MARINA KULICHIKHINA/NATALIA RUBAN

Semantisches Wörterbuch der deutschen Sprache für maschinelle Sprachverarbeitungssysteme

Im folgenden Beitrag handelt es sich um die Entwicklung eines semantischen Wörterbuches der deutschen Sprache für maschinelle Sprachverarbeitungssysteme im Rahmen des Projektes „Compreno" bei dem russischen IT-Unternehmen ABBYY. Es wird eine kurze Übersicht über andere elektronische Quellen zur deutschen Sprache gegeben, ferner werden ihre Unterschiede im Vergleich zum Projektwörterbuch analysiert. An einigen Beispielen werden aktuelle Probleme der Computerlexikografie (Bedeutungsunterscheidung, Komposita-Analyse u. a.) und ihre mögliche Lösung in Bezug auf das Projektwörterbuch betrachtet.

Die Erstellung von lexikografischen Quellen, die für sprachverarbeitende Systeme bestimmt sind, gehört zu den aktuellsten Aufgaben der modernen Lexikografie. So stellen C. Kunze und L. Lemnitzer in ihrer Monografie Computerlexikographie. Eine Einführung fest:

> Schließlich ist der Computer selbst zum „Konsumenten" lexikographischer Daten geworden, genauer: sprachtechnologische Software, die umfassende linguistische und lexikalische Informationen benötigt. Diese Informationen sind für viele sprachtechnologische Anwendungen essenziell, und es gibt einen wachsenden Markt für lexikalische Daten, die für diese neue „Zielgruppe" geeignet sind. Die Herausforderung liegt darin, die lexikographischen Daten in einer so strikt formalen Weise zu präsentieren, dass sprachtechnologische Anwendungen sie nutzen können. (KUNZE/LEMNITZER 2007: 5)

Lexikografische Quellen, die für maschinelle Sprachverarbeitungssysteme geeignet sind, werden in mehreren Ländern und in verschiedenen Sprachen entwickelt[1]: Das sind Online-Thesauri, semantische Netze, Ontologien, Framenetze und Quellen, die in sich Eigenschaften eines Wörterbuches, eines semantischen Netzes und einer Ontologie vereinigen. In diesem Beitrag möchten wir eine lexikografische Quelle für die deutsche Sprache vorstellen, die vom russischen Unternehmen ABBYY entwickelt wird.

1 WordNet, Lexipedia, SNOMED CT, ConceptNet, FrameNet, GermaNet et al.

1 Die Entwicklung des semantischen Wörterbuches der deutschen Sprache

Im Jahre 2009 wurde die lexikografische Beschreibung der deutschen Sprache im Rahmen des multilingualen Sprachverarbeitung-Projekts „Compreno“ gestartet. Dieses Projekt orientiert sich an Aufgaben der maschinellen Sprachverarbeitung wie maschinelle Übersetzung, semantische Suche, Dokumentenklassifizierung u. a. Zurzeit ist im Rahmen des Projektes die grundlegende Beschreibung der Wortschätze der englischen und russischen Sprachen beendet, Wortschätze der deutschen, französischen und chinesischen Sprache sind in Bearbeitung. Die mehrsprachige Struktur des Projektes ermöglicht die Erweiterung des Projektes durch andere Sprachen.

2 Wörterbuch- und NLP-Quellen

Bei der Arbeit am Projektwörterbuch der deutschen Sprache berücksichtigen die Lexikografen von ABBYY verschiedene Typen von lexikografischen und linguistischen Quellen. Die traditionelle deutsche Lexikografie bietet mehrere einsprachige Wörterbücher an (z. B. *Wörterbuch Deutsch als Fremdsprache* von de Gruyter, vgl. KEMPCKE 2000), Wörterbücher von Duden, Wahrig, Langenscheidt und Pons); es werden auch elektronische Quellen wie DWDS[2], Deutsches Wortschatz-Portal[3], Limas[4], OWID[5], das elektronische Valenzwörterbuch deutscher Verben[6] u. a. in Betracht gezogen.

Es gibt auch solche lexikografische Quellen der deutschen Sprache, die vorwiegend auf maschinelle Sprachverarbeitung orientiert sind. Eine wichtige Quelle ist das deutsche semantische Netz GermaNet[7], das an der Universität Tübingen entwickelt wird und auf Prinzipien des englischsprachigen semantischen Netzes WordNet[8] basiert.

GermaNet hat vom englischen Netz sein Strukturprinzip übernommen: Die Bedeutungen sind zu Synsets vereinigt und durch semantische Beziehungen verbunden. Im Mai 2012 beinhaltete das Netz ca. 75.000 Synsets mit ca. 100.000 lexikalischen Einheiten. GermaNet wird wie folgt definiert: „A resource for

2 http://www.dwds.de/

3 www.wortschatz.uni-leipzig.de

4 www.korpora.org/Limas/

5 www.owid.de

6 http://hypermedia2.ids-mannheim.de/evalbu/index.html

7 http://www.sfs.uni-tuebingen.de/lsd/

8 http://wordnet.princeton.edu

word sense disambiguation which is crucial for natural language applications like information retrieval, the construction of various language technology tools and the annotation of corpora“[9].

