Computerlexikographie-Tutorium 16.05.2008

- Themen für heute:
  - Teil I: Parsing von Wörterbuchartikeln
    - Parsing von Wörterbuchartikeln: Ziele und Anforderungen
    - Forschung von Vossen et al. 1989 mit LDOCE
  - Teil II: Das LexParse-System

Teil I: Parsing von Wörterbuchartikeln

Parsing von Wörterbuchartikeln

- Ziele:
  - funktionale Textsegmente durch **Segmentierung** der Wörterbuchartikel erkennen
  - digitale **Repräsentation** des Wörterbuchtextes erzeugen
- Anhand der Strukturanzeiger wird die initiale Segmentierung des Wörterbuchartikels in funktionale Textsegmente durchgeführt.
  - Problematische Fälle: indefinite Strukturanzeiger, die im WA nicht nur als Strukturanzeiger auftreten (↔ definite Strukturanzeiger).
- Das Parserprogramm
  - definiert **Wohlgeformtheitsbedingungen** für Wörterbucheinträge und Komponenten,
  - überprüft **Konsistenz** der Wörterbuchkodierung.

Parsing von Wörterbuchartikeln

- Anforderungen an einem Wörterbuchparser:

  Die **Parsergrammatik** muss

  - möglichst einfach und leicht adaptierbar sein.
  - mit typographischen Markern und Durchnummerierungen zurecht kommen (nicht kontextfrei!);
  - Iterationen behandeln können;
  - den Aufbau eigener Wörterbuchgrammatiken für die Nutzer ermöglichen;
  - das Format der Parsebäume konfigurierbar halten;
  - spezifische Funktionen vorsehen;
  - erweiterbar sein;
  - speicherbare Einstellungen enthalten;
  - möglichst plattformneutral implementiert sein.

LDOCE

- Longman Dictionary of Contemporary English (ab 1978)
- maschinenlesbares Wörterbuch
- LDOCE benutzt ein eingeschränktes Beschreibungsvokabular von ca. 2000 Vokabeln – sogenanntes **kontrolliertes Vokabular** (Controlled Vocabulary = CV).
  - Simon Dik (1978): kompositioneller Ansatz: schrittweise lexikalische Bedeutungen auf ein Basisvokabular reduzieren.
- Piek Vossen et al. (1989): LINKS-Projekt: Ziel: eine semantische Datenbank der systematischen Bedeutungsbeschreibungen zusammenzustellen.
  - Verarbeitungsschritte: (siehe E-Buch S. 108 f. (Kap.5/2.4))

    → syntaktische und semantische Annotation der Wörterbuchartikel
  - Das Parsing der Wörterbuchartikel des LDOCE dient der **semantischen** Rekonstruktion von Bedeutungsbeschreibungen.

LDOCE: Annotation

- Jeder Wörterbucheintrag wird mit 5 Angaben spezifiziert:
  - **Subject field code**: gibt das Sachbereichsfeld des Wörterbucheintrages an
  - **Box code**: stilistische, soziolinguistische, semantische Informationen über die Bedeutung des Wörterbucheintrages
  - **orthographic form**: orthographische Form des Wörterbucheintrages
  - **POS-code**: Wortart der Wörter in der Bedeutungsbeschreibung als syntaktische Kategorie
  - **meaning description (MD)**: die Bedeutungsbeschreibung innerhalb des Wörterbuchartikels

- Mithilfe dieser Struktur kann man auf verschiedene Elemente des Wörterbucheintrages systematisch zugreifen.

LDOCE: Annotation

- Beispiel: Suche nach „all meaning descriptions of nouns with the semantic label 'human', the subject field 'medicine' and 'occupation', and with the initial words *a doctor who*" – Ausgabe: (Vossen 1989: 173)

subject fields	semantic label	entry(POS)	marker of beginning of MD		meaning description
⇓ ⇓	⇓	⇓ ⇓	⇓	⇓	⇓
mdon–	....h...y.	anaesthetist(*n*)	ul	UL	a doctor who gives an anaesthetic to a patient before he is treated by another doctor.
mdon–	....h.....	specialist(*n*)	ul	UL	a doctor who gives treatment in a particular way or to certain kinds of people or diseases.

