     $self->_printSet($pos, $LCSOffset);
        $self->{'traceString'} .= "  Path length: $minDist.\n";
    }
    if($minDist == 100)
    {
290     $self->{'errorString'} .= "\nWarning
                    (WordNet::Similarity::edge->getRelatedness()) - ";
        $self->{'errorString'}
                      .= "A path length of 100... is that possible??";
        $self->{'error'} =
295         ($self->{'error'} < 1) ? 1 : $self->{'error'};
        return undef;
    }
    elsif($minDist > 0)
    {
300     $score = 2*$maxdist -($minDist-1); # AENDERUNG
```

```
        if($self->{'doCache'})
        {
            $self->{'simCache'}->{"${wps1}::$wps2"} = $score;
            $self->{'traceCache'}->{"${wps1}::$wps2"} =
                $self->{'traceString'} if($self->{'doCache'}
                                          && $self->{'trace'});
            push(@{$self->{'cacheQ'}}, "${wps1}::$wps2");
            if($self->{'maxCacheSize'} >= 0)
            {
                while(scalar(@{$self->{'cacheQ'}})
                      > $self->{'maxCacheSize'})
                {
                    my $delItem = shift(@{$self->{'cacheQ'}});
                    delete $self->{'simCache'}->{$delItem};
                    delete $self->{'traceCache'}->{$delItem};
                }
            }
        }
        return $score;
    }
    else
    {
        $self->{'errorString'} .= "\nWarning
                (WordNet::Similarity::edge->getRelatedness()) - ";
        $self->{'errorString'}
                .= "Internal error while finding relatedness.";
        $self->{'error'} =
            ($self->{'error'} < 1) ? 1 : $self->{'error'};
        return undef;
    }
}

sub getTraceString # return the current trace string
{
    my $self = shift;
    my $returnString = $self->{'traceString'};
    $self->{'traceString'} = "" if($self->{'trace'});
    $returnString =~ s/\n+$/\n/;
    return $returnString;
}

sub getError # Method to return recent error/warning condition
{
    my $self = shift;
    my $error = $self->{'error'};
```

```
    my $errorString = $self->{'errorString'};
    $self->{'error'} = 0;
    $self->{'errorString'} = "";
    $errorString =~ s/^\n//;
    return ($error, $errorString);
}

sub getHypernymTrees # returns an array of hypernym trees
{
    my $wn;
    my $offset;
    my $pos;
    my $wordForm;
    my $element;
    my $hypernym;
    my @hypernyms;
    my @returnArray;
    my @tmpArray;
    $wn = shift;
    $offset = shift;
    $pos = shift;
    $wordForm = $wn->getSense($offset, $pos);
    @hypernyms = $wn->querySense($wordForm, "hype");
    @returnArray = ();
    if($#hypernyms < 0)
    {
        @tmpArray = (0);
        push @returnArray, [@tmpArray];
    }
    else
    {
        foreach $hypernym (@hypernyms)
        {
            @tmpArray =
                &getHypernymTrees($wn, $wn->offset($hypernym), $pos);
            foreach $element (@tmpArray)
            {
                push @{$element}, $wn->offset($hypernym);
                push @returnArray, [@{$element}];
            }
        }
    }
    return @returnArray;
}
```

```
sub getLCSfromTrees #get Least Common Subsumer of two hypernym trees
{
    my $array1;
    my $array2;
    my $element;
    my $tmpString;
    my @tree1;
    my @tree2;

    $array1 = shift;
    $array2 = shift;
    @tree1 = reverse @{$array1};
    @tree2 = reverse @{$array2};
    $tmpString = " ".join(" ", @tree2)." ";
    foreach $element (@tree1)
    {
        if($tmpString =~ / $element /)
        {
            return $element;
        }
    }
    return 0;
}

sub _printSet # prints to traceString the WORD#POS#(SENSE/OFFSET)
{
    my $self;
    my $wn;
    my $offset;
    my $pos;
    my $wps;
    my $opstr;
    my @offsets;

    $self = shift;
    $pos = shift;
    @offsets = @_;
    $wn = $self->{'wn'};
    $opstr = "";
    foreach $offset (@offsets)
    {
        if(defined $offset && $offset != 0)
        {
            $wps = $wn->getSense($offset, $pos);
        }
```

```
            else
            {
                $wps = "*Root*\#$pos\#1";
            }
            $wps =~ s/ +/_/g;
            if($self->{'trace'} == 2 && defined $offset && $offset != 0)
            {
                $wps =~ s/\#[0-9]*$/\#$offset/;
            }
            $opstr .= "$wps ";
        }
        $opstr =~ s/\s+$//;
        $self->{'traceString'} .= $opstr if($self->{'trace'});
    }
1;
```

Ebenfalls aus dem Paket `edge.pm` wurde durch eine kleine Änderung das im neuen Zusammenhangsmaß verwendete Paket `edge3.pm` erstellt. Die Inversion der Pfadlänge wird wie in `edge.pm` unterbunden und von der Knotenanzahl der Wert 1 subtrahiert, so daß als Rückgabewert die Zahl der durchlaufenen Kanten gegeben wird. Die Veränderungen sind so gering, daß der Quellcode hier nicht aufgeführt wird.

A.3.2 Das einfache Wahrscheinlichkeitsmaß

Aufbauend auf dem für die Berechnung des Zusammenhangsmaßes von Resnik bereitgestellten Modul (`res.pm`) wurde ein Paket für das einfache Wahrscheinlichkeitsmaß entwickelt (vgl. Abschnitt 5.2). Auch hier wurden die Kommentare aus dem Quellcode herausgenommen und die eigenen Änderungen kenntlich gemacht.

```
package WordNet::Similarity::einfWahrsch;
use strict;
use WordNet::Similarity::LCSFinder;
use WordNet::Similarity::ICFinder;                    # AENDERUNG
our @ISA = qw/WordNet::Similarity::LCSFinder/;
our $VERSION = '0.07';
sub getRelatedness
{
  my $self = shift;
  my $wps1 = shift;
  my $wps2 = shift;
```

```
  my $wn = $self->{wn};
  my $class = ref $self || $self;
  unless ($wn) {
    $self->{errorString} .= "\nError
                             (${class}::getRelatedness()) - ";
    $self->{errorString} .= "A WordNet::QueryData object
                                                 is required.";
    $self->{error} = 2;
    return undef;
  }
  my $ret = $self->parseWps ($wps1, $wps2);
  ref $ret or return $ret;
  my ($word1,$pos1,undef,$offset1,$word2,$pos2,undef,$offset2)
      = @{$ret};
  $self->{traceString} = "";
  my $pos = $pos1;
  my $relatedness =
    $self->{doCache} ? $self->fetchFromCache ($wps1, $wps2) : undef;
  defined $relatedness and return $relatedness;
  $self->{traceString} = "";
  unless ($offset1 and $offset2) {
    $self->{errorString} .= "\nWarning
                               (${class}::getRelatedness()) - ";
    $self->{errorString} .= "Input senses not found in WordNet.";
    $self->{error} = ($self->{'error'} < 1) ? 1 : $self->{'error'};
    return undef;
  }
  my @LCSs = $self->getLCSbyPath ($offset1,$offset2,$pos1,"offset");
                                                 # AENDERUNG
  my $ref = shift @LCSs;
  unless (defined $ref) {
    return $self->UNRELATED;
  }
  my ($lcs, undef) = @{$ref};                    # AENDERUNG
  my $score = 1 - probability($lcs);             # AENDERUNG
  $self->{doCache} and $self->storeToCache ($wps1, $wps2, $score);
  return $score;
}
1;
```

A.3.3 Das normalisierte Maß von Leacock und Chodorow

Da das Maß von Lin nicht auf das Intervall [0,1] abbildet, wird durch die Division durch den maximal zu erreichenden Wert das Ergebnis normalisiert. Die Veränderung ist im Quellcode kenntlich gemacht.

```
package WordNet::Similarity::lch2;
use strict;
use Exporter;
use WordNet::Similarity::LCSFinder;
our @ISA = qw/WordNet::Similarity::LCSFinder/;
our $VERSION = '0.07';
sub setPosList
{
  my $self = shift;
  $self->{n} = 1;
  $self->{v} = 1;
}
sub getRelatedness
{
    my $self = shift;
    my $wps1 = shift;
    my $wps2 = shift;
    my $wn = $self->{wn};
    my $class = ref $self || $self;
    unless ($wn) {
        $self->{errorString} .= "\nError
                                 (${class}::getRelatedness()) - ";
        $self->{errorString} .=
                        "A WordNet::QueryData object is required.";
        $self->{error} = 2;
        return undef;
    }
    $self->{traceString} = "";
    my $ret = $self->parseWps ($wps1, $wps2);
    ref $ret or return $ret;
    my ($word1,$pos1,$sense1,$offset1,$word2,$pos2,$sense2,$offset2)
      = @{$ret};
    my $pos = $pos1;
    my $relatedness =
      $self->{doCache} ? $self->fetchFromCache ($wps1,$wps2) : undef;
    defined $relatedness and return $relatedness;
    my @LCSs = $self->getLCSbyPath($offset1,$offset2,$pos1,'offset');
    unless (defined $LCSs[0]) {
```

```
        return $self->UNRELATED;
    }
    my $maxdepth = -1;
    my $length;
    foreach (@LCSs) {
        my $lcs;
        ($lcs, $length) = @{$_};
        my @roots = $self->getTaxonomies ($lcs, $pos1);
        foreach my $root (@roots) {
            my $depth = $self->getTaxonomyDepth ($root, $pos1);
            unless (defined $depth) {
                $self->{error} = $self->{error} < 1 ? 1 : $self->{error};
                $self->{errorString} .="\nWarning
                                        (${class}::getRelatedness()) - ";
                $self->{errorString} .=
                                    "Taxonomy depth for $root undefined.";
                return undef;
            }
            $maxdepth = $depth if $depth > $maxdepth;
        }
    }
    if ($maxdepth <= 0) {
        $self->{error} = $self->{error} < 1 ? 1 : $self->{error};
        $self->{errorString} .= "\nWarning
                                        (${class}::getRelatedness()) - ";
        $self->{errorString} .= "Max depth of taxonomy is not positive.";
        return undef;
    }
    my $score = (log(2*$maxdepth/$length))/(log(2*$maxdepth)); # AENDERUNG
    $self->storeToCache ($offset1, $offset2, $score);
    return $score;
}
1;
```

A.3.4 Das normalisierte Maß von Jiang und Conrath

Auch bei dem Maß von Jiang und Conrath wurden im Perl-Paket Änderungen vorgenommen, um Ergebnisse im Intervall [0,1] zu bekommen. Zusätzlich wird bei der Auswertung der berechnete Wert (die „Distanz“ im Netz) durch die Division durch 20 normalisiert und durch die Subtraktion von 1 zu einem Zusammenhangsmaß umgerechnet.

