6

Lexikalische und ontologische Ressourcen

In diesem Kapitel stellen wir Ihnen populäre maschinenverarbeitbare Ressourcen für die Lexikographie, Computerlinguistik und Sprachtechnologie vor. Mit Wortnetzen lernen Sie einen besonders beliebten Ressourcen-Typ kennen, der durch seine einfache Struktur und hohe konzeptuelle Abdeckung in zahlreichen Szenarien angewendet wird. Mittlerweile gehören Wortnetze zur Grundausstattung der Ressourcen-Infrastruktur einer Sprache. Mit FrameNet ist ein komplexerer, auf der Frame-Semantik basierender Ansatz entwickelt worden, welcher vor allem semantische und syntaktische Strukturen im Umfeld von Verbkonzepten erfasst und computertechnisch auch für theoretische Fragestellungen verfügbar macht. Wir geben Ihnen abschließend einen Überblick über Ontologien, die Wissensmodellierungen jeglicher Art zum Inhalt haben, und in vielen wissenschaftlichen und technischen Disziplinen sowie in Web-Anwendungen eine zunehmend wichtige Rolle spielen.

1 Überblick

In diesem Abschnitt stellen wir elektronische Lexikonressourcen vor, die als maschinenverarbeitbare Wörterbücher („machine-tractable dictionary “, kurz: MTD) und lexikalische Wissensbasen („lexical knowledge base “, kurz: LKB) in computerlinguistischen Anwendungen und sprachverarbeitenden Prozessen genutzt werden. Die wesentlichen Anwendungsszenarien in der Sprachverarbeitung umfassen

- die Lesartendisambiguierung;
- die Informationserschließung und Informationsextraktion;
- die linguistische Annotierung von Sprachdaten auf verschiedenen Beschreibungsebenen;
- die Textklassifikation und automatische Textzusammenfassung;
- die Entwicklung von Werkzeugen für die Sprachanalyse bzw. -generierung;
- die maschinelle oder maschinengestützte Übersetzung.

Wir beschreiben zunächst mit lexikalisch-semantischen Wortnetzen einen Typus semantischer Online-Lexika, der seit der Entwicklung des Princeton WordNet sehr populär geworden ist. Neben einer Vielzahl bereits existierender Ressourcen gibt es zahlreiche Initiativen zum Aufbau einzelsprachlicher und sprachübergreifender Wortnetze bzw. Wortnetzverbünde. Während Wortnetze auch in sprachtechnologischen Anwendungen, die nicht genuin (computer-)linguistisch motiviert sind, beliebte Hintergrundressourcen darstellen, ist mit dem aus der Fillmoreschen Frame-Theorie hervorgegangenen FrameNet ein Ressourcentyp entstanden, der vor allem für theoretische (computer-)linguistische Fragestellungen relevant ist. So werden die gegenüber Wortnetzen reichhaltigeren Frames für die Analyse und Annotierung von Sprachkorpora eingesetzt. Frames sind außerdem durch die stark konzeptuelle Ausprägung besser geeignet für Belange der Universalienforschung und der maschinellen Übersetzung bzw. Interlinguaforschung. In einem weiteren Abschnitt stellen wir mit Ontologien Begriffsnetze aus der künstlichen Intelligenz (KI), Informatik und Semantic-Web-Forschung vor, welche in zahlreichen (kommerziellen) Szenarien der Sprachverarbeitung eine zentrale Rolle einnehmen. Ontologien werden nach strengeren Kriterien formalisiert als Wortnetze und spielen in Bezug auf die Modellierung spezifischer Fachdomänen eine wichtige Rolle. Sie sind daher als Organisationsform für (fachsprachliche) Konzepte in vielen wissenschaftlichen Disziplinen, wie z.B. der Biotechnologie und Medizin, von großem Nutzen.