LDOCE: Parsing von NMD

- Es wurden in (Vossen et al. 1989) nur die nominalen Einträge betrachtet. Diese enthalten sogenannte **nominale Bedeutungsbeschreibungen** (nominal meaning description = **NMD**).
  - Annahme: Die meisten nominalen Bedeutungsbeschreibungen sind Nominalphrasen.

- 4 Ebenen der nominalen Bedeutungsbeschreibungen:
  a. **Wortsequenz** (word sequence)
  b. **POS-Sequenz** (POS sequence)
  c. **syntaktisches Muster** (syntactic pattern)
  d. **semantisches Muster** (semantic pattern)

LDOCE: Parsing von NMD

- Beispiel für eine nominale Bedeutungsbeschreibung: *a man who works*
  (Vossen 1989: 176)

a. Word sequence	*a*	*man*	*who*	*works*
b. POS sequence	Det	Noun	RelPronoun	Verb
c. Syntactic pattern	[NP [Det a]	[N (or KERNEL) man]	[REL_CLAUSE who works]]	
d. Semantic pattern	Quantor	category	activity-specifying modifier	

LDOCE: Parsing von NMD

- Strukturelle Eigenschaften der nominalen Bedeutungsbeschreibungen: Die Grundstruktur der Nominalphrasen:
  - optionaler **Determinierer**: drückt Aspekte der Definiertheit, Numerus, Zählbarkeit etc. aus.
  - optionale(r) Modifikator(en): schränken die vom syntaktischen Kern (Nomen) bezeichneten Entitäten semantisch ein.
    - **Prämodifikator** des Kernes: geht dem Kern voran
    - **Postmodifikator** des Kernes: steht nach dem Kern
  - obligatorischer syntaktischer **Kern** (kernel): Die Struktur der Bedeutungsbeschreibung wird grundsätzlich vom syntaktischen Kern der NP gesteuert.
- Jeder Teil kann einfach oder komplex (z.B. Koordination oder Adjunkte) vorkommen.
  - Determinierer Prämodifikator(en) Kern Postmodifikator(en)

LDOCE: Parsing von NMD

- Beispiel für eine nominale Bedeutungsbeschreibung: *a man who works*
  (Vossen 1989: 176)

a. Word sequence	*a*	*man*	*who* *works*
b. POS sequence	Det	Noun	RelPronoun Verb
c. Syntactic pattern	[NP [Det a]	[N (or KERNEL) man]	[REL_CLAUSE who works]]
d. Semantic pattern	Quantor	category	activity-specifying modifier

Determinierer Kern Postmodifikator

LDOCE: Kerntypen

Kerntypen nach der Verteilung der relevanten semantischen Informationen in der Bedeutungsbeschreibung:

- **Link**
- **Synonym**
- **Linker**
- **Shunter**

LDOCE: Kerntypen

Kerntypen nach der Verteilung der relevanten semantischen Informationen in der Bedeutungsbeschreibung:

- **Link**: Der syntaktische Kern der NP ist ein Hyperonym des Wörterbucheintrages. Die Bedeutung des Kerns ist durch Prä- und/oder Postmodifikator(en) eingeschränkt.
- **Synonym**: Alle semantischen Informationen werden auf ein Wort konzentriert, das ein Synonym des Wörterbucheintrages ist.
  - (~ impliziter Verweis, d.h. dieses Synonym muss als Eintrag im Wörterbuch erklärt werden.)

LDOCE: Kerntypen

- **Linker**: Der Kern ist relativ bedeutungsleer. Die syntaktische Struktur ist oft ein *of*-Komplement mit einem Nomen. Dieses Nomen trägt die meiste semantische Information über den Wörterbucheintrag. Die semantische Relation zwischen dem Wörterbucheintrag und diesem Nomen ist keine Hyponymie-Hyperonymie-Relation (sondern z.B. Meronymie).
  - Nicht alle Kerne mit *of*-Komplement sind Linker! – Wenn der semantische Inhalt des Kernes und des Komplements ausgeglichen(er) ist, d.h. der Kern ungefähr so viel semantische Information wie das Komplement trägt, ist der Kern ein Link.