```
package WordNet::Similarity::jcnnormalisiert;
use strict;
use warnings;
use Exporter;
use WordNet::Similarity::LCSFinder;
our (@ISA, @EXPORT, @EXPORT_OK, %EXPORT_TAGS);
@ISA = qw(WordNet::Similarity::LCSFinder);
%EXPORT_TAGS = ();
@EXPORT_OK = ();
@EXPORT = ();
our $VERSION = '0.07';
sub getRelatedness
{
  my $self = shift;
  my $wps1 = shift;
  my $wps2 = shift;
  my $wn = $self->{wn};
  my $class = ref $self || $self;
  unless ($wn) {
    $self->{errorString} .= "\nError
                             (${class}::getRelatedness()) - ";
    $self->{errorString} .=
                    "A WordNet::QueryData object is required.";
    $self->{error} = 2;
    return undef;
  }
  $self->{traceString} = "";
  my $ret = $self->parseWps ($wps1, $wps2);
  ref $ret or return $ret;
  my ($word1,$pos1,undef,$offset1,$word2,$pos2,undef,$offset2)
      = @{$ret};
  my $pos = $pos1;
  my $relatedness =
    $self->{doCache} ? $self->fetchFromCache ($wps1,$wps2):undef;
  defined $relatedness and return $relatedness;
  my $mode = 'offset';
  my @LCSs = $self->getLCSbyIC ($offset1,$offset2,$pos,'offset');
  my $ref = shift @LCSs;
  unless (ref $ref) {
    return $self->UNRELATED;
  }
  my ($lcs, $lcsic) = @{$ref};
  my $lcsfreq = $self->getFrequency ($lcs, $pos, 'offset');
  my $maxScore;
  my $rootFreq = $self->getFrequency (0, $pos, 'offset');
```

```
    if($rootFreq) {
      $maxScore = 2 * -log (0.001 / $rootFreq) + 1;
    }
    else {
      $self->{errorString} .= "\nWarning
                               (${class}::getRelatedness()) - ";
      $self->{errorString} .= "Root node has a zero
                                frequency count.";
      $self->{error} = ($self->{error} < 1) ? 1 : $self->{error};
      return 0;
    }
    my $ic1 = $self->IC($offset1 , $pos);
    my $ic2 = $self->IC($offset2 , $pos);
    if ($self->{trace}) {
      $self->{traceString} .= "Concept1: ";
      $self->printSet ($pos, $mode, $offset1);
      $self->{traceString} .= " (IC=";
      $self->{traceString} .= sprintf ("%.6f", $ic1);
      $self->{traceString} .= ")\n";
      $self->{traceString} .= "Concept2: ";
      $self->printSet ($pos, $mode, $offset2);
      $self->{traceString} .= " (IC=";
      $self->{traceString} .= sprintf ("%.6f", $ic2);
      $self->{traceString} .= ")\n";
    }
    my $distance;
    if($ic1 && $ic2) {
      my $ic3 = $self->IC($lcs , $pos);
      $distance = $ic1 + $ic2 - (2 * $ic3);
    }
    else {
      return 0;
    }
  # Folgende Berechnung des Zusammenhangs geloescht.
    return $distance; # Rueckgabe des Abstands
  }
  1;
```

A.3.5 Das Maß von Resnik

Das Maß von Resnik bildet nicht auf das Intervall [0,1] ab. Der oberste Wert für sein Maß ist $ln(N)$ mit N als der Summe aller Frequenzen aus der jeweils verwendeten Frequenzdatei. Im vorliegenden Fall ist es die Datei, die jedem

Knoten die Zahl seiner Unterknoten (er selbst eingeschlossen) zuordnet (es ergibt sich bei der Frequenzsumme 200.003 damit als maximaler Wert 12.20608765). Dieser maximale Wert wird allerdings nicht erreicht.

Die Normalisierung dieses Maßes, d.h. die Abbildung in das Intervall [0,1] macht keinen Sinn, da in der Rechnung der Informationsgehalt des kleinsten gemeinsamen Ahnenknotens verwendet wird, der sich nicht linear zur Länge des Verbindungspfades verhält. Eine lineare Abbildung in das Einheitsintervall würde die Werte wesentlich verfälschen. Es würde z.B. zwei Elementen desselben *Synsets* im oberen Teil der Hierarchie nicht der intuitive Zusammenhangswert 1 zugeordnet, sondern der Informationsgehalt ihres Konzeptknotens, der – wegen der Hierarchietiefe – eher gering ist. Den Zusammenhangswert 1 würden nur zwei Elemente desselben Blattknotens erhalten.

A.4 Das erstellte Korpusmaß

Mithilfe des Programms `korpusmass.pl` wird für Paare, die aus zwei Quelldateien zusammengestellt werden, der Zusammenhang im erweiterten Netz berechnet.[3] Dort liegen die zu vergleichenden Konzepte im *wps*-Format vor (*word*, *part of speech* und *sense number*).
Ein Ausschnitt aus `SOURCE.txt`:

```
voiture#n#1
fenêtre#n#1
table#n#2
```

Ein Ausschnitt aus `TARGET.txt`:

```
verre#n#1
ordinateur#n#1
```

In der Datei `Knoten.txt` (siehe Abschnitt A.5.2) findet sich die Erweiterung für das semantische Netz, dessen Datenbank durch die Erweiterung nicht verändert werden soll).

[3] Die detaillierte Beschreibung des Maßes findet sich in Kapitel 10.

Die Informationen aus den Quelldateien werden in verschiedene *Hashes* eingelesen und im Laufe des Programms ausgewertet. Dann folgt die eigentliche Berechnung des Zusammenhangsmaßes für die Elemente des *Sourcearray* und des *Targetarray*.

Für das neue Maß wird bei jedem Element im Knotenarray nachgeprüft, ob hier eine Erweiterung des Netzes besteht. Wenn für das zu berechnende Ausgangselement (z.B. <père#n#3>) im Knotenarray ein Eintrag existiert, wird die vorliegende Gewichtung eingelesen und als weiteres mögliches Ausgangselement mit dem verknüpften Eintrag (hier <maison#n#3>) ebenfalls eine Berechnung durchgeführt.

Für die Knotenpaare wird die Länge der Vebindungspfade und die Gewichtung durch die jeweilig vorliegenden Werte des *Informationcontent* berechnet. Aus diesen Werten wird der geringste Wert (d.h. die beste Kombination von *Source*- und *Target*-element) bestimmt und vom maximalen Abstand der Taxonomie (hier 32) abgezogen und normalisiert.

Schließlich werden alle Werte zu einem besseren Vergleich in eine einzige Auswertungstabelle geschrieben, die gleich als LaTeX-formatierte Tabelle vorliegt.

Programmcode zum Korpusmaß

```
#!/usr/bin/perl
#use strict;

use WordNet::Similarity::edge3;
use WordNet::Similarity::lin;
use WordNet::QueryData;

open(OUT, ">Auswertung.dat") or
    die("Kann Auswertung.dat nicht oeffnen."); # Ausgabe-DATEI
open(LOG, ">Auswertung.log") or
    die("Kann Logdatei nicht oeffnen.");       # LOG-DATEI
open(SOURCE, "<SOURCE.txt") or
    die("Kann SOURCE.txt nicht oeffnen.");    ;
open(TARGET, "<TARGET.txt") or
    die("Kann TARGET.txt nicht oeffnen.");    ;

my @sourcearray;
my @targetarray;
my $i = 0;
```

```
20
while(<SOURCE>)
{
    if ($_ =~ /^(.*?\#n|#|d*?)|n/)
    {
        $sourcearray[$i] = $1;
    }
    $i++;
}

$i = 0;
while(<TARGET>)
{
    if ($_ =~ /^(.*?\#n|#|d*?)|n/)
    {
        $targetarray[$i] = $1;
    }
    $i++;
}

close SOURCE;
close TARGET;

my %knotenhash;
$i = 0;
open(IN, "<Knoten.txt");
while (<IN>)
{ $i++;
  if ($_ =~ /^([\s\w\d\-\_\@\"áàéèîìóòúùâêîôûïç\/\:\+\.]*?
  \#n\#\d*?)\s([\s\w\d\-\_\@áàéèîìóòúùâêîôûïç\"\/\:\+\.]*?
                \#n\#\d*?)\s(\d)$/)
  {
      my $ausgangslemma = $1;
      my $ziellemma = $2;
      my $wert = $3;
      my $schluessel = $1."\$".$3;
      push(@{$knotenhash{$schluessel}}, $ziellemma);
  }
  else
  {
  }
}

$sourceinhalt = $#sourcearray +1;
$targetinhalt = $#targetarray + 1;
```

```
$produkt = $sourceinhalt * $targetinhalt;
print STDOUT "$sourceinhalt Sourceelemente,
     $targetinhalt Targetelemente: $produkt Durchgaenge.\n";
# Jedes Sourceelement wird mit jedem Targetelement
# verglichen.