In einem Synset sind Wörter gleicher Wortart aufgeführt. In „Compreno“ ist es dagegen möglich, Wörter verschiedener Wortarten in eine semantischen Klasse zu platzieren. Die Besonderheit von GermaNet ist die Verwendung von künstlichen Konzepten, z. B. **hierarchischer_Lehrer* (s. KUNZE 2007: 138–139), die die Lücken in der Hierarchie ergänzen und die Verhältnisse zwischen den Synsets besser konstruieren.

Eine weitere Quelle ist HaGenLex (HAgen GErmaN LEXicon), ein Computerlexikon der deutschen Sprache von der Fernuniversität Hagen. Diese Quelle hat 26.000 Einträge von verschiedenen Wortarten. Eine Besonderheit dieses Lexikons ist die Einbeziehung syntaktischer und semantischer Valenzrahmen in die Beschreibung. HaGenLex wird im Parser WOCADI[10] benutzt.

An der Universität des Saarlandes wird ein elektronisches lexikalisch-semantisches Lexikon SALSA II[11] entwickelt; das ist eine Quelle für die linguistische und maschinelle Analyse, die auf Prinzipien der Frame-Semantik basiert (vgl. FILLMORE 1982). Die Frame-Quellen beschreiben lexikalische Einheiten (meistens Verben) im Zusammenhang mit Wörtern, die mit ihnen semantisch verbunden sind; dieser Ansatz ist den Prinzipien der lexikografischen Beschreibung, die wir in „Compreno“ verwenden, sehr nah. Eine weitere Frame-Quelle ist German FrameNet[12], das an der University of Texas (Austin) auf Basis der englischsprachigen Quelle FrameNet und mit der Benutzung des deutschen Lexikons SALSA II entwickelt wird. Ein Vorteil von Frame-Quellen besteht in der aktiven Verwendung von Korpora für die Beschreibung der Wortkompatibilität.

Zu den mehrsprachigen lexikalischen Basen gehören das semantische Netz EuroWordNet[13] und das semantische Wörterbuch Lexipedia[14]. Die Besonderheit von EuroWordNet ist ein sprachunabhängiger Inter-Lingual-Index – ein

9 Eine Quelle für die Disambiguierung von Textwörtern, die für sprachverarbeitende Anwendungen wie die Informationsbeschaffung, die Entwicklung von verschiedenen Tools der Sprachtechnologie und die Annotation von Korpora äußerst wichtig ist, bildet http://www.sfs.uni-tuebingen.de/lsd/germanet_structure.shtml.

10 http://pi7.fernuni-hagen.de/research/wocadi/wocadi_demo_de.html

11 http://www.coli.uni-saarland.de/projects/salsa/page.php?id=overview

12 http://www.laits.utexas.edu/gframenet/

13 http://www.illc.uva.nl/EuroWordNet/

14 http://www.lexipedia.com/

Katalog von universellen Begriffen, die dem System von semantischen Klassen in der universellen semantischen Hierarchie (USH, s. unten) ähnlich sind. Lexipedia ist ein visuelles semantisches Online-Wörterbuch, das Englisch, Spanisch, Deutsch, Französisch und Niederländisch unterstützt.

Die elektronischen Ontologien streben nach der Entwicklung einer hierarchischen Taxonomie von Begriffen, die die Realien der gegenständlichen Welt beschreibt. Die erfolgreichsten Ontologien (SUMO, OpenCyc, DOLCE) sind in der englischen Sprache vorhanden. Die Verfasser von GermaNet weisen darauf hin, dass dieses semantische Netz als eine Ontologie betrachtet werden kann. Das im Rahmen von „Compreno" entwickelte deutsche Wörterbuch kann jedenfalls als eine Ontologie benutzt werden (besonders im Teil der Konkreta, z. B. bei einigen Tochterknoten des Teilbaums ENTITY_LIKE_CLASSES – wie PHYSICAL_OBJECT, FOOD u. a.), weil die Taxonomie dieser Begriffe der Hierarchie von Konkreta sehr nahe steht.

Bei lexikografischen Beschreibungen in unserem Projekt haben wir Erfahrungen folgender lexikografischer Quellen berücksichtigt: Unser Projektwörterbuch der deutschen Sprache verbindet die Eigenschaften von Thesaurus, semantischem Netz, FrameNetz und Ontologien; jedoch verlangen die Anforderungen von „Compreno" besondere Beschreibungsstandarde: Berücksichtigung des Erbens der Eigenschaften in der hierarchischen Struktur, möglichst ausführlicher lexikografischer Kommentar, der vollständige relevante Informationen über das Wort enthält, was das Funktionieren syntaktischer Beschreibungen ermöglicht, und andere Unterschiede, auf die wir später eingehen werden.