Nicht explizit beschreiben, sondern nur erwähnen wollen wir in unserer Darstellung die Open-Source-Lexika und -Enzyklopädien, die gemeinschaftlich durch eine Vielzahl von Nutzern bzw. Autoren aufgebaut wurden und sich in der stetigen Weiterentwicklung befinden. Diese dynamischen Formen der Lexikographie ebenso wie das „Web as Corpus“ - Projekt werden zukünftig eine wichtige Rolle spielen und die bislang hierarchisch-normativ geprägte Lexikographiepraxis nachhaltig verändern. Nachteilig zum gegenwärtigen Zeitpunkt sind die mangelnde Kontrolle und Konsistenz bei der Erstellung der Lexikonartikel.

Ebenfalls von der Betrachtung ausgenomen bleiben in dieser Einführung Lexika für multimodale Systeme (vgl. Gibbon (2001)), die aufgrund der Komplexität ihrer enthaltenen Daten auf mehreren Repräsentationsebenen, z.B. zur Kennzeichnung von Wortbetonungen oder begleitenden Gesten, bislang nicht über Prototypenstatus hinausgehen[1].

[1] Vgl. das Projekt MODELEX an der Universität Bielefeld, `http://coral.lili.uni-bielefeld.de/modelex/`.

Auch soll an dieser Stelle auf die für Dialogsysteme wichtigen Lexika für gesprochene Sprache lediglich verwiesen werden[2].

[2] Vgl. das Bayrische Archiv für Sprachsignale: `http://www.phonetik.uni-muenchen.de/Bas/BasHomedeu.html`.

2 Lexikalisch-semantische Wortnetze

2.1 Einleitung

In diesem Kapitel werden lexikalisch-semantische Wortnetze im Stile des Princeton WordNet (vgl. Miller (1990), Fellbaum (1998)) als eine besondere Spielart elektronischer Ressourcen, als so genannte Online-Thesauri, vorgestellt. Solche Wortnetze bilden die häufigsten und wichtigsten Wörter einer Sprache und ihre bedeutungstragenden Beziehungen zu anderen Wörtern der Sprache ab. Im Wortnetz ist ein Wort als Konzeptknoten mit seinen semantischen Verknüpfungen repräsentiert: z.B. *Stuhl* mit dem Oberbegriff *Sitzmöbel* und seinen Unterbegriffen *Drehstuhl*, *Klappstuhl*, *Kinderstuhl* etc. Der Oberbegriff ist darüber hinaus mit den Konzepten *Lehne*, *Sitzfläche* und *Bein* verbunden, die Teile eines Sitzmöbels repräsentieren, vgl. Abb. 23. Ein Konzept ist al-