$v = $produkt;
for ($s = 0; $s <= $#sourcearray; $s++)
{
    my $lex1 = $sourcearray[$s];
    my $ewn = WordNet::QueryData->new;
    for ($m = 0; $m <= $#targetarray; $m++)
    {
        my $zwischenwert = 50;  # Anzahl Kanten
        my $lex2 = $targetarray[$m];
        print STDOUT "$v: ";
        $v--;
        print STDOUT "Berechnung: $lex1 und $lex2.\n";
        my $wn = WordNet::QueryData->new();
        my @synset1array = ();
        my @synset2array = ();

# aus knotenhash werden entsprechende arrays herausgeholt
        foreach (keys(%knotenhash))
        {
            $arraystring = join(" ", @{$knotenhash{$_}});
            if ($arraystring =~ /^.*?$lex1.*?$/)
# evt. soll fuer $lex1 mit schluessel gerechnet werden
            {
                push (@synset1array, $_);
            }
            elsif ($arraystring =~ /^.*?$lex2.*?$/)
# evt. soll fuer $lex2 mit schluessel gerechnet werden
            {
                push (@synset2array, $_);
            }
            else
            {
            }
        }

# zu Arrays jeweils noch Originaleintraege holen
# haben schon Bedeutungsnummer
        push (@synset1array, $lex1);
        push (@synset2array, $lex2);
```

```
# liste der knoten fuer 1. wort (Original und Neu)
        for ($i = 0; $i<=$#synset1array; $i++)
        {
            $werteschalter1 = 0;
            $wertgewicht1 = 0;
            my $syn1;

            if ($synset1array[$i]
                =~ /^([\s\w\dáàéèîíóòúùâêîôûïç\-\/\:\+\.]*?
                    \#n\#\d*?)\$(\d)/)
            {
# evt. Wert der Verknuepfung aufnehmen, zu schalter addieren
                $syn1 = $1;
                $wertgewicht1 =+ scalar($2);
                $werteschalter1++;
            }
            else
            {
# ist Kante ohne Gewicht
                $syn1 = $synset1array[$i];
            }

            for ($j = 0; $j<=$#synset2array; $j++)
            {
                $wertgewicht2 = 0;
                $werteschalter2 = 0;
                my $syn2;
                if ($synset2array[$j]
                  =~ /^([\s\w\dáàéèîíóòúùâêîôûïç\-\/\:\+\.]*?
                      \#n\#\d*?)\$(\d)/)
                {
# evt. Wert der Verknuepfung aufnehmen, zu schalter addieren
                    $syn2 = $1;
                    $wertgewicht2 =+ scalar($2);
                    $werteschalter2++;
                }
                else
                {
                    $wertgewicht2 = 0;
                    $syn2 = $synset2array[$j];
                }

                my $endwert = $wertgewicht1 + $wertgewicht2;
# umgeandertes edge
                my $mymeasure =
```

```
155                     WordNet::Similarity::edge3->new($wn);
                    my $sum =
                      $mymeasure->getRelatedness("$syn1","$syn2");
  # noch die zwei zusaetzlichen Kanten zuzaehlen
                    $sum = $sum+$werteschalter1+$werteschalter2;
160
                    ($error,$errorString)=$mymeasure->getError();
                    die "$errorString\n" if ($error);

                    my $wn2 = WordNet::QueryData->new();
165                 my $mymeasure2 =
                      WordNet::Similarity::lin->new($wn2);
                    my $faktor =
                     $mymeasure2->getRelatedness("$syn1","$syn2");

170                 if (scalar($sum) <  scalar($zwischenwert))
                    {
                        @zwischenarray = (); # wieder leeren
                        $zwischenwert = $sum;
                        my $zwischensyn1 = $syn1;
175                     my $zwischensyn2 = $syn2;
                        @zwischenarray[0] = "$zwischensyn1
                    $zwischensyn2 $zwischenwert $faktor $endwert";
                    }
                    elsif (scalar($sum) ==  scalar($zwischenwert))
180                 {
                        my $zwischensyn1 = $syn1;
                        my $zwischensyn2 = $syn2;
                        push (@zwischenarray, "$zwischensyn1
                   $zwischensyn2 $zwischenwert $faktor $endwert");
185                 }
                    else
                    {
                    }
                }# in synset2array-for-schleife
190             }# in synset1array-for-schleife

                    my $zgew = -1;
            my $zpfad = 33;
            my $l= 0;
195         foreach ($k = 0; $k<=$#zwischenarray; $k++)
            {
                (undef, undef, $pfad, $gew, undef) =
                    split(" ",$zwischenarray[$k]);
                if ($zpfad > $pfad)  # kuerzesten Pfad suchen
```

```
            {
                $zpfad = $pfad;  # als neuer Wert
                if ($zgew < $gew)
                {
                    $zgew = $gew;  # hoechstes Gewicht
                    $l = $k;
                }
                else
                {
                }
            }
            else
            {
            }
        }

        ($lex1, $lex2, $laenge, $fak, $gewicht) =
            split(" ",$zwischenarray[$l]);
        $mass = (32 - $laenge -$gewicht/3)*$fak/32;

        $lex1 = $sourcearray[$s];
        $lex2 = $targetarray[$m];
        $wortt1 = synsetausgabe("$lex1");
        $wortt2 = synsetausgabe("$lex2");
        print OUT "$lex1 und $lex2: ".runden($mass)."\n";
        $lex1 = $sourcearray[$s];
    }
}

close LOG;
close OUT;

sub runden
{
    my ($a) = @_;
    $b = sprintf("%.4f", $a);
    return $b;
}

sub synsetausgabe
{
    my($lex) = @_;
    if ($lex =~ /^(.*?)\#n\#(\d*?)$/)
    {
        $lex = $1.":".$2;
```

```
245     }
        return $lex;
}
```

A.5 Vergleich mit anderen Maßen

Für den Vergleich der diskutierten Maße mit dem neu erstellten Maß (vgl. detaillierte Diskussion in 11.2) werden die Dateien `FranzLexeme.dat` und `Knoten.txt` (vgl. Abschnitt A.5.2) verwendet.

Auch hier werden die Ergebnisse der besseren Vergleichbarkeit halber in eine LaTeX-formatierte Tabelle geschrieben. Die Werte, die Miller und Charles ermittelt haben, werden in ein eigenes Array eingelesen. Dann werden aus der Quelldatei `FranzLexeme.dat` die zu untersuchenden Paare eingelesen und für jedes Maß die Zusammenhangsberechnung durchgeführt.

Für die Berechnung der Abweichungen werden die einzelnen Werte mit den Werten von Miller und Charles verglichen. Die Ergebnisse werden in eine eigene Tabelle geschrieben.[4]

A.5.1 Programmcode für den Vergleich

```
#!/usr/bin/perl
#use strict;

use WordNet::Similarity::edge; # LADEN DER MODULE
use WordNet::Similarity::edge3;
use WordNet::Similarity::res;
use WordNet::Similarity::hso;
use WordNet::Similarity::jcnnormalisiert;
use WordNet::Similarity::lchneu;
use WordNet::Similarity::lin;
use WordNet::QueryData;
use Time::Local;
use POSIX qw(strftime);
my $day;
my $month;
my $year;
$tm = localtime;
($sec, $min, $stunde, $day, $month, $year) = (localtime)[0..5];
```

[4] Für den Vergleich der 25 Paare mit allen 10 Maßen muß das vorliegende Programm nur wenig ergänzt werden.

```
$anfang = timelocal($sec, $min, $stunde, $day, $month, $year);

open(OUT, ">MassVergleich.tex")
    or die("Kann Zieldatei nicht oeffnen.");
print OUT "\\NeedsTeXFormat{LaTeX2e}\\input{Vorspann}
                                        \\begin{document}";
print OUT "\%";
printf OUT ("Auswertungsbeginn: vom %02d.%02d.%04d,
             %02d:%02d:%02d Uhr.\n",
             $day, $month+1, $year+1900, $stunde, $min, $sec);
print OUT "\\small\n";
print OUT "\\setlongtables\n";
print OUT "\\begin{longtable}{ |l|l|l|l|l|l|l| }\n";
print OUT "\\caption*{Auswertungsdatei}
                                          \\\\ \\hline \n";
print OUT "\\textbf{Graphien} \& \\textbf{MC} \&
           \\textbf{HSO} \& \\textbf{LCH} \&
           \\textbf{JCN} \& \\textbf{LIN} \&
           \\textbf{Korpus}\\\\ \\hline \\endhead \n";
print OUT " \\endfoot\n";
open(VERGLEICH, "<FranzLexeme.dat")
 or die("Kann FranzLexeme.dat fuer Auswertungspaare nicht oeffnen.");

my @sourcearray;
my @targetarray;
my $i = 0;
my @MCarray = (0.98, 0.96, 0.94, 0.925, 0.9025, 0.875, 0.7375,
               0.275, 0.2375, 0.2225, 0.1575);
my @hsoarray;
my @lcharray;
my @jcnarray;
my @linarray;
my @korpusarray;


while(<VERGLEICH>)
{
    if ($_ =~ /^(.*?\#n\#|d*?)\s(.*?\#n\#|d*?)$/)
    {
        $sourcearray[$i] = $1;
        $targetarray[$i] = $2;
    }
    $i++;
}
```

```
close VERGLEICH;