3 Universelle semantische Hierarchie

Die Grundlage des Systems „Compreno" bildet die universelle semantische Hierarchie (weiter als USH bezeichnet) – der hierarchische Thesaurus-Baum, dessen Knoten semantische Klassen sind. Die semantische Klasse (weiter SK) ist eine Einheit, die einen bestimmten Universalbegriff beinhaltet. Zwei Hauptteilbäume der USH werden jeweils mit prädikativer Lexik (situationsbedingter, attributiver Lexik) und der Entität-Lexik (beinhaltet sowohl konkrete als auch abstrakte Gegenstände) gebildet. Je weiter nach unten in der Hierarchie die jeweilige semantische Klasse gesetzt wird, einen desto konkreteren und spezifischeren Begriff repräsentiert sie. Jede semantische Klasse kann mit einer konkreten sprachlichen Bedeutung – einer lexikalischen Klasse – ergänzt werden. Die lexikalische Klasse (weiter als LK bezeichnet) enthält konkrete Lexeme: Zum Beispiel die LK *Klassik* aus der semantischen Klasse *CLASSICAL_PERIOD* erhält zwei Lexeme – *Klassik* und *klassisch*.

Bei der lexikografischen Beschreibung werden deutsche lexikalische Klassen in die semantische Hierarchie eingeführt und mit einem detaillierten Kommentar versehen.

Zum September 2013 sind über 13.300 deutsche lexikalische Klassen (mit über 30.000 Lexemen) beschrieben, weiterhin streben wir eine möglichst vollständige Beschreibung von deutscher Lexik an.

Die USH ist ein Bindeglied zwischen verschiedenen Sprachen. Indem man von einer Sprache zu einer anderen geht, kann man sehen, wie der jeweilige universelle Begriff in einer konkreten natürlichen Sprache kodiert wird.

4 Der Vergleich des Projektwörterbuchs mit einigen anderen lexikografischen Quellen der deutschen Sprache

Lexikografische Beschreibungen für „Compreno", wie auch für andere maschinenorientierte Quellen, unterscheiden sich in vieler Hinsicht von den Wörterbüchern, die für Menschen bestimmt sind. Die Verfasser von klassischen Wörterbüchern gehen davon aus, dass der Mensch beim Lesen eines Wörterbuchartikels die Bedeutung des jeweiligen Wortes versteht. Daraus folgend kann er es für Bildung von freien sprachlichen Konstruktionen benutzen. Damit der Computer die Bedeutung des Wortes in unserem System „verstehen" kann, soll das Wort in die Sprache der universellen Begriffe übersetzt und mit verschiedenen Instrumenten der Bedeutungserkennung versehen werden.

Zu einer der wichtigsten Aufgaben von „Compreno" gehört eine möglichst vollständige und ausführliche Beschreibung der Kompatibilität jedes Wortes vom Standpunkt der *semantischen Rollen* (engl. semantic slots) her – universellen metasprachlichen Einheiten, die dem Begriff *Valenz* nahe stehen. Das unterscheidet unser Wörterbuch von vielen anderen Quellen (es ist zu bemerken, dass ein ähnliches Prinzip der lexikografischen Beschreibung im *Valenzwörterbuch der deutschen Verben*[15] verwendet wird. Dort werden jedoch nur Verben beschrieben, und der Umfang des Wörterbuches ist viel geringer als der von „Compreno"). Außerdem werden erweiterte grammatische Informationen, die für die jeweilige Bedeutung des Wortes relevant sind, gegeben, z. B. die Angaben zu Lokativpräpositionen für Substantive.

Somit hat unser Projektwörterbuch der deutschen Sprache folgende distinktive Eigenschaften:

- Es berücksichtigt die Errungenschaften der klassischen und modernen Lexikografie, indem jedoch verschiedene Ansätze in der Beschreibung verwendet

15 http://hypermedia2.ids-mannheim.de/evalbu/index.html

werden. Im Vergleich zum *New Oxford Thesaurus Of English* (vgl. WAITE 2000), wo in einem semantischen Feld (z. B. *Accord*) nur Synonyme beschrieben werden, können in unserer Hierarchie auch Antonyme in gleicher semantischen Klasse platziert werden, indem entsprechende Semanteme der Polarität (z. B. Einigkeit – PolarityPlus, Uneinigkeit – PolarityMinus, vgl. ANISIMOVICH et al. 2012: 100) zugeschrieben werden.

- Es ist ein Teil eines mehrsprachigen Projektes, das auf universellen Sprachmodellen basiert.
- Es hat hierarchische Struktur, an den Gipfeln sind universelle Konzepte aufgeführt. Die Tochterknoten erben entsprechende Eigenschaften von den eigenen Mutterknoten; dabei ist das Wortartenprinzip nicht systembildend.
- Die USH beinhaltet auch die semantischen Klassen, die keine konkrete sprachliche Bedeutung haben (z. B. AREA_OF_HUMAN_ACTIVITY), ähnlich den künstlichen Konzepten von GermaNet.
- Es hat Eigenschaften einer linguistischen Ontologie.
- Es beschreibt ausführlich die Kompatibilität der Wörter verschiedener Wortarten.
- Besonderer Wert wird auf ausführliche Beschreibung von Kollokationen, Komposita und Mehrwortlexemen gelegt.

5 Die Methodik der Beschreibung des deutschen Wortschatzes für das multilinguale NLP-System „Compreno“

Das Projekt „Compreno“ startete mit den lexikografischen Beschreibungen der englischen und russischen Sprache. Die Erweiterung der USH durch die deutsche Sprache wurde zur Prüfung ihrer Universalität. Für die neue Sprache sollten wir einige neue SK hinzufügen. Es wurden jedoch nur solche Klassen hinzugefügt, die keine Tochter-SK haben: z. B. *TO_GLORIFY_IN_VERSES*. Diese semantische Klasse wurde speziell für das deutsche Verb *bedichten* eingeführt, weil es weder im Russischen noch im Englischen ein entsprechendes Wort für diesen Begriff gibt.