LDOCE: Kerntypen

- **Shunter**: haben oft ein *of*-Komplement oder ein Relativsatz als Postmodifikator. Der Kern ist – wie bei Linkern – relativ uninformativ, (zu) allgemein. Unterschiede zu Linkern:
  - Der Kern hat keine *direkte* semantische Relation zum Wörterbucheintrag.
  - Das *of*-Komplement in der Bedeutungsbeschreibung
    - ist eine Verbalphrase (Verb typischerweise in Verlaufsform). Der Kern dieser VP ist i.d.R. das Verb oder ein Adjektiv. Das heißt, es wird ein Wortartenwechsel in die Bedeutungsbeschreibung eingeführt.
    - kann mit einem Relativsatz paraphrasiert werden.
  - Der Postmodifikator kann auch ein Relativsatz sein.
  - Der Wörterbucheintrag ist i.d.R. ein Derivat (Nominalisierung).
- Aber: Wenn der Kern und der Modifikator ungefähr genauso viel zur Bedeutungsbeschreibung beitragen, spricht man vom Link. Wenn der Kern selbst ein nominalisiertes Element ist, dann trägt der Kern auch mehr Bedeutung, deswegen spricht man hier auch von Links.

Was sind die Kerne in den folgenden Bedeutungsbeschreibungen?

1. **abattoir:** slaughterhouse
2. **accuracy**: the quality of being accurate
3. **actuality**: the state of being real
4. **agony column**: personal column
5. **angling**: a sport of catching fish with a hook and line
6. **aspiration**: the pronunciation of the letter *h*
7. **attraction**: something which attracts
8. **beef**: the meat of farm cattle
9. **blackout**: a period of darkness caused by a failure of the electric power supply
10. **captain**: the leader of a team or group
11. **flamingo**: a tall tropical water bird with long thin legs, pink an red feater, and a broad beak curved downwards
12. **high seas**: the oceans of the world which do not belong to any particular country
13. **stomach**: the front part of the body below the chest

(Weitere Beispiele: E-Buch S. 109 ff. (Kap.5/2.4/Typen nominaler Bed.beschr.))

Was sind die Kerne in den folgenden Bedeutungsbeschreibungen?

1. **abattoir:** slaughterhouse
2. **accuracy**: the quality of being accurate
3. **actuality**: the state of being real
4. **agony column**: personal column
5. **angling**: a sport of catching fish with a hook and line
6. **aspiration**: the pronunciation of the letter *h*
7. **attraction**: something which attracts
8. **beef**: the meat of farm cattle
9. **blackout**: a period of darkness caused by a failure of the electric power supply
10. **captain**: the leader of a team or group
11. **flamingo**: a tall tropical water bird with long thin legs, pink an red feater, and a broad beak curved downwards
12. **high seas**: the oceans of the world which do not belong to any particular country
13. **stomach**: the front part of the body below the chest

(Weitere Beispiele: E-Buch S. 109 ff. (Kap.5/2.4/Typen nominaler Bed.beschr.))

Was sind die Kerntypen? Argumentieren Sie!

1. **abattoir:** slaughterhouse
2. **accuracy**: the quality of being accurate
3. **actuality**: the state of being real
4. **agony column**: personal column
5. **angling**: a sport of catching fish with a hook and line
6. **aspiration**: the pronunciation of the letter *h*
7. **attraction**: something which attracts
8. **beef**: the meat of farm cattle
9. **blackout**: a period of darkness caused by a failure of the electric power supply
10. **captain**: the leader of a team or group
11. **flamingo**: a tall tropical water bird with long thin legs, pink an red feater, and a broad beak curved downwards
12. **high seas**: the oceans of the world which do not belong to any particular country
13. **stomach**: the front part of the body below the chest