$sourceinhalt = $#sourcearray +1;
$targetinhalt = $#targetarray + 1;
print STDOUT "$sourceinhalt Sourceelemente, $targetinhalt
              Targetelemente: $sourceinhalt Durchgaenge.\n";
print OUT "\%";
print OUT "$sourceinhalt Sourceelemente, $targetinhalt
           Targetelemente: $sourceinhalt Durchgaenge.\n";
$v = $sourceinhalt;

for ($s = 0; $s <= $#sourcearray; $s++)
{
    my $lex1 = $sourcearray[$s];
    my $ewn = WordNet::QueryData->new;
    my $lex2 = $targetarray[$s];
    print STDOUT "$v: ";
    $v--;
    print STDOUT "Berechnung von $lex1 und $lex2.\n";

# BERECHNUNG DER WERTE
# LIN: [0,1] nach Aenderung in lch.pm
    my $wn = WordNet::QueryData->new();
    my $lchmeasure = WordNet::Similarity::lchneu->new($wn);
    my $lch = $lchmeasure->getRelatedness("$lex1", "$lex2");
    ($error, $errorString) = $lchmeasure->getError();
    die "$errorString\n" if ($error);
    my $zwischenlch = $lch;
    $lch = $zwischenlch;

# JCN: [0,1] nach Aenderung in jcn.pm
    my $jcnmeasure = WordNet::Similarity::jcn0512->new($wn);
    my $jcn = $jcnmeasure->getRelatedness("$lex1", "$lex2");
    ($error, $errorString) = $jcnmeasure->getError();
    die "$errorString\n" if ($error);
    my $zwischenjcn = $jcn/20;
    my $zwischenjcn2 = 1-$zwischenjcn;
    $jcn = $zwischenjcn2;

# HSO: [0,1] nach Aenderung
    my $hsomeasure = WordNet::Similarity::hso->new($wn);
    my $hso = $hsomeasure->getRelatedness("$lex1", "$lex2");
    ($error, $errorString) = $hsomeasure->getError();
    die "$errorString\n" if ($error);
    my $zwischenhso = $hso/16;
```

```
    $hso = $zwischenhso;

# LIN:
    my $linmeasure = WordNet::Similarity::lin->new($wn);
    my $lin = $linmeasure->getRelatedness("$lex1", "$lex2");
    ($error, $errorString) = $linmeasure->getError();
    die "$errorString\n" if ($error);
    my $zwischenlin = $lin;
    $lin = $zwischenlin;

# KORPUS:
    my $mymeasure = WordNet::Similarity::edge3->new($wn);
    my $sum = $mymeasure->getRelatedness("$lex1", "$lex2");
    $sum = $sum-1;
    ($error, $errorString) = $mymeasure->getError();
    die "$errorString\n" if ($error);
    my $faktor = $lin;  # schon berechnet
    $korpus = (32 - $sum)*$faktor/32;

    $lex1 = $sourcearray[$s];
    $lex2 = $targetarray[$s];
    $lexx1 = synsetausgabe("$lex1");
    $lexx2 = synsetausgabe("$lex2");

# AUSGABE
    print OUT "$lexx1 $lexx2 "
        ." \& ".$MCarray[$s]
            ." \& ".runden($hso)
                ." \& ".runden($lch)
                    ." \& ".runden($jcn)
                        ." \& ".runden($lin)
                            ." \& ".runden($korpus)
                                ." \\\\ \\hline\n";

    push (@hsoarray, runden($hso));
    push (@linarray, runden($lin));
    push (@lcharray, runden($lch));
    push (@jcnarray, runden($jcn));
    push (@korpusarray, runden($korpus));
}


print OUT "\\end{longtable}\n";
$tm = localtime;
($sec, $min, $stunde, $day, $month, $year) = (localtime)[0..5];
```

```
$ende = timelocal($sec, $min, $stunde, $day, $month, $year);
print OUT "\%";
printf OUT ("Auswertungsende: vom %02d.%02d.%04d,
             %02d:%02d:%02d Uhr.\n", $day, $month+1,
            $year+1900, $stunde, $min, $sec);

# BERECHNUNG DER ABWEICHUNGEN

print OUT "\\small\n";
print OUT "\\setlongtables\n";
print OUT "\\begin\{longtable\}\{ \|l\|l\| \}\n";
print OUT "\\caption\*\{Zusammenstellung der Abweichung\}
           \\\\ \\hline \n";
print OUT "\\textbf\{Zusammenhangsma\{\\ss\}e\} \&
           \\textbf\{Abweichung von den Werten von
           Miller/Charles\}\\\\ \\hline \\endhead \n";
print OUT " \\endfoot\n";

$hsoabw = abweichung(@hsoarray);
$lchabw = abweichung(@lcharray);
$jcnabw = abweichung(@jcnarray);
$linabw = abweichung(@linarray);
$korpusabw = abweichung(@korpusarray);

print OUT "Hirst/St-Onge \& "      .runden($hsoabw)
                                   ." \\\\ \\hline\n";
print OUT "Leacock/Chodorow \& "   .runden($lchabw)
                                   ." \\\\ \\hline\n";
print OUT "Jiang/Conrath \& "      .runden($jcnabw)
                                   ." \\\\ \\hline\n";
print OUT "Lin \& "                .runden($linabw)
                                   ." \\\\ \\hline\n";
print OUT "Korpus \& "             .runden($korpusabw)
                                   ." \\\\ \\hline\n";
print OUT "\\end\{longtable\}\n";


print OUT "\%";
$insgesamt = $ende - $anfang;
$seconds = $insgesamt %60;
$insgesamt = ($insgesamt - $seconds) /60;
$minutes = $insgesamt %60;
$insgesamt = ($insgesamt - $minutes) /60;
$hours = $insgesamt %24;
$insgesamt = ($insgesamt - $hours) /24;
```

```
$days = $insgesamt %7;
$weeks = ($insgesamt - $days) /7;
print OUT "Insgesamt: $weeks Wochen, $days Tage,
           $hours:$minutes;$seconds Stunden)\n";
print OUT "\\end\{document\}";
close OUT;


sub runden
{
    my ($a) = @_;
    $b = sprintf("%.4f", $a);
    return $b;
}

sub synsetausgabe
{
    my($lex) = @_;
    if ($lex =~ /^(.*?)\#n\#(\d*?)$/)
    {
        $lex = $1.":".$2;
    }
    return $lex;
}

sub abweichung
{
    @array = @_;
    print STDOUT join(" ", @_)."\n";
    my $zaehler = 0;
    my $nenner = 0;
    for ($i = 0; $i<=$#array; $i++)
    {
        $einzeln = $array[$i] - $MCarray[$i];
        if ($einzeln < 0)
        {
            $einzeln = $einzeln*(-1);
        }
        $quadrat = $einzeln*$einzeln;
        $zaehler += $einzeln;
        $nenner += $quadrat;
    }
    $abw = 1-$zaehler/11;
    return $abw;
}
```

A.5.2 Quelldateien

In diesem Abschnitt findet sich die Beschreibung der Quelldateien, die von den Auswertungsprogrammen aufgerufen werden.

Knoten.txt

In der Datei `Knoten.txt` sind die neuen Knoten des Netzwerkes (die aus Kompabilitätsgründen nicht in die Datenbank eingearbeitet werden) extra mit ihren jeweiligen Verknüpfungsknoten aufgeführt. Der erste Eintrag ist der Netzknoten im *wps*-Format, nach einem Leerzeichen folgt der neue Knoten ebenfalls im *wps*-Format. Nach einem weiteren Leerzeichen folgt die Gewichtung der Kante (nach den Auswertungen im Kapitel 8.7).

Vollständige Datei `Knoten.txt`

maison#n#1 chambre#n#4 1	maison#n#2 ascendant#n#2 2
maison#n#1 étage#n#1 1	maison#n#2 astrologie#n#1 2
maison#n#1 salle#n#4 1	maison#n#2 astre#n#2 2
maison#n#1 bois#n#2 1	maison#n#3 enfant#n#1 1
maison#n#1 fenêtre#n#1 1	maison#n#3 lit#n#2 1
maison#n#1 quartier#n#1 2	maison#n#3 père#n#1 1
maison#n#1 pièce#n#8 2	maison#n#3 classe#n#1 2
maison#n#1 mur#n#2 2	maison#n#4 patron#n#2 1
maison#n#2 soleil#n#2 1	maison#n#4 duc#n#1 1
maison#n#2 suite#n#4 1	maison#n#4 correction#n#1 1
maison#n#2 ciel#n#2 1	maison#n#4 directeur#n#9 2
maison#n#2 mars#n#1 2	maison#n#4 production#n#2 2
maison#n#2 milieu#n#5 2	maison#n#4 contrat#n#2 2

FranzLexeme.dat

Für den Vergleich aller Maße und die Berechnung des jeweiligen Korrelationskoeffizienten, wird das Programm `massvergleich.pl` ausgeführt. Es ruft die Datei `FranzLexeme.dat` auf und berechnet für die Paare den Zusammenhangswert mit jedem hier diskutierten Maß. In dieser Datei befinden sich die von Miller

und Charles überprüften Paare in der Lesart, die von den einzelnen Maßen für die Berechnung des Zusammenhangs verwendet wurden. Eine Übernahme ohne Lesartendifferenzierung ist aus bekannten Gründen nicht möglich.