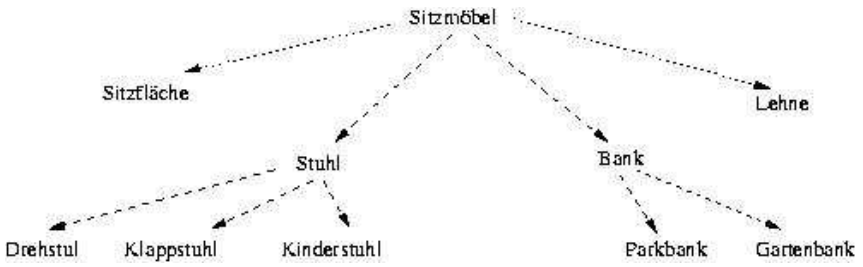

Abbildung 23: Ausschnitt aus der GermaNet-Hierarchie, Teilbaum *Sitzmöbel*

so nicht nur über seinen Knoten, sondern auch über seine Relationen charakterisierbar. Da die zugrunde liegende Repräsentationseinheit, das so genannte SYNSET, gleiche Bedeutungen, die Synonyme, zu einem Konzeptknoten zusammenfasst und nicht etwa gleiche Wörter, werden in Wortnetzen Lesarten unterschieden. Diese Lesartendisambiguierung ist eine unabdingbare Voraussetzung für Anwendungen im Bereich der maschinellen Übersetzung und der Informationserschließung, zur semantischen Annotierung von Sprachkorpora und für die Entwicklung verschiedener Werkzeuge zum Sprach- und Informationserwerb und für die Übersetzung. Wortnetze bilden natürlich-sprachliche Hierarchien ab und sind – zumindest vorläufig – von den ONTOLOGIEN aus dem Umfeld der künstlichen Intelligenz zu unterscheiden, die (meist sprachunabhängige oder domänenspezifische) konzeptuelle Begriffsnetze konstituieren. Der folgende Abschnitt beschreibt detailliert das lexikalisch-semantische Wortnetz GermaNet (vgl. Kunze und Naumann (1999-2007)) und des-

sen Einbindung in das polylinguale EuroWordNet, das im Rahmen eines europäischen Projektes 1996-1999 für acht Sprachen aufgebaut wurde (vgl. Vossen (1999)).

2.2 GermaNet – ein deutsches Wortnetz

Mit GermaNet (`http://www.sfs.uni-tuebingen.de/lsd`) ist ein computertechnisch verfügbares semantisches Lexikon aufgebaut und ein wichtiger Beitrag zur wissensbasierten Ressourcenbildung für das Deutsche geleistet worden. Im Wesentlichen orientiert sich das deutschsprachige Wortnetz am Datenbankformat und an den Strukturierungsprinzipien des Princeton WordNet 1.5, das als „Mutter aller Netze“ eine initiale Rolle für viele einzelsprachliche Wortnetz-Initiativen spielte.[3] GermaNet ist jedoch keine pure Übersetzung des WordNet, sondern setzt eigene Schwerpunkte in der Konzeptrepräsentation (vgl. Hamp und Feldweg (1997)). GermaNet ist aus verschiedenen lexikographischen Quellen, z.B. dem Wehrle und Eggers (1989) und dem Brockhaus-Wahrig (1980-1984), und unter der Berücksichtigung von Korpusfrequenzen von Hand aufgebaut worden. In GermaNet sind die bedeutungstragenden Kategorien der Nomina, Verben und Adjektive modelliert. Zentrales Repräsentationskonzept ist das Synset, welches die Synonymenmenge eines gegebenen Konzeptes bereitstellt, z.B. {*Streichholz, Zündholz*}, {*fleißig, eifrig, emsig, tüchtig*} und {*vergeben, verzeihen*}. Im Wortnetz sind semantische Relationen zwischen den Konzepten (Synsets) oder einzelnen Varianten (Synonymen aus den Synsets) kodiert. Zur Zeit enthält GermaNet ca. 53 500 Synsets mit ca. 76 500 Lexical Units, davon knapp 39 000 Nomen, 9 000 Verben und 5 500 Adjektive. Das deutsche Wortnetz wird durch den Abgleich der Datenbankeinträge mit Frequenzlisten aus Korpora systematisch um fehlende Konzepte ergänzt. GermaNet repräsentiert nur wenige Mehrwortlexeme wie *gesprochene Sprache* oder *Neues Testament*. Eigennamen treten hauptsächlich im Wortfeld der Geographie auf, z.B. als Städtenamen, und werden speziell markiert.

Relationstypen in GermaNet

Die Aussagekraft semantischer Netze liegt in den zahlreichen sinnhaften Verknüpfungen zwischen den repräsentierten Knoten. GermaNet unterscheidet zwischen LEXIKALISCHEN und KONZEPTUELLEN RELATIONEN:

- Lexikalische Relationen sind bidirektionale Beziehungen zwischen Wortbedeutungen wie die Synset-interne SYNONYMIE (Bedeutungs-

[3] Das Urmodell semantischer Netze entwickelte Quillian (vgl. Quillian (1966)) zur Modellierung des semantischen Gedächtnisses innerhalb der KI.

gleichheit zwischen *Ruf* und *Leumund*) und die ANTONYMIE (Gegenteiligkeit), etwa zwischen *Geburt* und *Tod*, *glauben* und *zweifeln*, *schön* und *hässlich*.