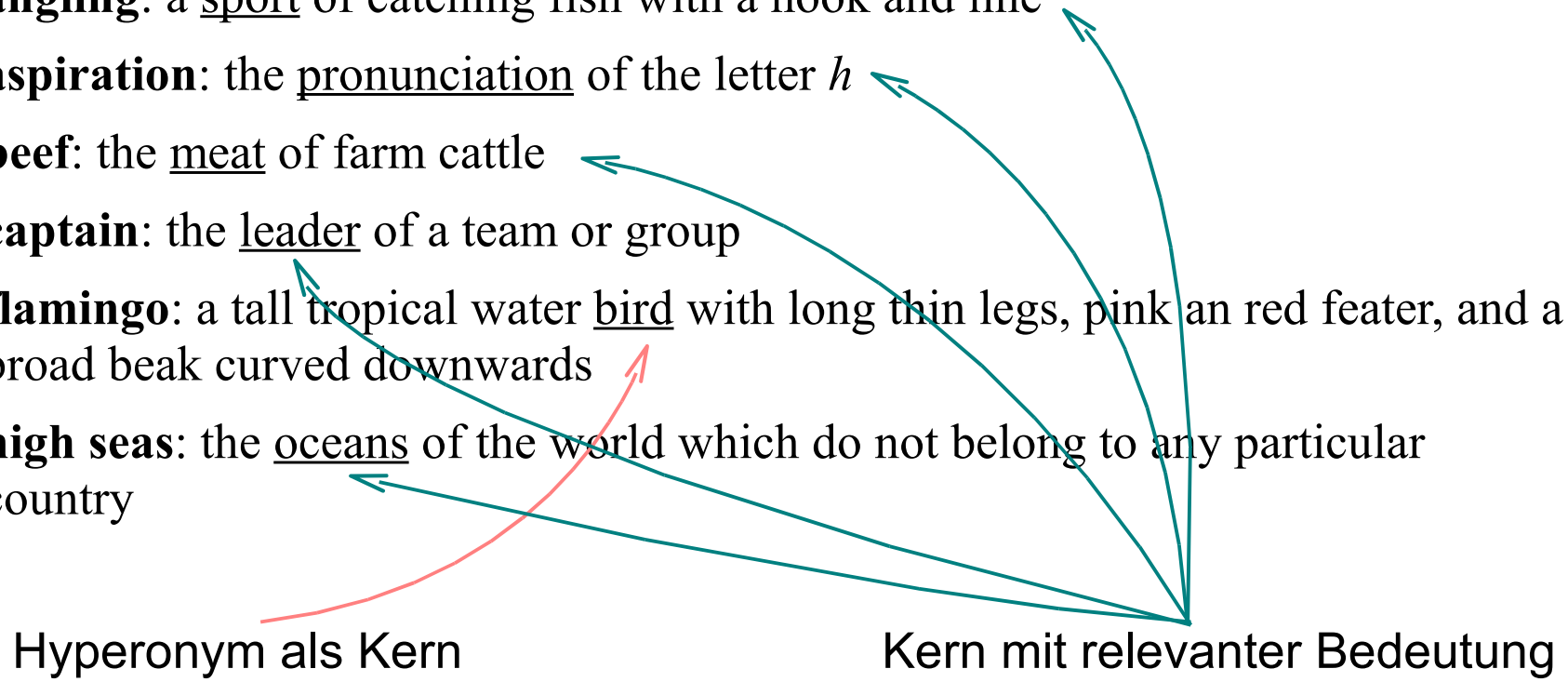
(Weitere Beispiele: E-Buch S. 109 ff. (Kap.5/2.4/Typen nominaler Bed.beschr.))

Was sind die Kerntypen? Argumentieren Sie!

- **Link:**
  - **angling**: a sport of catching fish with a hook and line
  - **aspiration**: the pronunciation of the letter *h*
  - **beef**: the meat of farm cattle
  - **captain**: the leader of a team or group
  - **flamingo**: a tall tropical water bird with long thin legs, pink an red feater, and a broad beak curved downwards
  - **high seas**: the oceans of the world which do not belong to any particular country

Was sind die Kerntypen? Argumentieren Sie!

- **Link:**
  - **angling**: a sport of catching fish with a hook and line
  - **aspiration**: the pronunciation of the letter *h*
  - **beef**: the meat of farm cattle
  - **captain**: the leader of a team or group
  - **flamingo**: a tall tropical water bird with long thin legs, pink an red feater, and a broad beak curved downwards
  - **high seas**: the oceans of the world which do not belong to any particular country

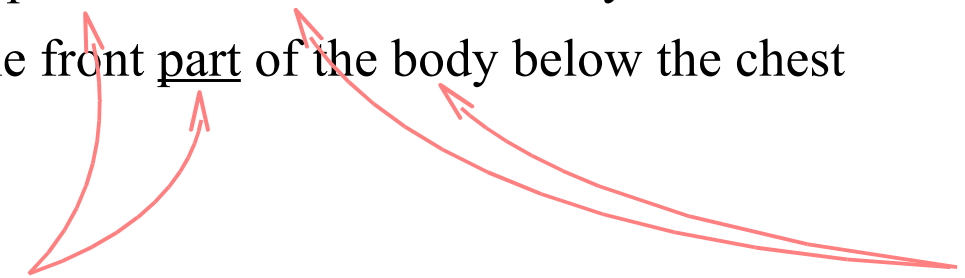


Was sind die Kerntypen? Argumentieren Sie!