```
voiture#n#2 automobile#n#1
bijou#n#1 pierre_précieuse#n#1
croisière#n#1 voyage#n#2
garçon#n#2 gars#n#1
côte#n#5 littoral#n#2
asile_de_fous#n#1 asile_de_fous#n#2
sorcier#n#1 magicien#n#2
calorifère#n#1 cuisinière#n#3
nourriture#n#1 fruit#n#2
oiseau#n#1 coq#n#1
outil#n#1 instrument#n#4
frère#n#1 moine#n#1
mec#n#2 frère#n#1
croisière#n#1 voiture#n#2
moine#n#1 prophète#n#1
cimetière#n#2 forêt#n#1
nourriture#n#1 coq#n#1
forêt#n#1 cimetière#n#2
littoral#n#2 forêt#n#1
moine#n#1 esclave#n#6
littoral#n#1 forêt#n#1
mec#n#2 crack#n#2
verre#n#6 prestidigitateur#n#2
coq#n#1 voyage#n#2
midi#n#1 cordon#n#2
```

Anhang B

Tabellen

Interne Verknüpfungen	**Substantiv**		**Verb**		**insgesamt**	
has_hyperonym	18013	46,0%	4728	45,8%	22741	45,9%
has_hyponym	18013	46,0%	4728	45,8%	22741	45,9%
has_holonym	50	0,1%	0	0%	50	0,1%
has_holo_madeof	51	0,1%	0	0%	51	0,1%
has_holo_member	131	0,3%	0	0%	131	0,3%
has_holo_part	1067	2,7%	0	0%	1067	2,2%
has_holo_portion	1	0%	0	0%	1	0%
has_meronym	50	0,1%	0	0%	50	0,1%
has_mero_madeof	51	0,1%	0	0%	51	0,1%
has_mero_member	131	0,3%	0	0%	131	0,3%
has_mero_part	1067	2,7%	0	0%	1067	2,2%
has_mero_portion	1	0%	0	0%	1	0%
involved	2	0%	0	0%	2	0%
involved_agent	4	0%	0	0%	4	0%
involved_instrument	10	0%	0	0%	10	0%
involved_location	1	0%	0	0%	10	0%
role	2	0%	0	0%	2	0%
role_agent	4	0%	0	0%	4	0%
role_instrument	10	0%	0	0%	10	0%
role_location	1	0%	0	0%	1	0%
causes	0	0%	311	3,0%	311	0,6%
is_caused_by	0	0%	311	3,0%	311	0,6%
has_subevent	0	0%	1	0%	1	0%
is_subevent_of	0	0%	1	0%	1	0%
near_antonym	512	1,3%	242	2,3%	754	1,5%
Insgesamt		39172		10322		49494
Synsets		17826		4919		22745
Durchschnitt pro Synset		2,20		2,10		2,18

Tabelle B.1: Verbindungen in EWN. (Die nicht aufgeführten Verbindungen sind nicht eingearbeitet. Vgl. Catherin 1999, 4f..)

A	Adjectif (sauf cas Aca, Apr, Aps)
Aca	Adjectif cardinal
APr	Adjectif/participe présent
APs	Adjectif/participe passé
Adv	Adverbe
Avn	Partie d'une négation (*ne*, *n'* ou *pas*, *point*, *guère*, ... associés à *ne* ou *n'*)
Cc	Conjonction coordination
Cs	Conjonction subordination
D	Déterminant (sauf cas Dca, Dg)
Dca	Nombre cardinal déterminatif
Dg	Amalgamés (*au*/*aux*/*du*/*des*)
E	Exclamatif
Ep	Présentatif (*voici*, *voilà*, ...)
Ger	Gérondif (*en* lié à un participe présent)
Inf	Infinitif
Inj	Interjection (*ah*, *oh*, *ha*, *ho*, ...)
Int	Interrogatif
Np	Nom propre
Nu	Numéral cardinal
Ono	Onomatopée
P	Pronom (sauf cas Per, X)
Per	Pronom personnel
Pp	Préposition
Pr	Participe présent (sauf cas APr, Ger)
Ps	Participe passé (sauf cas APs)
S	Substantif
V	Verbe (sauf participes et infinitif)
R	Mot inconnu du logiciel
X	Mot non traité (*que*/*qu'*, *où*, *sinon*)

Tabelle B.2: Annotation in *Frantext*

241	≫	1	sirprise	1	gloi
134	que	1	shirts	1	già
91	qu’	1	serê	1	gemütlich
69	où	1	scénariquement	1	frustrantes
38	comme	1	scampi	1	flicaille
7	mémé	1	sacristes	1	ez
4	vécue	1	s	1	eye
4	tou	1	répétaisje	1	extra
4	etc	1	risi	1	espingo
4	e	1	ridero	1	entaulé
3	the	1	restau	1	déssirer
3	plou	1	refoutes	1	désinvoltement
3	gratos	1	ragoteuse	1	désamour
2	ê	1	r	1	dérangeante
2	émile	1	queûh	1	déje
2	°	1	pécho	1	déconnades
2	sniffer	1	purple	1	desservantes
2	poulaga	1	prépes	1	dealer
2	gnagna	1	provises	1	dam’
2	cra	1	printannières	1	d
2	cabitus	1	prende	1	cé
2	%xixe	1	praevalebunt	1	cré
1	éva	1	poële	1	cravetouse
1	étienne	1	plaza	1	cosa
1	épiscopat	1	pelotonnement	1	clébard
1	élitistes	1	pel	1	city
1	élisabeth	1	parrapluy	1	cinocheux
1	élie	1	palazzo	1	chorba
1	éh	1	p	1	chiatique
1	égypte	1	overdosés	1	che
1	édouard	1	néos	1	castratrices
1	zéphir	1	non	1	castagnent
1	zé	1	n	1	cassone
1	z’	1	motor	1	cara
1	z	1	meûdeûmeûh	1	buon
1	yeshivas	1	maternantes	1	buisenesslike
1	verso	1	marroniers	1	buccino
1	vape	1	maniaquement	1	bouzillé
1	uom	1	maizhon	1	bordélique
1	téchou	1	macabs	1	bonniche
1	tuperwares	1	lincélum	1	bombinettes
1	treillissé	1	labeel	1	baltes
1	trees	1	kong	1	babouchkas
1	toung	1	into	1	appart’
1	tornato	1	indédence	1	appart
1	tommettes	1	imprésentable	1	and
1	tintin	1	imaginairement	1	aliens
1	ti	1	hében	1	(?)
1	tertulias	1	hanoum	1	%xvi
1	tacete	1	habillantes	1	%xiv
1	sérait	1	h	1	%xii
1	sécu	1	gyrophare	1	%x
1	subclaquant	1	gym	1	%ii
1	street	1	guéable	1	%i
1	stersund	1	guinguois	1	%

Tabelle B.3: Nichterkannte Zeichenketten in *Frantext*

.

Anhang C

Manuelle Annotation im Subkorpus

Im folgenden werden die während der Annotation getroffenenen Unterscheidungen nach Lesartenkategorien, denen sie zugeordnet werden, zusammengefaßt.[1]

In der Lesart als **Gebäude [maison:1]** (auch im übertragenen Sinn oder als Herkunft): *sa maison était ouverte*, *le seuil de la maison*, *le genre de la maison*, *la maternité de la maison*, *la douceur de la maison*, *la maison de qn*, *entrer dans la maison*, *sortir de la maison*, *s'occuper de la maison*, *garder la maison*, *tenir la maison*.

Als Konzept aus der **Astrologie**:
Da dieses Konzept nur insgesamt zweimal auftaucht, ist hier nichts anzumerken.

Als Synonym zu **[chez-soi:2]**:
venir à la maison, *rentrer à la maison*, *le retour à la maison* (nicht ohne Einschränkung).

Als Äquivalent zum deutschen Konzept **[Firma:1]**:
maison-mère, *maison de couture*, *maison de disques*, *maison de production*, *maison de confection*, *offert par la maison*, *Maison de Mademoiselle*, *administrateur de maison*.

Innerhalb von **Komposita**:

[1] Die Groß- und Kleinschreibung wurde aus den Textstellen übernommen.

Hauszugehörige: *maîtresse de maison*, *maître de la maison*, *mère de maison*, *femme de charge de la maison*, *la Maison du Père*, *fille de la maison*, *l'homme de la maison*, *employée de la maison*, *cuisinière de la maison*, *Gens de Maison*, *dame de la maison*, *fils de la maison*, *l'enfant de la maison*, *travail de maison*, *charges de la maison*, *linge de maison*, *devoirs maison*, *carcasse de maison*.

Hausgemacht: *fabriqué maison*, *fromage maison*, *parfum maison*, *cigarres maison*, *cousu maison*, *tartes-maison*, *dessert maison*, *pot-au-feu maison*, *salaire maison*, *teinture-maison*, *bousculade maison*, *la spécialité maison*, *salsa maison*, *eye-liner maison*, *Kir maison*, *Secret maison*.

Andere: *maison-coquillage*, *tube-maison*, *passe-maison*, *bombinettes-maison*, *extra-maison*, *pelle-maison*.

Verben: *faire maison commune*, *faire grande maison*, *faire les honneurs de la maison*, *faire sa maison*.

Lexikalisierte Kombinationen: *maison chapelle*, *maison-forte*, *maison-d'édition*, *Maison de la Presse*, *Maison de la Radio*, *maison de la culture*, *Maison des Arts*, *maison de jeu*, *maison de vacances*, *maison de campagne*, *maison de paysan*, *maison de garde-chasse*, *maison de province*, *maison des palettes*, *maison des nains*, *maison de repos*, *maison de rendez-vous*, *maison de fous*, *maison de fonction de Sadec*, *maison du poulet*, *maison de maître*, *maison de ville*, *maison du gouverneur*, *maison de putains*, *Maison des écrivains*, *maison du peuple*, *maison des rendez-vous*, *maison de maître*, *maison de plaisir*, *maison de passe*, *maison de la poupée*, *Maison de l'enfant*, *maison des enfants*, *Maison des hommes*, *maison des petits moins de dix-huits ans*, *maison des jeunes*, *maison de garde-barrière*, *maison de santé*, *la maison d'arrêt*, *maison de correction et d'arrêt*, *maison (psychiatrique)*, *la maison des naissances*, *la maison de l'éternel*, *Maison des morts*.

Innerhalb von **Eigennamen**:
la Maison dorée, *Maison-Carrée*, *Maison-Blanche*, *Maison de l'Unesco*, *Maison Verte*, *Maison de Nanterre*, *Maison de France*, *Maison des Sciences*, *la maison de l'Ave Maria*, *Maison des Dames du Sacré Coeur de Jésus*, *la Maison des Trois Filles*, *la Grande Maison de Blanc*, *Grande Maison*, *maison Dargaud*, *Maison Zalapore*, *Maison Krauss*, *maison Grandet*, *maison Usher*, *maison Tellier*, *maison Marx*, *maison Gerit Wanhorsgstraten*, *maison poulaga*, *maison des Tascher*,

maison Lalière, *maison Montauquier*, *Maison de Molière*, *Maison Rose*, *Maison Lancia*, *Maison Pottier*, *maison Royco*, *Maison de Bernarda*, *Maison Dieu*, *Maison Malaussène*, *Maison Kiravi*, *Maison Staline*, *maison Chenue*, *Maison Deberny-Peignot*, *La Maison Grise*, *maison capétitienne*, *Maison royale de Suède*. Auch die metasprachliche Verwendung von *maison* wurde hier dazugezählt.

Anhang D

Korpusnachweis

Die hier aufgeführten Angaben sind die Informationen aus dem *Frantext*-Korpus. Die verwendeten Abkürzungen stehen für die Textgattung: *roman* (R), *essai* (E) und *traité* (T). Zusätzlich gibt *Frantext* noch weitere Angaben zur Gattung: *prose* (P) und *critique littéraire* (CL). In eckigen Klammern sind die internen Korpuskürzel angegeben. Bei mehrfachen Jahresangaben ist eine Neuauflage für das *Frantext*-Korpus verwendet worden.[1]

1. **Aragon, L.**: *Œuvre poétique, Livre III* (1926) 1982 in: Œuvre Poétique, T. 1., Paris: Livre Club Diderot, 1989. [S077] (P, E)
2. **Aventin, C.**: *Le Cœur en poche*, Paris: Mercure de France, 1988. [R817] (P, R)
3. **Bayon**: *Le lycéen*, Paris: Quai Voltaire, 1987. [R727] (P, R)
4. **Beck, B.**: *La prunelle des yeux*, Paris: Grasset, 1986. [S373] (P, R)
5. **Beck, B.**: *Stella Corfou*, Paris: Grasset, 1988. [R827] (P, R)
6. **Belloc, D.**: *Neons*, Paris: Lieu Commun, 1987. [R668] (P, R)
7. **Belloc, D.**: *Kepas*, Paris: Lieu Commun, 1989. [R669] (P, R)
8. **Benoziglio, J.-L.**: *Cabinet portrait*, 1980, Paris: Editions du Seuil, 1981. [R670] (P, R)

[1] Die bibliographischen Angaben werden unverändert aus *Frantext* übernommen, sind daher nicht immer vollständig.

9. **Bianciotti, H.**: *Sans la miséricorde du Christ*, 1985, Paris: Gallimard, 1996. [S321] (P, R)

10. **Bianciotti, H.**: *Le pas si lent de l'amour*, Paris: Grasset, 1995. [S307] (P, R)

11. **Bienne, G.**: *Le silence de la ferme*, Etrepilly: C. de Bartillat: Presses du Village, 1986. [R725] (P, R)

12. **Bienne, G.**: *Les jouets de la nuit*, Paris: Gallimard, 1990. [R968] (P, R)

13. **Boudard, A.**: *Les Enfants de chœur*, 1982, Paris: Gallimard, 1984. [R528] (P, R)

14. **Boudard, A.**: *Mourir d'enfance*, 1995, Paris: Pocket, 1997. [S312] (P, R)

15. **Brisac, G.**: *Week-end de chasse à la mère*, Paris: Ed. de l'Olivier, 1996. [S314] (P, R)

16. **Caradec, F.**: *La Compagnie des zincs*, Paris: Ramsay, 1986. [R815] (P, R)

17. **Carrère, E.**: *La Classe de neige*, Paris: POL, 1995. [S327] (P, R)

18. **Charef, M.**: *Le thé au harem*, 1983, Paris: Gallimard, 1991. [R617] (P, R)

19. **Cluny, C. M.**: *Un jeune homme de Venise*, Paris: Gallimard, 1983. [R532] (P, R)

20. **Degaudenzi, J.-L.**: *Zone*, Paris: Fixot, 1987. [R765] (P, R)

21. **Djian, P.**: *37,2 Le matin*, 1985, Paris: J'ai Lu, 1989. [R813] (P, R)

22. **Dolto, F.**: *La cause des enfants*, 1985, Paris: R. Laffont, 1995. [S318] (P, T, E)

23. **Duras, M.**: *L'Amant*, 1984, Paris: Ed. de Minuit, 1993. [S038] (P, R)

24. **Duras, M.**: *La douleur*, 1985, Paris: Gallimard, 1993. [R882] (P, R)

25. **Embareck, M.**: *Sur la ligne blanche*, Paris: Autrement, 1984. [R752] (P, R)

26. **Ernaux, A.**: *La femme Gelée* 1981, Paris: Gallimard, 1989. [R758] (P, R)

27. **Forlani, R.**: *Gouttière*, 1989, Paris: Gallimard, 1992. [R971] (P, R)

28. **Germain, S.**: *La pleurante des rues de Prague*, 1992, Paris: Gallimard, 1994. [S284] (P, R)

29. **Giraud, R.**: *Carrefour Buci*, Paris: Le Dilettante, 1987. [R818] (P, R)

30. **Gracq, J.**: *En lisant en écrivant*, 1980, Paris: Corti, 1991. [R869] (P, E, CL)

31. **Gracq, J.**: *La forme d'une ville*, 1985, Paris: Corti, 1990. [R870] (P, E)

32. **Gracq, J.**: *Autour des sept collines*, 1988, Paris: Corti, 1991. [R866] (P, E)

33. **Gracq, J.**: *Carnets du grand chemin*, Paris: Corti, 1992. [R675] (P, E)

34. **Grèce, M.**: *De la nuit du sérail*, 1982, Paris: Gallimard, 1986. [R539] (P, R)

35. **Guibert, H.**: *Voyage avec deux enfants*, 1982, Paris: Ed. de Minuit, 1992. [R722] (P, R)

36. **Guibert, H.**: *Des aveugles*, 1985, Paris: Gallimard, 1993. [R721] (P, R)

37. **Guibert, H.**: *A l'ami qui ne m'a pas sauvé la vie*, 1990, Paris: Gallimard, 1993. [R720] (P, R)

38. **Hanska, E.**: *J'arrête pas de t'aimer*, Paris: Balland, 1981. [R768] (P, R)

39. **Hanska, E.**: *Les amants foudroyés*, Paris: Mazarine, 1984. [R769] (P, R)

40. **Hermary-Vieille, C.**: *L'Épiphanie des Dieux*, 1983, Paris: Gallimard, 1984. [R540] (P, R)

41. **Jardin, A.**: *Bille en tête*, 1986, Paris: Gallimard, 1991. [R820] (P, R)

42. **Kristeva, J.**: *Les Samouraïs*, 1990, Paris: Gallimard, 1992. [S325] (P, R)

43. **Labro, P.**: *Des bateaux dans la nuit*, 1982, Paris: Gallimard, 1995. [S313] (P, R)

44. **Lange, M.**: *Les cabines de bain*, 1982, Paris: Gallimard, 1987. [R542] (P, R)

45. **Lanzmann, J.**: *La Horde d'or*, 1994, Paris: Pocket, 1995. [S269] (P, R)

46. **Lasaygues, F.**: *Vache noire hannetons et autres insectes*, Paris: B. Barrault, 1985. [R759] (P, R)

47. **Makine, A.**: *Le testament français*, Paris: Mercure de France, 1995. [S229] (P, R)

48. **Manœuvre, P.**: *L'enfant du rock*, ARTS, Paris: J.-C. Lattes, 1985. [R823] (P, R)

49. **Matzneff, G.**: *Ivre du vin perdu*, 1981, Paris: Gallimard, 1985. [R544] (P, R)

50. **Mohrt, M.**: *Vers l'Ouest*, 1988 in: *La Maison du père suivi de vers l'Ouest*, Paris: Gallimard, 1990. [S286] (P, R)

51. **Mordillat, G.**: *Vive la sociale*, 1981, Paris: Ed. du Seuil, 1987. [R757] (P, R)

52. **Ollivier, E.**: *L'orphelin de mer*, 1982, Paris: Gallimard, 1984. [R551] (P, R)

53. **Ormesson, J. D'**: *Le vent du soir*, 1985, Paris: Le Livre de Poche, 1988. [S299] (P, R)

54. **Ormesson, J. D'**: *Tous les hommes sont fous*, 1986, Paris: Le Livre de Poche, 1989. [S300] (P, R)

55. **Ormesson, J. D'**: *Le bonheur a San Miniato*, 1987, Paris: Le Livre de Poche, 1994. [S301] (P, R)

56. **Ormesson, J. D'**: *La Douane de mer*, 1993, Paris: Gallimard, 1995. [S246] (P, R)

57. **Orsenna, E.**: *Grand amour*, 1993, Paris: Ed. du Seuil, 1995. [S037] (P, R)

58. **Page, A.**: *Tchao pantin*, Paris: Denoël, 1982. [R764] (P, R)

59. **Pennac, D.**: *La petit marchande de prose*, 1989, Paris: Gallimard, 1995. [S249] (P, R)

60. **Perec, G.**: *Ellis Island*, 1980, Paris: POL, 1995. [S290] (P, R)

61. **Perec, G.**: *Quel petit vélo à guidon chromé au fond de la cour?*, 1996, Paris: Gallimard, 1966. [S288] (P, R)

62. **Poirot-Delpech, B.**: *L'Été 36*, 1984, Paris: Gallimard, 1994. [S248] (P, R)

63. **Pouy, J.-B.**: *La clef des mensonges*, Paris: Gallimard, 1988. [R957] (P, R)

64. **Queffélec, Y.**: *Les Noces barbares*, 1985, Paris: Gallimard, 1989. [R819] (P, R)

65. **Rambaud, P.**: *La Bataille*, Paris: Grasset, 1997. [S361] (P, R)

66. **Rheims, M.**: *Les greniers de Sienne*, 1987, Paris: Gallimard, 1990. [S302] (P, R)

67. **Rochant, E.**: *Un monde sans pitié*, Paris: Gallimard, 1990. [R824] (P, R)

68. **Rolin, J.**: *L'Organisation*, Paris: Gallimard, 1996. [S328] (P, R)

69. **Romilly, J. DE**: *Les œufs de Pâques*, Paris: Ed. de Fallois, 1993. [S322] (P, R)

70. **Rouad, J.**: *Les Champs d'honneurs*, 1990, Paris: Les Editions de Minuit, 1996. [S250] (P, R)

71. **Roze, P.**: *Le chasseur Zéro*, Paris: Albin Michel, 1996. [S305] (P, R)

72. **Sabatier, R.**: *Les fillettes chantantes*, Paris: Albin Michel, 1980. [R762] (P, R)