- Konzeptuelle Relationen wie HYPONYMIE, HYPERONYMIE, MERONYMIE, IMPLIKATION und KAUSATION bestehen zwischen Konzepten, gelten also für alle Realisierungen innerhalb eines Synsets. Hyponymie und Hyperonymie konstituieren KONVERSE RELATIONSPAARE: so ist *Gebäude* das Hyperonym zu *Haus* und *Haus* ein Hyponym von *Gebäude*.

Das wichtigste Strukturierungsprinzip in semantischen Netzen stellt die hierarchiebildende Hyponymierelation, wie sie z.B. zwischen *Rotkehlchen* und *Vogel* besteht, dar. Besonders die Nomina haben Ketten mit tiefen Hierarchien, wie z.B. das Konzept *Kieferchirurg* mit 15 Dominanzstufen. In GermaNet sind auch die Verben und Adjektive taxonomisch (d.h. unter Rückgriff auf die Hyponymierelation) gegliedert. Die Meronymierelation (Teil-Ganzes-Beziehung) wird nur für Nomina angenommen: Ein *Dach* kann nicht angemessen als eine Art *Gebäude* klassifiziert werden, sondern ist Teil eines *Gebäudes*. Teil-Ganzes-Beziehungen können auch abstrakter Natur sein, z.B. in Bezug auf die Mitgliedschaft in einer Gruppe (*Vorsitzender* einer *Partei*) oder als Material in einer Komposition (*Fensterscheibe* aus *Glas*). Typischerweise wird die Verknüpfung zwischen lexikalischen Resultativen wie *töten* und *sterben* oder *öffnen* und *offen* als KAUSATIONSRELATION spezifiziert. Die kausale Relation kann klassenübergreifend zwischen allen Kategorien kodiert werden. Seltener hingegen wird von der Implikationsbeziehung oder dem ENTAILMENT Gebrauch gemacht, wie etwa zwischen *gelingen* und *versuchen*. Die Bedeutung eines Wortes ist durch die Gesamtheit der Relationen, die sie zu anderen Wortbedeutungen aufweist, gekennzeichnet. Es gibt in GermaNet über die ausführlich beschriebenen Relationen hinausgehend noch die PERTONYMIE (eine Art semantischer Derivationsbeziehung wie z.B. zwischen *finanziell* und *Finanzen*) und eine ÄHNLICHKEITSRELATION (SEE ALSO), die assoziativen Verknüpfungen Rechnung trägt wie zwischen *Weltrangliste* und *Tennis* oder *Talmud* und *Judentum*. Abbildung 24 zeigt das kausative Verb *öffnen* mit allen semantisch korrelierten Konzepten. Synsets und Varianten sind mit den entsprechenden Lesartennummern aus GermaNet aufgeführt. Die Verbindung des Synset *öffnen_3*, *aufmachen_2* mit seinem Hyperonym *wandeln_4*, *verändern_2* wird durch den nach oben weisenden Pfeil repräsentiert, mit den drei Hyponymen *aufstoßen_2*,*aufbrechen_1* und *aufsperren_1* durch jeweils abwärts gerichtete Pfeilspitzen, und die kausale Relation zum intransitiven Konzept *öffnen_1*, *aufgehen_1*