Die Ausrichtung an NLP-Systeme bestimmte eine spezielle Methode der lexikografischen Beschreibung des deutschen Wortschatzes, die in der Projektdokumentation festgelegt wurde. Die lexikografische Beschreibung besteht aus der Einführung der deutschen lexikalischen Klassen in die semantische Hierarchie und einem detaillierten Kommentar zu diesen Klassen. Den Kern des Kommentars bilden die vom Standpunkt des „semantischen Modells“ (das erweiterte semantische Valenzmodell im Projekt) her kommentierten Beispiele, die aus zuverlässigen Quellen stammen. Im Kommentar wird die Kompatibi-

lität des Wortes illustriert. Außerdem beinhaltet der Kommentar vollständige grammatische und semantische Informationen, die für diese Bedeutung relevant sind: Besonderheiten von Genus- und Numerus-Paradigma, (In-)Kompatibilität mit finiten und infiniten Nebensätzen, Kombinierbarkeit mit den Präpositionen, Kollokationen etc.

Seit April 2012 wird an den syntaktischen Beschreibungen für die deutsche Sprache gearbeitet, was das Erkennen und Übersetzen von Phrasen und Sätzen ermöglicht. Der lexikografische Kommentar dient dabei als ausführliche und grundlegende Quelle für Informationen über Wörter, deren Bedeutungen und Valenz.

5.1 Das Verhältnis der Bedeutungen im Wörterbuch und lexikalischen Klassen in der USH

Die Bedeutungen eines Wortes können in den Hauptwörterbüchern der deutschen Sprache unterschiedlich gegliedert werden. Bei der Bedeutungsunterscheidung für „Compreno" richten wir uns nach den Forderungen der Beschreibung im Projekt. So wird manchmal eine Wörterbuchbedeutung in 2 LK in der USH gegliedert (z. B. Landschaft), wenn es unmöglich ist, in einer LK alle Kompatibilitätsvarianten zu beschreiben. In einigen Fällen ist es dagegen rational, zwei oder mehrere Wörterbuchbedeutungen in eine LK zusammenzulegen, weil sie im Rahmen desselben semantischen Modells beschrieben sein können. Manche Bedeutungsschattierungen, die im realen Wortgebrauch berücksichtigt werden können, sind für maschinelle Sprachverarbeitung irrelevant: Das Substantiv *Gefühl* hat 4–5 Bedeutungen in akademischen Wörterbüchern der deutschen Sprache, während in der USH nur 2 LK *Gefühl* in den semantischen Klassen FEELING_AS_CONDITION und PERCEPTION_ABILITY vorhanden sind.

5.2 Der Inhalt des lexikografischen Kommentars

Nachdem eine neue lexikalische Klasse in die USH eingefügt ist, beginnt der Lexikograf mit dem Verfassen eines Kommentars dazu. Dieser Kommentar hat folgende Bereiche: 1) Erläuterung der Bedeutung; 2) möglichst universelle Übersetzung in die russische Sprache; 3) Beispiele für die Kompatibilität mit der Übersetzung ins Russische; 4) ungewöhnliche Kompatibilität (Kollokationen, feste Redewendungen); 5) grammatische Besonderheiten; 6) Komposita; 7) Derivate (falls vorhanden).

Kompatibilität der lexikalischen Einheiten wird mit Hilfe der semantischen Rollen (weiter SR*)* beschrieben: Das sind z. B. Agent, Object, Time u. a. Zurzeit wird in „Compreno" 365 SR verwendet, was die Beschreibung aller möglichen Relationstypen zwischen dem Hauptwort und abhängigen Wörtern ermöglicht.

Ein bestimmtes Set von semantischen Rollen stellt eine Eigenschaft der universellen USH-Einheit, der semantischen Klasse dar, dabei erben die Tochterknoten dieser SK den ganzen SR-Set automatisch von ihren Mutterknoten. Der SR-Satz kann an der jeweiligen SK modifiziert werden. Jede semantische Rolle hat einen Satz von den sog. „Argumentklassen“: Das sind semantische Klassen, die diese semantische Rolle ergänzen können. Dieses Set kann an einer konkreten SK modifiziert werden, was die Homonymie der lexikalischen Klassen zu lösen hilft.

Wir führen weiter einige Beispiele für die Beschreibung des Modells vom Verb *aussehen* (von der SK TO_LOOK_AS_TO_SEEM) mit Hilfe der semantischen Rollen an:

Object, State:
[er] sieht [gesund] aus
Sphere:
[In der Politik] sieht das anders aus
Standpoint:
[wirtschaftlich] sieht das anders aus
MetaphoricLocative
[Im Buch] sieht es anders aus
Object, PartComplement_EntityLike:
[der rote Pullover] sieht [zu dem grauen Rock] sehr gut aus
usw.