- **Synonym:**
  - **abattoir:** slaughterhouse
  - **agony column**: personal column
- **Linker:**
  - **blackout**: a period of darkness caused by a failure of the electric power supply
  - **stomach**: the front part of the body below the chest

Was sind die Kerntypen? Argumentieren Sie!

- **Synonym:**
  - **abattoir:** slaughterhouse
  - **agony column**: personal column
- **Linker:**
  - **blackout**: a period of darkness caused by a failure of the electric power supply
  - **stomach**: the front part of the body below the chest



relativ bedeutungsleerer Kern + nominales *of*-Komplement

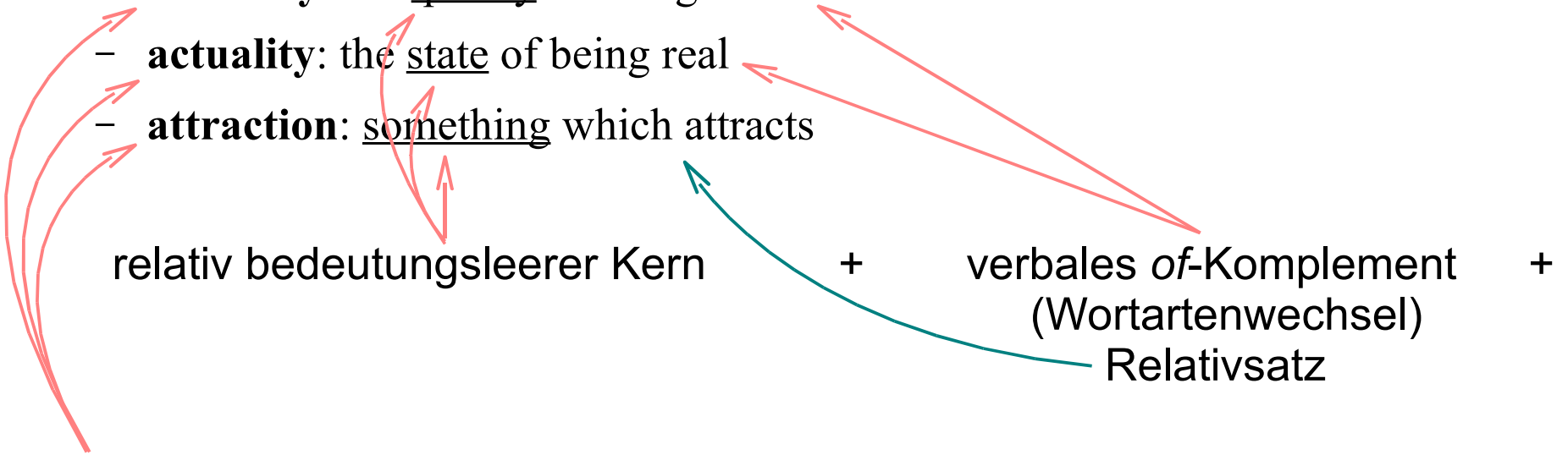
! Die Bedeutungsbeschreibung von **stomach** kann auch als ein Link betrachtet werden, weil „part of the body" ~ „body part" eine feste Wendung ist und deswegen kein bedeutungsleerer Kern in der NMD.

Was sind die Kerntypen? Argumentieren Sie!

- **Shunter:**
  - **accuracy**: the quality of being accurate
  - **actuality**: the state of being real
  - **attraction**: something which attracts

Was sind die Kerntypen? Argumentieren Sie!

- **Shunter:**
  - **accuracy**: the quality of being accurate
  - **actuality**: the state of being real
  - **attraction**: something which attracts



LDOCE: Problematische Phänomene

- Im LDOCE waren 55 % der Kerne in nominalen Bedeutungsbeschreibungen vom Typ Linker oder Shunter.

  ↔ Die allgemeine Erwartung war, dass man viel Links mit Hyperonym in der Bedeutungsbeschreibung findet.