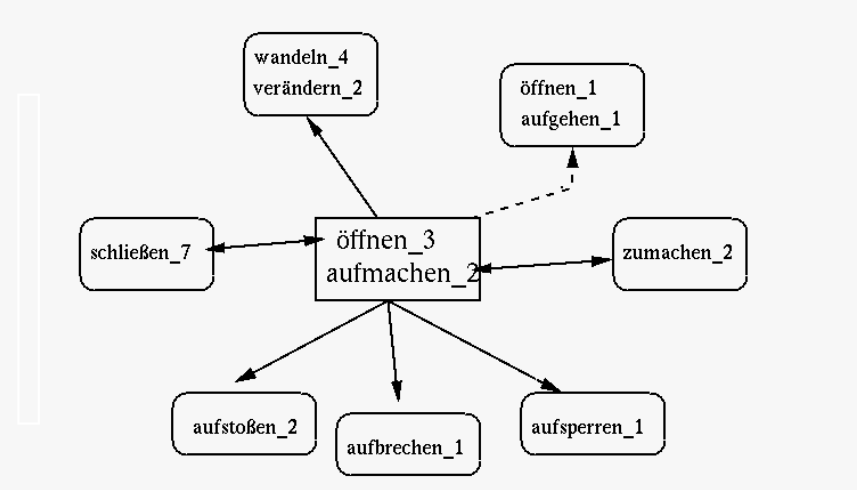

Abbildung 24: Ausschnitt aus der GermaNet-Hierarchie, Teilbaum *öffnen*

durch den Pfeil mit gestrichelter Linienführung. Die beiden Varianten im Synset haben unterschiedliche Antonyme: *öffnen_3* hat als Antonym *schließen_7*, und *aufmachen_2* das Antonym *zumachen_2*. Die Antonymierelation ist durch den Doppelpfeil gekennzeichnet.

Kreuzklassifikation und künstliche Konzepte

Ein Konzept wie *Banane* kann ebenso wie eine Reihe weiterer Früchte gleichermaßen als *Pflanze* und als *Nahrungsmittel* klassifiziert und somit unterschiedlichen semantischen Feldern zugeordnet werden. Um diese Information zugreifbar zu machen, empfiehlt sich die KREUZKLASSIFIKATION solcher Konzepte in verschiedenen Hierarchien, vgl. Abbildung 25. Wortnetze sollen nur tatsächlich vorkommende lexikalische

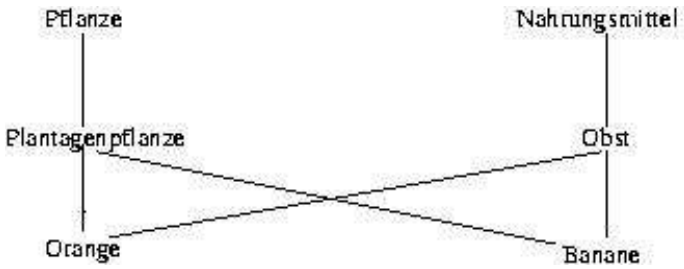

Abbildung 25: Beispiel für eine Kreuzklassifikation

Einheiten einer Sprache abbilden. In GermaNet wird jedoch Gebrauch von KÜNSTLICHEN KONZEPTEN gemacht, wenn diese geeignet sind, die Hierarchie besser zu strukturieren und unmotivierte Ko-Hyponymie zu vermeiden. Nach Cruse (1986) sollten Ko-Hyponyme auf einer Basis von Ähnlichkeit, die durch den gemeinsamen Mutterknoten gegeben ist, möglichst inkompatibel zueinander sein, vgl. *Säugling*, *Kleinkind*, *Vorschulkind*, *Schulkind* als Unterbegriffe zu *Kind*, die einander wechselseitig ausschließen. Im Wortfeld Lehrer sind Unterbegriffe wie *Fachlehrer*, *Berufsschullehrer* und *Konrektor* nicht sinnvoll auf einer gemeinsamen Hierarchieebene anzusiedeln. Um das Teilnetz symmetrischer zu gestalten, werden mit *?Schullehrer* und *?hierarchischer_Lehrer* zwei künstliche Konzepte eingeführt, vgl. Abbildung 26. GermaNet kodiert dar-