Für die Beschreibung der festen Kompatibilität der deutschen Lexik verwenden wir *lexikalische Funktion* (vgl. МЕЛЬЧУК et al. 1984) und *Wortverbindungen*. In „Compreno“ wird die lexikalische Funktion folgenderweise realisiert: Wenn eine SK die Eigenschaft „lexikalische Funktion“ hat, kann man einer konkreten semantischen Rolle eine lexikalische bzw. semantische Klasse vorgeben. Somit kann man durch die lexikalische Funktion für zwei Synonyme aus der SK TO_TAKE_PLACE *spielen* und *geschehen* Folgendes bestimmen: *Die Handlung des Romans spielt in Spanien*, aber **Die Handlung des Romans geschieht in Spanien.*

Durch die Funktion „Wortverbindung“ kann man in die USH nicht nur einzelne Wörter, sondern auch Wortverbindungen einführen und solche feste Wortverbindungen wie *Schi laufen*, *zu Mittag essen* u. a. beschreiben.

Es werden auch grammatische Besonderheiten des Wortes in der jeweiligen Bedeutung beschrieben, z. B. die Fähigkeit eines Wortes, einen Infinitivsatz anzuschließen. Bei vielen Wörtern ist diese Eigenschaft lexikalisiert und kann bei ihren nahen Synonymen fehlen. Das Substantiv *Vorteil* besitzt diese Eigenschaft, während sein Synonym *Plus* sie nicht hat:

- *Das neue Auto hat den Vorteil, weniger Benzin zu verbrauchen* und *dass es weniger Benzin verbraucht.*
- *Das neue Auto hat das Plus, *weniger Benzin zu verbrauchen*, aber *dass es weniger Benzin verbraucht.*

5.3 Die Derivation der deutschen Verben und ihre Darstellung im Rahmen der USH

Die Produktivität der Wortbildung durch Präfixe ist kennzeichnend für deutsche Verben: „Den weitaus größten Anteil an der Verbbildung haben Präfixe und Halbpräfixe" (DROSDOWSKI 1995: 434). Dabei wird oft die Bedeutung des Verbs durch das Hinzufügen eines Präfixes nicht völlig geändert, sondern nur modifiziert und präzisiert. Die Präfigierung ist dermaßen produktiv, dass das «Ausgangsverb» Dutzende von Derivaten haben kann. LK *wachsen* (von der SK TO_GROW) z. B. hat 13 Derivate: *durchwachsen*, *einwachsen* u. a.

In der USH werden die Derivate mit ähnlicher Bedeutung in gleicher LK wie das Ausgangsverb beschrieben, indem alle allgemeinen Eigenschaften übernommen werden und die Bedeutungsschattierungen durch Semanteme kodiert werden. Die Beschreibung aller Derivate im Rahmen einer LK macht die lexikografische Beschreibung bequemer und rationaler, d. h. es ist nicht notwendig, die gesamte Valenz/Kombinierbarkeit der Derivate vollständig zu beschreiben – es ist vielmehr ausreichend, auf spezifische Eigenschaften dieser Derivate hinzuweisen.

Regelmäßige Wortbildungsmodelle werden in „Compreno" mit Hilfe von Derivatemen beschrieben. Unter Derivatemen versteht man im Projekt abstrakte Einheiten, die durch Semanteme und Grammeme (spezielle Funktionen, die eine bestimmte grammatische Bedeutung angeben) für mehrere Derivate die allgemeine semantische Komponente und das syntaktische Verhalten/Modell bezeichnen. Momentan gibt es in „Compreno" 125 Derivateme, davon 199 Verbderivateme, für die Beschreibung der deutschen regelmäßigen Derivation. Zum Beispiel ist das Derivatem *An_Approach* solchen Derivaten wie *anlaufen*, *anfliegen*, *anfahren* u. a. zugeschrieben; das Derivatem *Fort_Progressive* ist den Verben *fortarbeiten*, *fortentwickeln*, *fortfahren* u. a. zugeschrieben.

6 Deutsche Komposita in „Compreno“[16]

Komposita werfen in Bezug auf maschinelle Textverarbeitung folgende Fragen auf:

- In welchen Fällen sind Komposita für das Verständnis ihrer Bedeutung zu spalten und in welchen nicht?
- Wie wird von der Maschine bestimmt, wo die Grenze zwischen Kompositа-Teilen liegt; wie sind die Fugenelemente auszugliedern?
- Wie ist die Bedeutung der Zusammensetzung insgesamt zu verstehen?

Diese und ähnliche Fragen werden in vielen Artikeln behandelt (s. GOLDSMITH/REUTTER 1998, BARONI et al. 2002, HENRICH/HINRICHS 2010). Oft wird die Meinung zum Ausdruck gebracht, dass eine Zusammensetzung nicht als eine selbständige lexikalische Einheit, sondern als eine zusammengeschriebene Wortverbindung, derer linker Teil vom rechten Teil semantisch abhängig ist, betrachtet werden soll:

In German as in some other languages, compounds are commonly written as single orthographic strings. Because compounding is a very productive process, this leads to a considerable amount of orthographic words that cannot, even in principle, be listed in a lexicon. We present a solution to this problem based on the idea that compounds should not be predicted as units, but as the concatenation of their components. (BARONI et al. 2002: 470)