73. **Sabatier, R.**: *David et Olivier*, Paris: Albin Michel, 1985. [R763] (P, R)

74. **Salvayre, L.**: *La puissance des mouches*, 1995, Paris: Seuil, 1997. [S326] (P, R)

75. **Sarraute, N.**: *Enfance*, 1983, Paris: Gallimard, 1995. [S126] (P, R)

76. **Seguin, F.**: *L'Arme à gauche*, Paris: Julliard, 1990. [R760] (P, R)

77. **Simon, C.**: *Les Géorgiques*, 1981, Paris: Ed. de Minuit, 1992. [S128] (P, R)

78. **Simon, C.**: *L'Acacia*, Paris: Ed. de Minuit, 1989. [S129] (P, R)

79. **Sollers, P.**: *Le cœur absolu*, 1987, Paris: Gallimard, 1991. [S317] (P, R)

80. **Sollers, P.**: *Le secret*, 1993, Paris: Gallimard, 1995. [S009] (P, R)

81. **Thérame, V.**: *Bastienne*, Paris: Flammarion, 1985. [R674] (P, R)

82. **Thorez, P.**: *Les enfants modèles*, 1982, Paris: Gallimard, 1986. [R557] (P, R)

83. **Tournier, M.**: *Le medianoche amoureux*, 1989, Paris: Gallimard, 1996. [S320] (P, R)

84. **Vergne, A.**: *L'Innocence du boucher*, Paris: J.-C. Lattes, 1984. [R828] (P, R)

85. **Weyergans, F.**: *Macaire le Copte*, 1981, Paris: Gallimard, 1984. [R559] (P, R)

86. **Yourcenar, M.**: *Un homme obscur*, 1982, in: *Œuvres romanesques*, Paris: Gallimard, 1991. [S207] (P, R)

87. **Yourcenar, M.**: *Une belle matinée*, 1982, in: *Œuvres romanesques*, Paris: Gallimard, 1991. [S208] (P, R)

Literatur

Agirre, E. und G. Rigau. 1995. A Proposal for Word Sense Disambiguation using Conceptual Distance. In: *Proceedings of the First International Conference on Recent Advances in Natural Language Processing.* Tzigov Chark, Bulgarien. 258–264. URL `http://citeseer.ist.psu.edu/agirre95proposal.html. 12.01.2005`.

Agirre, E. und G. Rigau. 1996. Word sense disambiguation using conceptual density. In: *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96).* Kopenhagen, Dänemark. 16–22. URL `http://citeseer.ist.psu.edu/agirre96word.html`.

Brockhaus, F. A. 2000. *Brockhaus - Die Enzyklopädie: in 24 Bänden (1996-1999). Online-Version.* 20., neu bearbeitete Aufl. Bibliographisches Institut, F.A. Brockhaus AB und xipolis.net.

Budanitsky, A. 1999. Lexical semantic relatedness and its application in natural language processing. Technischer Bericht CSRG390. Toronto: University of Toronto. URL `ftp.cs.toronto.edu/csrg-technical-reports/390/tr390. ps.12.01.2005`.

Budanitsky, A. und G. Hirst. 2001. Semantic Distance in WordNet: An Experimental, Application-oriented Evaluation of Five Measures. In: *Workshop on WordNet and Other Lexical Resources, in the NAACL-2000.* Pittsburgh, Pennsylvania. Ohne Paginierung. URL `http://citeseer.ist. psu.edu/budanitsky01semantic.html.12.01.2005`.

Catherin, L. 1999. The French Wordnet. Technischer Bericht 2D014, Part B3. University of Amsterdam: EuroWordNet (LE-8328). URL `http://www.illc. uva.nl/EuroWordNet/docs/FrenchWordnetPS.zip.12.01.2005`.

Catherin, L. und A. Wagner. 1998. Specification of German and French WordNets. Technischer Bericht 2D0021998. University of Amsterdam: EuroWordNet (LE-8328). URL http://www.illc.uva.nl/EuroWordNet/docs/2D002.ai,12.01.2005.

Climent, S., H. Rodríguez und J. Gonzalo. 1996. Definition of the links and subsets for nouns of the EuroWordNet project, Version 6, Final. Technischer Bericht D005, WP3.1. University of Amsterdam: EuroWordNet (LE2-4003). URL http://www.illc.uva.nl/EuroWordNet/docs/D005.ai,12.01.2005.

Collins, A. M. und E. F. Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological Review* 82(6): 407–428.

Eizirik, L. M. R., V. C. Barbosa und S. B. T. Mendes. 1993. A Bayesian-Network Approach to Lexical Disambiguation. *Cognitive Science* 17: 257–283.

Fellbaum, C. (Hg.). 1998. *WordNet: An Electronical Lexical Database and some of its applications.* Cambridge, Massachusetts: The MIT Press.

Gale, W. A., K. W. Church und D. Yarowsky. 1993. Work on statistical methods for word sense disambiguation. In: *Probabilistic Approaches to Natural Language: Papers from the 1992 AAAI Fall Symposium.* Cambridge, Massachusetts. 54–60.

Gauger, H. M. 1961. *Über die Anfänge der französischen Synonymik und das Problem der Synonymie.* Dissertation, Universität Tübingen.

Geckeler, H. 1971. *Strukturelle Semantik und Wortfeldtheorie.* München: Wilhelm Fink.

Hamp, B. und H. Feldweg. 1997. GermaNet – a Lexical-Semantic Net for German. In: *Proceedings of the ACL/EACL-97 Workshop on Automatic Information Extraction and Building Lexical Semantic Resources for NLP applications.* Madrid. 9–14. URL http://citeseer.ist.psu.edu/hamp97germanet.html.12.01.2005.

Hirst, G. 1987. Resolving lexical ambiguity computationally with spreading activation and polaroid words. In: S. Small, G. Cottrell und M. Tanenhaus (Hg.). *Lexical Ambiguity Resolution.* Los Altos: Morgan Kaufman. 73–107.

Hirst, G. und D. St.Onge. 1997. Lexical Chains as representation of context for the detection and correction malapropisms. In: C. Fellbaum (Hg.). *WordNet: An Electronic Lexical Database and some of its applications.* Cambridge, Massachusetts: The MIT Press. 305–332. URL `http://citeseer.ist.psu.edu/hirst97lexical.html.12.01.2005`.

Jiang, J. J. und D. W. Conrath. 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In: *Proceedings of the International Conference Research on Computational Linguistics (ROCLING X).* 19–33.

Kilgariff, A. und M. Palmer. 2000. Introduction to the Special Issue on SENSEVAL. *Computers and the Humanities* 34: 1–13.

Kunze, C., A. Wagner, D. Dutoit, L. Catherin, K. Pala, P. Sevecek, K. Vider, L. Paldre, H. Orav und H. Oim. 1998. First Wordnets for Base Concepts in French, German, Czeck and Estonia, Version 1, Final. Technischer Bericht 2D007, WP 3.2, WP 4.2. University of Amsterdam: EuroWordNet (LE2-4003 or LE4-8283). URL `http://www.illc.uva.nl/EuroWordNet/docs/2D007.ai,27.08.2004`.

Leacock, C. und M. Chodorow. 1998. Combining Local Context and WordNet Similarity for Word Sense Identification. In: C. Fellbaum (Hg.). *WordNet: An Electronical Lexical Database and some of its applications.* Cambridge, Massachusetts: The MIT Press. 264–283.

Legrain, M. und Y. Garnier (Hg.). 2000. *Le Petit Larousse illustré.* Paris: Larousse.

Lin, D. 1997. Using Syntactic Dependency as Local Context to Resolve Word Sense Ambiguity. In: *35th Meeting of the Association for Computational Linguistics: Proceedings of the Conference, July 1997.* 64–71. URL `http://citeseer.ist.psu.edu/lin97using.html.12.01.2005`.

Lyons, J. 1968. *Introduction to Theoretical Linguistics.* Cambridge: Cambridge University Press.

Miller, G. A. 1986a. Dictionaries in the mind. *Language and cognitive processes* 1: 171–185.

Miller, G. A. 1986b. WordNet: a dictionary browser. In: *Proceedings of the conference of the University of Waterloo Centre for the New Oxford English Dictionary: Information in Data, 1.* Waterloo, Ontario. 25–28.

Miller, G. A. 1993. Nouns in WordNet: A Lexical Inheritance System. In: G. A. Miller, C. Fellbaum, R. Beckwith, D. Gross und K. J. Miller (Hg.). *Five Papers on WordNet.* Technischer Bericht 43. Princeton University: Cognitive Science Laboratory. 10–25. URL `ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps`.

Miller, G. A. 1995. WordNet: A Lexical Database of English. *Communications of the Association for Computing Machinery (ACM)* 38(11): 39–41.

Miller, G. A. und W. G. Charles. 1991. Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes* 6(1): 1–28.

Miller, G. A. und C. Fellbaum. 1991. Semantic networks of English. *Cognition* 41: 197–229.

Miller, G. A., C. Fellbaum, R. Beckwith, D. Gross und K. J. Miller. 1993a. Five Papers on WordNet. Technischer Bericht 43. Princeton University: Cognitive Science Laboratory. URL `ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps`.