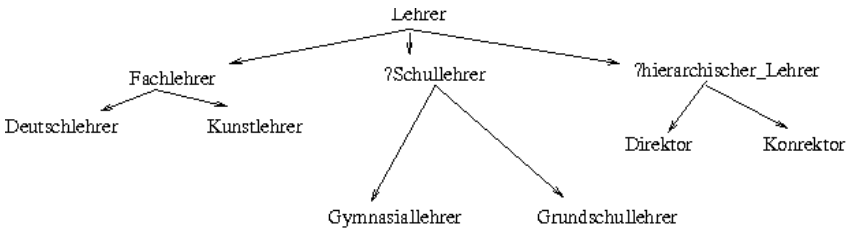

Abbildung 26: Beispiel für die Verwendung artifizieller Konzepte

über hinaus Subkategorisierungsrahmen zur Kennzeichnung des syntaktischen Komplementierungsverhaltens von Verben. Da in dieser Darstellung der Schwerpunkt auf den semantischen Relationen in GermaNet liegt, mögen an den Verbrahmen Interessierte die GermaNet-Homepage konsultieren, zur Erklärung der Notation, Verwendung der Rahmentypen und der Illustration mit entsprechenden Beispielsätzen[4].

2.3 EuroWordNet, ein polylinguales Wortnetz

Das Basisvokabular des GermaNet, etwa 15 000 Synsets, ist in das polylinguale EuroWordNet[5] für acht europäische Sprachen integriert worden, vgl. Vossen (1999). EuroWordNet modelliert die wichtigsten Konzepte des Englischen, Spanischen, Holländischen, Italienischen, Französischen, Deutschen, Tschechischen und Estnischen mit ihren semantischen Relationen. Kernkomponente der Datenbankarchitektur ist der INTERLINGUALE INDEX (ILI), an den die einzelsprachlichen Wortnetze geknüpft sind. Der ILI fungiert als sprachunabhängige Kompo-

[4] S. `http://www.sfs.uni-tuebingen.de/lsd/`.
[5] S. `http://www.hum.uva.nl/~ewn/`.

nente und besteht aus einer unstrukturierten Liste von ILI-Records, die an WordNet Synsets (und somit englischen Konzepten) orientiert und durch einen eindeutigen Code (‚unique identifier'), gekennzeichnet sind. Konzepte der einzelnen Sprachen werden mit sprachübergreifenden Relationen an passende Übersetzungsäquivalente aus dem ILI angebunden. Über den ILI können dann mittelbar spezifische Sprachpaare zu erfragten Konzepten gebildet werden, z.B. *guidare:conducir* (Italienisch:Spanisch) für das Konzept *drive* in Abbildung 27. Zu den sprachunabhängigen Komponenten zählen neben dem ILI die TOP-ONTOLOGIE mit 63 semantischen Merkmalen und die DOMÄNEN-ONTOLOGIE, die semantische Felder zur Verfügung stellt. Alle einzelsprachlichen Wort-