Im Rahmen unseres Systems unterscheiden wir zwischen solchen Komposita wie *Tierpark*, *Friedhof*, *Fleischwolf* einerseits und *Tischschublade*, *ressourceneffizient*, *wasserhaltig*, *Nominalkomposita*, *denkfähig* andererseits. Die erste Gruppe von Zusammensetzungen sind solche Komposita, die einen einzelnen Begriff, der sich nicht aus der Summe von Bedeutungen seiner Teile entschließen lässt, darstellen. Ihre Gliederung in Bestandteile gibt keine richtige Vorstellung über ihre Bedeutung. So ist z. B. *Tierpark* nicht ein beliebiger Park mit Tieren, sondern eine spezifische Art des zoologischen Gartens. Solche Komposita sind meistens in Wörterbüchern präsent und verfügen über ihr eigenes Kompatibilitätsmodell, in der USH haben sie ihre eigene semantische Klasse, die in anderen Sprachen in der Regel mit den Wörtern, die diesen Begriff bezeichnen, ergänzt ist: z. B. *Fleischwolf* – engl. *mincer*, russ. *мясорубка*.

Die zweite Gruppe von Komposita bilden zusammengeschriebene Wortverbindungen. Um solche zu verstehen, soll man sie in Teile gliedern können.

16 Wir danken unseren Kolleg(inn)en Olga Loginova, Polina Antonova und Alexander Golovin für nützliche Hinweise zu diesem Thema.

Die Art der semantischen Verbindung zwischen dem rechten und linken Teil soll auch bestimmt werden. Solche Komposita sind sehr wenig in den Wörterbüchern präsent: Das sind meistens relativ feste Bildungen. In der USH platzieren wir solche Komposita nicht, sondern bearbeiten sie als normale Wortverbindungen, z. B.

Lebensqualität: #[Object_Situation:Lebens“leben:leben:LIVE“]Predicate_Noun:qualität „Qualität:Qualität:QUALITY_CLASSIFICATION“

Um eine richtige Analyse von Komposita zu ermöglichen, werden der USH nicht nur einzelne Wörter, sondern auch Präfixoide und Suffixoide hinzugefügt:

wasserhaltig: #[Object:wasser“Wasser:WATER“] ParticipleRelativeClause:haltig „haltig:TO_CONTAIN“ – die Suffixoide *-haltig*

Endstation: #[OrderInTimeAndSpace:End“End:RELATIVE_ORDER“] Predicate_Noun:station „Station: SHORT_TERM_STOP_FOR_TRANSPORT“ – die Präfixoide *End-*

Die Komposita der zweiten Gruppe werden der USH nicht hinzugefügt, da ihre Zahl potenziell unvorhersehbar ist. Sie sind ihren Teilen nach zu analysieren. Die Komposita der ersten Gruppe kann man nach den Teilen analysieren, wir lassen jedoch diese Analyse nicht zu, um der Fehldeutung vorzubeugen.

Bei der Analyse der Komposita der zweiten Gruppe hat man folgendes Problem: Ein und dasselbe Kompositum kann mehrere semantische Interpretationen haben, dabei ist nur eine davon richtig. Das Problem stellt ein Sonderfall dar: Es handelt sich dabei um ein für maschinelle Sprachverarbeitung allgemeines Problem der Homonymie. So ist es für das NLC-System problematisch zu bestimmen, wie man *Jugendarbeit* korrekt analysieren kann, da mindestens zwei semantische Interpretationen möglich sind:

#[Agent:Jugend“Jugend:YOUTH“] Predicate_Noun:arbeit“arbeiten: TO_WORK“ – die Arbeit der Jugend

#[ContrAgent:Jugend“Jugend:YOUTH“] Predicate_Noun:arbeit“arbeiten: TO_WORK“ – die Arbeit mit der Jugend

Mit Hilfe von präzisen semantischen Begrenzungen wird meistens die richtige Analyse erreicht. In den Fällen, wo es schwierig ist (wo semantische Begrenzungen nicht effektiv sind), wird die richtige Analyse durch die Funktion „Wortverbindung“ gesichert.

Es ist nicht immer möglich, aufgrund der semantischen Kriterien zwischen den Komposita der ersten und der zweiten Gruppe zu unterscheiden. Deswegen lassen wir uns zuerst von einem formalen Kriterium leiten: Ob der linke Teil vom Kompositum sich beiordnen kann. Wenn ja, dann interpretieren wir dieses Wort als Kompositum der zweiten Gruppe und analysieren es nach Teilen, indem das semantische Verhältnis zwischen dem linken und dem rechten

Teil festgestellt wird (genauer, die semantische Rolle, die der linke Teil beim rechten ergänzt). Ansonsten kann das NLP-Programm nicht bestimmen, dass z. B. *Abfall- und Ressourcennutzung = Abfallnutzung und Ressourcennutzung* und *Bildungspolitikerinnen und -politiker = Bildungspolitikerinnen und Bildungspolitiker.* Manchmal tendiert ein Kompositum zur ersten Gruppe, wir betrachten es aber als eine Wortverbindung wegen der Beiordnung.