- Viele Strukturen sind ambig.
  - z.B. *of*-Komponenten
- Idiome in der Bedeutungsbeschreibung erschweren die Erkennung.
- Typographische Markierungen wie Klammerung und Schrägstriche sind schwierig handhabbar.
- Zirkuläre Verweise sollten umgearbeitet werden. Beispiel:

  object: a thing

  thing: any material object

Teil II: Das LexParse-System

LexParse-System

- Tool und Beschreibung: http://www.sfs.uni-tuebingen.de/de_nf_asc_resources.shtml ; http://www.cl.uni-heidelberg.de/kurs/apparat/colex5.pdf ; http://www.sfs.uni-tuebingen.de/~lothar/LEXPARSE/lexparse_info.pdf

- „LexParse ist ein benutzerfreundliches, schnelles und flexibles Programm zum Parsen maschinenlesbarer Lexika" von Ralf Hauser (im Rahmen des ELWIS-Projektes, Uni Tübingen).
- LexParse „erkennt die einzelnen Lexikoneinträge und ihren internen Aufbau. Ebenso kann es die Ergebnisse in verschiedenen Formaten (als Baumstruktur, in SGML, in LaTeX) darstellen."
- Wörterbuchgrammatiken definieren Wohlgeformtheitsbedingungen für den Zugriff auf die Ordnungsstrukturen.
- Basiseinheit: Wörterbucheintrag
- Untereinträge werden geclustert.

LexParse-System

- Das System wendet die Theorie Wiegands für die Wörterbuchgrammatik an. (Siehe letztes Tutorium.)
- Strukturelle Relationen zwischen den Teilen eines Wörterbuchartikels:
  - **partitive Relationen**: für Angabeklassen; ~ partitive Mikrostruktur, hierarchische Aufteilung
  - **Präzedenzrelationen**: Abfolge der elementaren (terminalen) Angaben innerhalb einer Mikrostruktur
- Kontextfreie **W**örterbuch**a**rtikel**g**rammatik ist definiert als:
  WAG = <CEI, CNI, R, WA>
  - CEI: terminales Alphabet von WAG: Menge der elementaren Angabeklassen
  - CNI: Menge der nichtterminalen Symbole von WAG
  - R: Menge der kontextfreien Ersetzungsregeln
  - WA: Wörterbuchartikel; das initiale Symbol aus CNI

Architektur von LexParse

- Modularer Aufbau. (Siehe E-Buch: S. 102 f. (Kap.5/2.2))
- Konfigurationsdatei + auszuführendes Programm
- Vorangehensweise:
  - der Wörterbucheintrag wird in Segmente zerteilt
  - Parsing: die hierarchische Struktur des Wörterbuchartikels wird festgestellt.
  - Alle implizite Informationen werden explizit gemacht, Abkürzungen werden aufgelöst.
  - Überprüfung der Wohlgeformtheit.
  - Daten werden in einer wohlgeformten Form repräsentiert.
- Beispielgrammatik + geparster Eintrag: E-Buch: S. 104–107 (Kap.5/2.3)

Quellen

- Engelke, Sabine, Ralf Hauser & Angelika Storrer (o.J.): *LexParse V1.00. A Dictionary Entry Parser*. Link: http://www.sfs.uni-tuebingen.de/~lothar/LEXPARSE/lexparse_info.pdf (Stand: 15.05.2008, 15:40)
- Lemnitzer, Lothar, Claudia Kunze (2005): *Dictionary Entry Parsing*. Link: www.cl.uni-heidelberg.de/courses/ss08/lexicography/slidesDictPars.pdf (Stand: 28.04.2008, 22:02)
- Lemnitzer, Lothar, Claudia Kunze (2006): *Parsen von Wörterbuchartikeln.* Link: http://www.cl.uni-heidelberg.de/kurs/apparat/colex5.pdf (Stand: 15.05.2008, 15:41)
- Tools von Universität Tübingen, Seminar für Sprachwissenschaft. Link: http://www.sfs.uni-tuebingen.de/de_nf_asc_resources.shtml (Stand: 15.05.2008, 16:42)
- Vossen, Piek, Willem Meijs & Marianne den Broeder (1989): Meaning and structure in dictionary definitions. In: Boguraev, Bran & Ted Briscoe (Hrsg.): *Computational Lexicography for Natural Language Processing.* London / New York : Longman. S. 171–192.