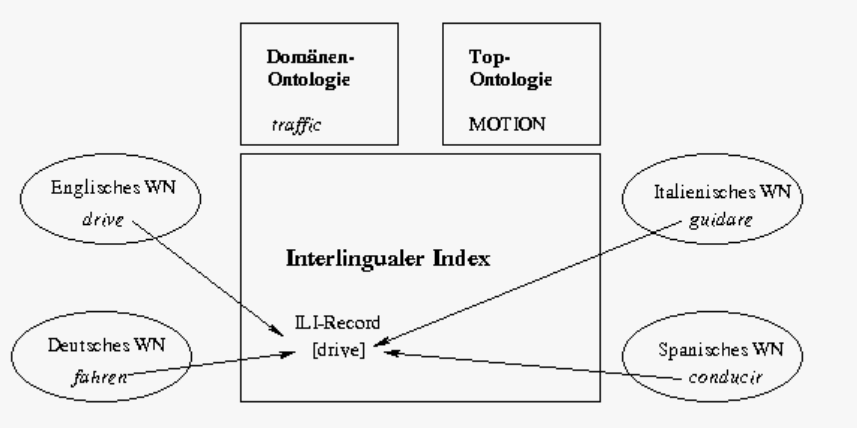

Abbildung 27: Architektur des EuroWordNet

netze enthalten eine gemeinsame Menge so genannter BASE CONCEPTS, 1000 Nomen und 300 Verben, die als zentrales Vokabular des polylingualen Wortnetzaufbaus fungieren und die Kompatibilität der einzelnen Sprachnetze gewährleisten. Base Concepts werden durch semantische Merkmale oder Merkmalskombinationen aus der Top-Ontologie charakterisiert, z.B. *Werkzeug* durch die Merkmale ARTEFACT, INSTRUMENT, OBJECT. Base Concepts dominieren viele Knoten und/oder eine hierarchisch vielstufige Kette von Unterbegriffen oder sie sind häufig auftretende Konzepte in mindestens zwei Sprachen. Sie sollen konkreter als die semantischen Merkmale der Top-Ontologie wie DYNAMIC, FUNCTION und PROPERTY sein, aber wiederum abstrakter als die von Rosch (1978) postulierten BASIC LEVEL CONCEPTS, z.B. *Tisch* und *Hammer*. Der angemessene Abstraktionsgrad für Base Concepts wird von den jeweiligen Oberbegriffen der Basic Level Concepts, z.B. *Möbel* für

Tisch und *Werkzeug* für *Hammer* erreicht. Nachdem das Inventar der Base Concepts mit dem ILI verknüpft worden war, sind Top-Konzepte und Hyponyme erster Ordnung gelinkt worden, was zu einem ersten Datenensemble von ca. 7 500 Synsets führte. Der Aufbau einzelsprachlicher Netze konnte dann unabhängig erfolgen, zumal die Vererbung der semantischen Merkmale der Top-Ontologie ermöglicht, die Abdeckung der Netze in einzelnen semantischen Feldern statistisch zu untersuchen und damit eine gewisse Ausgewogenheit zwischen den Sprachen sicherzustellen. Aufgrund unterschiedlicher Lexikalisierungsmuster der einzelnen Sprachen, die auf sprachliche und kulturelle Unterschiede zurückgehen, und aufgrund von Kodierungslücken im Princeton-WordNet (das ja die Basisressource für den ILI darstellt), können nicht immer angemessene Übersetzungen der einzelsprachlichen Konzepte gefunden werden. Daher sind auch nicht-synonymische sprachübergreifende Verknüpfungen sowie die Kombination mehrerer nicht-synonymischer Links möglich. Z.B. ist für das Konzept *Sportbekleidung* kein synonymisches Targetkonzept *sports garment* im ILI verfügbar. Ersatzweise können zwei sprachübergreifende Links zum Hyperonym *garment* (‚Kleidung') und zum Holonym *sports equipment* (‚Sportausrüstung') etabliert werden. Die internationale Zusammenarbeit zum Aufbau eines polylingualen Wortnetzes hat geholfen, einen Quasi-Standard für Wortnetze zu entwickeln und weist somit Modellfunktion für die Integrierung weiterer Sprachen auf. In diesem Zusammenhang ist im Sommer 2000 die ‚Global WordNet Association' (`http://globalwordnet.org/`) gegründet worden. Mittlerweile gibt es mehrere polylinguale Architekturen, die auf den EuroWordNet ILI zurückgreifen, wie z.B. in BalkaNet, einem Verband (süd-)osteuropäischer Sprachen[6] und CoreNet (für das Chinesische, Koreanische und Japanische)[7] realisiert.

[6] Vgl. Tufiş et al. (2004).
[7] `http://bola.or.kr/CoreNet_Project/`.