Das Problem der Fugenelemente wird dadurch gelöst, dass jedes Lexem eine bestimmte Form für Komposita hat. Manchmal haben Homonyme verschiedene Formen für Komposita, was auch die Homonymie als Problem lösen kann. Zum Beispiel hat die LK *DEADLY_DEGREE: Mord* nur eine Form *Mords-: Mordsglück*, *Mordshunger*, und ihm homonyme *Mord* aus der LK *morden:TO_KILL(1:3204567)* – nur die Form *Mord: Morddrohung*, *Mordinstrument*. Dementsprechend kann die zweite Reihe von Komposita nie über die SK *DEADLY_DEGREE: Morddrohung* – *‚sehr starke Drohung' usw. analysiert werden.

Die Zusammensetzungen, die aus drei und mehr Teilen bestehen, sind für die Analyse auch problematisch, da die Verbindung der Teile sowohl konsequent als auch parallel erfolgen kann. Meistens ist es jedoch eine konsequente Verbindung, z. B. *(Daten + Verarbeitung) + Gerät = Datenverarbeitungsgerät.* Wenn der erste linke Teil von einem Kompositum wegen semantischer Begrenzung nicht an den nächsten rechten Nachbarteil gebunden werden kann, wird er an den letzten rechten Teil gebunden, und dabei betrachten wir das als parallele Verbindung: Im Kompositum *Hochschulsommerkurs* wird das Element *Hochschul-* semantisch nicht an *Sommer-*, sondern an *-kurs* gebunden.

7 Die Bedeutung der Vollständigkeit der Beschreibung für Disambiguierung von Textwörtern

Die Schwierigkeit bei der Analyse der deutschen Komposita ist ein Beispiel für Homonymie, die zu den zentralen Problemen von NLP-Systemen gehört (vgl. KUNZE/LEMNITZER 2007: 58). Im Rahmen von „Compreno" ist unter der lexikalischen Homonymie die „Zugehörigkeit eines Lexems zu mehreren lexikalischen Klassen" zu verstehen (ANISIMOVICH et al. 2012: 102). Die Tatsache, dass es in einem oder anderen Kontext schwierig ist, zu bestimmen, zu welcher lexikalischen Klasse ein Lexem gehört, stellt ein Problem dar.

Das System der semantischen Rollen hilft in vielen Fällen, die Textwörter zu disambiguieren. Homonyme aus verschiedenen USH-Teilbäumen unterscheiden sich durch verschiedene Sets von semantischen Rollen, ihre Argumentklassen und Distribution (semantisches Modell ihrer syntaktischen Hauptwörter).

In der USH gibt es drei homonyme Lexeme für *Information*, die in folgenden LK dargestellt sind:

TO_INFORM : informieren (‚das Informieren')
INFORMATION : Information (‚Ereignisse, Fakten')
INFORMATION_BUREAU : Information (‚Informationsstand')

Diese Homonymie ist in meisten Kontexten unschwer zu unterscheiden, da das semantische Modell dieser LK und ihre Distribution nicht übereinstimmen. So wird z. B. im Satz *Der Bürgermeister verhindert Information der Bevölkerung über Verkabelung* das Lexem *Information* durch die erste LK interpretiert, da die Objektposition bei diesem Verb nur Substantive, die Situationen bezeichnen, ergänzen können.

Schwer zu unterscheiden ist die Bedeutung in Kontexten wie *Die Information*, *dass der Irak des Saddam Hussein über Massenvernichtungswaffen verfügt*, *war ein wesentliches Argument*, da *Information* hier sowohl über die erste, als auch über die zweite LK interpretiert werden kann. In diesem Fall soll der Lexikograf die Häufigkeit des *dass*-Satzgliedes für beide LK, höhere Relevanz der abhängigen Wörter für erste LK (z. B. in der semantischen Rolle *Addressee*[17]) usw. analysieren.

Obwohl die Homonyme oft schwierig zu unterscheiden sind, kann man dieses Problem lösen, indem man die Homonyme nach der Kompatibilität disambiguiert. Deswegen legen die Lexikografen der deutschen Abteilung von „Compreno" besonderen Wert auf eine möglichst vollständige und akkurate Beschreibung der Kompatibilität der Wörter.

8 Fazit

Das deutsche semantische Wörterbuch, das Teilprojekt von „Compreno", hat zum Ziel, den realen Wortgebrauch in der gegenwärtigen deutschen Sprache mit computerlinguistischen Methoden zu beschreiben. Das Projektwörterbuch richtet sich an wichtige Prinzipien der maschinellen Sprachverarbeitung aus: Man strebt nach der möglichst vollständigen Beschreibung. Das stimmt für den geplanten Wörterbuchumfang, der sowohl Kernwortschatz als auch sprachliche Peripherie, einschließlich Fachwörter, umfassen soll, sowie auch in Bezug auf die ausführliche Illustration der Kompatibilität für jedes Wort in jeder Bedeutung mit der Unterscheidung von Homonymen und Lesartendisambiguierung.

Dank der ausführlichen lexikografischen Bearbeitung entsprechend der Projektdokumentation kann das Wörterbuch wichtige Probleme der maschinel-

17 Adressat der genannten Handlung.

len Sprachverarbeitung, die Bedeutungsunterscheidung, Komposita-Analyse u. a. lösen. Obwohl sich das deutsche semantische Wörterbuch „Compreno“ ursprünglich auf NLP-Systeme richtet, kann es jedoch auch für Deutschlernende nützlich sein, da es vollständig und gründlich die Kompatibilität der Wörter illustriert und zusätzliche Informationen gibt, die in traditionellen Wörterbüchern nicht vorhanden sind. Somit halten wir es für möglich, dass dieses semantische Wörterbuch sowohl im Rahmen des NLP-Projektes als auch als selbständige lexikografische Quelle benutzt werden kann.