Miller, G. A., C. Fellbaum, R. Beckwith, D. Gross und K. J. Miller. 1993b. Introduction to WordNet: An On-line Lexical Database. In: G. A. Miller, C. Fellbaum, R. Beckwith, D. Gross und K. J. Miller (Hg.). *Five Papers on WordNet.* Technischer Bericht 43. Princeton University: Cognitive Science Laboratory. 1–9. URL `ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps`.

Müller, W. 1965. Probleme und Aufgaben deutscher Synonymik. *Die wissenschaftliche Redaktion. Beiträge, Aufsätze, Vorträge aus dem Bibliografischen Institut in zwangloser Folge* 1: 90–101.

Pearl, J. und S. Russel. 2004. Bayesian Networks. Technischer Bericht R-277. November 2000. University of California, Los Angeles: Cognitive Systems Laboratory. In: M. A. Arbib (Hg.). *Handbook of Brain Theory and Neural Networks.* Cambridge, Massachussets: MIT Press. 157–160.

Quillian, R. M. 1968. Semantic memory. In: M. Minsky (Hg.). *Semantic Information Processing.* Cambridge, Massachussets: MIT Press. 227–270.

Resnik, P. 1995a. Disambiguating noun groupings with respect to WordNet sense. In: D. Yarowsky und K. W. Church (Hg.). *Proceedings of the Third Workshop on Very Large Corpora.* Cambridge, Massachusetts. 54–68. URL `http://acl.ldc.upenn.edu/W/W95/W95-0105.pdf.12.01.2005`.

Resnik, P. 1995b. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, August 1995, IJCAI.* 448–453. URL `http://citeseer.ist.psu.edu/resnik95using.html.12.01.2005`.

Resnik, P. 1998. WordNet and Class-Based Probabilities. In: C. Fellbaum (Hg.). *WordNet: An Electronical Lexical Database and some of its applications.* Cambridge, Massachusetts; London, England: The MIT Press. 239 –263.

Resnik, P. 1999. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research* 11: 95–130. URL `http://citeseer.ist.psu.edu/resnik99semantic.html.12.01.2005`.

Resnik, P. und D. Yarowsky. 1997. A perspective on word sense disambiguation methods and their evaluation. In: M. Light (Hg.). *Tagging Text with Lexical Semantics: Why, What and How?* 79–86. URL `http://citeseer.ist.psu.edu/resnik97perspective.html.12.01.2005`.

Resnik, P. und D. Yarowsky. 1999. Distinguishing systems and distinguishing senses: new evaluation methods for Word Sense Disambiguation. *Natural Language Engineering* 5(2): 113–134. URL `http://citeseer.ist.psu.edu/resnik98distinguishing.html.12.01.2005`.

Rey, A. (Hg.). 2003. *Le Grand Robert de la langue française. Version CD-ROM.* Emme Interactive.

Rey, A. und J. Rey-Debove (Hg.). 1993. *Le Nouveau Petit Robert. Dictionnaire de la langue française.* Paris: Dictionnaires Le Robert.

Richardson, R., A. F. Smeaton und J. Murphy. 1994. Using WordNet as a Knowledge Base for Measuring Semantic Similarity between Words. Technischer Bericht CA-1294. Dublin, Ireland: Trinity College. URL `http://citeseer.ist.psu.edu/richardson94using.html.12.01.2005`.

Rubenstein, H. und J. B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM* 8(10): 627–633.

Sussna, M. 1993. Word sense disambiguation for free-text indexing using a massive semantic network. In: B. K. Bhargava, T. W. Finin und Y. Yesha (Hg.). *Proceedings of the Second International Conference on Information and Knowledge Management, Washington, DC, USA, November 1-5, 1993*. ACM Press. 67–74.

Ullmann, S. 1967, Reprint der 2. Ausgabe von 1957. *The Principles of Semantics*. Oxford: Basil Blackwell.

Voorhees, E. M. 1993. Using WordNet to disambiguate word senses for text retrieval. In: *Sixteenth Annual International Conference on Research and Development in Information Retrieval*. Pittsburgh, Pennsylvania: Association for Computing Machinery. 171–180.

Vossen, P. 1997. EuroWordNet: a multilingual database for information retrieval. In: *Proceedings of the DELOS workshop on Cross-language Information Retrieval*. Zürich. URL `http://citeseer.ist.psu.edu/vossen97eurowordnet.html.12.01.2005`.

Vossen, P., L. Bloksma, P. Boersma, F. Verdejo, J. Gonzalo, H. Rodriguez, G. Rigau, N. Calzolari, C. Peters, E. Picchi, S. Montenagni und W. Peters. 1998a. EuroWordNet Tools and Resources Report, Version 1, Final. Technischer Bericht D021D025. University of Amsterdam: EuroWordNet (LE-4003). URL `http://www.illc.uva.nl/EuroWordNet/docs/D021D025PS.zip.12.01.2005`.

Vossen, P., L. Bloksma, S. Climent, M. A. Marti, G. Oreggioni, G. Escudero, G. Rigau, H. Rodriguez, A. Roventini, F. Bertagna, A. Alonge, C. Peters und W. Peters. 1998b. The Restructured Core Wordnets in EuroWordNet: Subset 1. Version 3, Final. Technischer Bericht D014, D015, WP3, WP4. University

of Amsterdam: EuroWordNet (LE2-4003). URL http://www.illc.uva.nl/EuroWordNet/docs/D014D015PS.zip.12.01.2005.

Vossen, P., L. Bloksma, H. Rodriguez, S. Climent, N. Calzolari, A. Roventini, F. Bertagna, A. Alonge und W. Peters. 1997. The EuroWordNet Base Concepts and Top Ontology. Technischer Bericht D017, D034, D036. University of Amsterdam: EuroWordNet (LE2-4003). URL http://www.illc.uva.nl/EuroWordNet/docs/D017PS.zip.12.01.2005.

Vossen, P. und G. Escudero. 1999. Comparison of the Final Wordnets German, French, Czech and Estonian Version 2, Final. Technischer Bericht 2D011D01, WP3, WP4. University of Amsterdam: EuroWordNet (LE4-8328). URL http://www.illc.uva.nl/EuroWordNet/docs/2D011D012PS.zip.12.01.2005.

Véronis, J. und N. Ide. 1998. Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. *Computational Linguistics. Special Issue on Word Sense Disambiguation* 24(1): 1–40.

Wagner, A., D. Dutoit und L. Catherin. 1999. Tools and Resources for the French and German Wordnets EuroWordNet. Technischer Bericht 2D005. University of Amsterdam: EuroWordNet (LE4-8283). URL http://www.illc.uva.nl/EuroWordNet/docs/2D005.ai.12.01.2005.

Wagner, A. und C. Kunze. 1999. Integrating GermaNet into EuroWordNet, a Multilingual Lexical-Semantic Database. *Sprache und Datenverarbeitung* 23(2): 5–20.

Wiebe, J., J. Maples, L. Duan und R. Bruce. 1997. Experience in WordNet Sense Tagging in the Wall Street Journal. In: *Proceedings of the ANLP 1997 Workshop, Tagging Text with Lexical Semantics: Why, What, How?* Washington D.C.: Association for Computational Linguistics SIGLEX. 8–11. URL http://citeseer.ist.psu.edu/551394.html.12.01.2005.

Wiebe, J., T. O'Hara und R. Bruce. 1998. Constructing Bayesian Networks from WordNet for Word-Sense Disambiguation: Representational and Processing Issues. In: S. Harabagiu (Hg.). *Use of WordNet in Natural Language Processing Systems: Proceedings of the Conference.* Somerset, New Jersey:

Association for Computational Linguistics. 23–30. URL http://citeseer.ist.psu.edu/wiebe98constructing.html.12.01.2005.

Wu, Z. und M. Palmer. 1994. Verb semantics and lexical selection. In: *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*. New Mexico State University, Las Cruces, New Mexico. 133 – 138. URL http://citeseer.ist.psu.edu/579319.html.12.01.2005.

Yarowsky, D. 1994a. A comparison of corpus-based techniques for restoring accents in Spanish and French text. In: *2nd Annual Workshop on Very Large Corpora*. Kyoto. 19–32. URL http://citeseer.ist.psu.edu/yarowsky94comparison.html.12.01.2005.

Yarowsky, D. 1994b. Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. In: *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*. New Mexico State University, Las Cruces, New Mexico. 88–95. URL http://citeseer.ist.psu.edu/yarowsky94decision.html.12.01.2005.

Daten der mündlichen Prüfung:
Hauptfach: Romanistik (Schwerpunkt Französisch), 14.07.2005
1. Nebenfach: Spanisch, 18.10.2005
2. Nebenfach: Englisch, 23.03.2005
Dekan: Prof. Dr. Dr. h. c. Wichard Woyke
Erstkorrektor: Prof. Dr. Wolf Dietrich
Zweitkorrektor: Prof. Dr. Wolf Paprotté