Literaturverzeichnis:

ANISIMOWICH, K. V./DRUZHKIN, K. Y./MINLOS, P. R./PETROVA, M. A./SELEGEY, V. P./ZUEV, K. A. (2012); Syntactic and Semantic Parser based on ABBYY „Compreno“ Linguistic Technologies. In: Kompyuternaia Lingvistika I Intellektual'nye Technologii: Trudi Mezhdunarodnoj Konferentsii „Dialog 2012“T. 2 [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference „Dialog 2012“Vol. 2]. Moskau: Izd-vo RGGU, S. 91–103.

BARONI, Marco/MATIASEK, Johannes/TROST, Harald (2002): Predicting the Components of German Nominal Compounds. In: ECAI 2002, Proceedings of the 15th European Conference on Artificial Intelligence. Amsterdam: IOS Press, S. 470–474.

Digitales Wörterbuch der deutschen Sprache. URL: http://www.dwds.de/ [26.02.2013]

DROSDOWSKI, Günther (1995): DUDEN Grammatik der deutschen Gegenwartssprache. Mannheim: Dudenverlag.

EuroWordNet.URL: http://www.illc.uva.nl/EuroWordNet/ [26.02.2013]

FILLMORE, Charles J. (1982): Frame Semantics. In: Linguistics in the Morning Calm. Seoul: Hanshin, S. 111–138.

FrameNet. URL: https://framenet.icsi.berkeley.edu/fndrupal/ [26.02.2013]

GÖTZ, Dieter/HAENSCH, Günther/WELLMANN, Hans (2008): Langenscheidt Großwörterbuch Deutsch als Fremdsprache. Berlin u. a.: Langenscheidt.

GermaNet. A German Wordnet. URL: http://www.sfs.uni-tuebingen.de/lsd/ [26.02.2013]

German FrameNet. URL: http://www.laits.utexas.edu/gframenet/ [26.02.2013]

GOLDSMITH, John/REUTTER, Tom (1998): Automatic collection and analysis of German Compounds. URL: http://acl.ldc.upenn.edu/W/W98/W98-0609.pdf [26.02.2013]

HaGenLex. URL: http://pi7.fernuni-hagen.de/research/hagenlex/hagenlex-de.html [26.02.2013]

HENRICH, Verena/HINRICHS, Erhard (2010): Standardizing Wordnets in the ISO Standard LMF: Wordnet-LMF for GermaNet. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), August 2010. Beijing: Coling 2010 Organizing Committee, S. 456–464.

KEMPCKE, Günter (2000): Wörterbuch Deutsch als Fremdsprache. Berlin, New York: de Gruyter.

KROTT, Andrea (2007): Analogical effects on linking elements in German compounds. URL: http://www.sfs.uni-tuebingen.de/~hbaayen/publications/KrottEtAlLCP.pdf [26.02.2013]

KUBCZAK, Jacqueline (2011): E-VALBU – Das elektronische Valenzwörterbuch deutscher Verben. URL: http://hypermedia2.ids-mannheim.de/evalbu/index.html [26.02.2013]

KUNZE, Claudia/LEMNITZER, Lothar (2007): Computerlexikographie. Eine Einführung. Tübingen: Gunter Narr Verlag.

Lexipedia. URL: http://www.lexipedia.com/ [26.02.2013]

LIBBEN, Gary (1998): Semantic transparency in the processing of compounds: consequences for representation, processing, and impairment. In: Brain and Language. 61(1), S. 30–44.

Limas-Korpus. URL: http://www.korpora.org/Limas/ [26.02.2013]

NOY, Natalya F./McGUINNESS, Deborah L. (2001): Ontology Development 101: A Guide to Creating Your First Ontology. URL: http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness-abstract.html [26.02.2013]

Online-Wortschatz-Informationsystem Deutsch. URL: www.owid.de [26.02.2013]

Salsa II. URL: http://www.coli.uni-saarland.de/projects/salsa/page.php?id=overview [26.02.2013]

SCHOLZE-STUBENRECHT, Werner (1999): Duden. Das große Wörterbuch der deutschen Sprache. Mannheim: Dudenverl.

WAITE, Maurice (2000): New Oxford Thesaurus of English. Oxford: Oxford University Press.

WordNet. A lexical database for English. URL: http://wordnet.princeton.edu/ [26.02.2013]

Wortschatz. Universität Leipzig. URL: www.wortschatz.uni-leipzig.de [26.02.2013]

ДОБРОВ Б.В./ЛУКАШЕВИЧ Б.В.(2006): Онтологии для автоматической обработки текстов: описание понятий и лексических значений. URL: http://www.dialog-21.ru/digests/dialog2006/materials/html/Dobrov.htm [26.02.2013]

МЕЛЬЧУК И.А./ЖОЛКОВСКИЙ А.К. и др. (1984): Толково-комбинаторный словарь современного русского языка. Опыты семантико-синтаксического описания русской лексики. Wien: Wiener Slavistischer Almanach.