Institut für Computerlinguistik
Universität Zürich Sommersemester 2002

Dr. Kai-Uwe Carstensen
Seminar
Aspekte der Wissensrepräsentation für die Computerlinguistik

WISSENSREPRÄSENTATION FÜR DIE MASCHINELLE ÜBERSETZUNG AM BEISPIEL SENSUS

Madeleine Hussmann
Widderfeld 13
6386 Wolfenschiessen
Tel. 041 628 23 52
hussmann@access.unizh.ch 30. Juli 2002

Inhaltsverzeichnis

1. Einführung

Die maschinelle Übersetzung (MÜ), der Ursprung der Computerlinguistik (CL), hat sich im Laufe der Zeit neben vielen neueren Anliegen der CL wie Wissensmanagement und Textgenerierung einreihen müssen. Trotzdem nimmt sie noch immer viele Ressourcen in Anspruch. Sie kennt zwei grundsätzlich verschiedene Techniken, die Transfer-Technik und das Interlingua-Verfahren. Während die Transfer-Technik auf der Syntax basiert, stützt sich das Interlingua-Verfahren auf die Semantik, und zwar auf eine sprachunabhängige Repräsentation von Konzepten. Diese Konzepte sollen die Welt abbilden und der Maschine das Welt- und evtl. sogar Situationswissen zur Verfügung stellen, das eine intelligente Verarbeitung von Sprache voraussetzt. Verschiedene Projekte haben mit grossem Aufwand ein Weltmodell erstellt, das der maschinellen Verarbeitung von natürlicher Sprache (NLP) oder im Bereich der künstlichen Intelligenz (KI) dienen soll. Ein solches Projekt ist z.B. CYC, das bis heute über 600 Personenjahre Aufwand gekostet hat und für die NLP nur in geringem Masse funktioniert (Egger;2002). Die beachtliche Konzeptualisierung der Welt in CYC hat einen dermassen hohen Formalisierungsgrad, dass sie nur für Experten des Systems zugänglich ist. Andere Projekte benutzen eine Ontologie, die spezifisch auf eine Domäne ausgerichtet ist und für grössere, domänenübergreifende Vorhaben zu begrenzt sind. Ein solches Vorhaben wäre die MÜ, welche mit einer sprachunabhängigen Konzeptualisierung erlaubt, grössere Wörterbücher von verschiedenen Sprachen sozusagen anzuhängen. Wie eine solche Wissensrepräsentation für die MÜ aussehen kann, werde ich in dieser Arbeit anhand von SENSUS illustrieren. SENSUS ist die Ontologie des MÜ-Systems PANGLOSS. Zuerst werde ich kurz dieses System vorstellen und auf die Problematik zwischen Theorie und Praxis eingehen. Dann werden die Komponenten von SENSUS kurz beschrieben und dargestellt, wie sie miteinander verbunden werden. Mit Beispielen wird die SENSUS-Struktur näher untersucht und mit dem WordNet verglichen. Am Schluss wird der Frage nachgegangen, wie es zu der so genannten Sprachunabhängigkeit der Interlingua kommt.

2. Das MÜ-System PANGLOSS

Laut Kevin Knight und Steve Luk (1994) gibt es MÜ-Systeme, die für beschränkte Domänen recht gut funktionieren. PANGLOSS ist nun ein Projekt, das sich das Ziel gesetzt hat, in unbeschränkten Anwendungsbereichen Texte, anfänglich von Spanisch, später von Japanisch

auf Englisch zu übersetzen. Dafür braucht es eine sehr breite, umfassende Ontologie. Die Erstellung einer solchen Ontologie von Grund auf würde Jahre und Ressourcen beanspruchen, die niemand zu bezahlen bereit ist. PANGLOSS ist denn ein Projekt, das sich aus vorhandenen Projekten zusammensetzt und vergangene Investitionen nutzt. Finanziert von der US Defense Advanced Research Projects Agency (DARPA) startete PANGLOSS Ende 1991 für erstmal drei Jahre als Partnerprojekt von drei Forschungsinstitutionen. In den Artikeln von Eduard Hovy und Knight (1993), sowie von Knight und Luk (1994) wird beschrieben, welche Projekte PANGLOSS beinhaltet. Es sind dies

- *Panglyzer Parser* vom CRL[1] in der New Mexico State University, Las Cruces
- *Translator's Workstation* vom CMT[2] in der Carnegie Mellon's University, Pittsburgh
- *PENMAN English Generation System* vom ISI[3] in der University of Southern California, Marina del Rey.

Die einzelnen Projekte sind aufgrund unterschiedlicher Theorien entstanden. Bei PENMAN z.B. ist Hallidays *Systemic Grammar* die Grundlage für das System. Die Herausforderung für ein System, das sich aus schon vorhandenen zusammensetzt, ist, dass der Vorgang in erster Linie **prozedural** ausgerichtet ist. Eine Theorie, die diesen Vorgang und das Ergebnis umschreibt, muss nahe an den technischen Möglichkeiten und Experimenten sein. Hovy und Knight (1993) argumentieren in ihrem Artikel, "that well-motivated knowledge sharing is indeed possible, as long as one recognizes that a new kind of theory is needed". Grundsätzlich geht es darum, dass Programmierung und Theorie nicht auseinanderklaffen und am Schluss ein System vorliegt, das intuitiv und ad hoc entstanden ist, das zwar möglicherweise funktioniert, bei dem jedoch niemand mehr nachvollziehen kann warum. Die Herausforderung ist jedoch nicht auf die Schnittstelle Theorie und Implementierung beschränkt. Wie Matthiessen und Halliday (1997) in ihrer Einführung in die in der NLP angewandten *Systemic functional Grammar* darlegen, ist es in der Wissenschaft allgemein so, dass sich ausschliessende Theorien Erkenntnisse bringen, die sich ergänzen. Sie meinen: „[W]e need complementary theoretical perspectives to account for the rich diversity of properties we uncover in the phenomena being studied." Dies scheint ganz besonders auf die Komplexität der Sprache zuzutreffen, wenn es darum geht, sie für die MÜ verfügbar zu machen.

[1] Computing Research Laboratory
[2] Center for Machine Translation
[3] Information Sciences Institute

Wie schon erwähnt, sollte PANGLOSS anfänglich Texte von Spanisch auf Englisch übersetzen. Japanisch als zweite Ausgangssprache ist später hinzugefügt worden. Der ursprüngliche Anwendungsbereich waren Zeitungsartikel. Eine Schnittstelle für manuelle Unterstützung des Systems war vorgesehen. Die Übersetzungen sollten anfänglich manuellen Arbeiten Hilfe leisten und allmählich mehr und mehr automatisiert werden. PANGLOSS heute ist ein Prototyp für die Forschung.

Das System wird von einer Interlingua unterstützt, einer Ontologie, die ca. 50'000 Konzepte in einem breiten Klassifikationssystem vernetzt. Wie die Übersetzungsmaschine selbst ist auch diese Ontologie eine Verbindung und Anpassung von vorhandenen Projekten. Die PANGLOSS-Ontologie wird SENSUS genannt. Im nächsten Abschnitt wird sie näher beschrieben.

3. SENSUS: eine breit angelegte Ontologie

3.1 Die Bestandteile von SENSUS

Die Hauptkomponenten von SENSUS sind das Lexikon, die Hierarchie-Kategorien des WordNets und die Ontologie-Basis (OB), welche mit sehr abstrakten Ausdrücken die sog. Interlingua darstellt. Abbildung 1 (Knight/Luk;1994) zeigt einen Überblick, wie SENSUS zusammengesetzt worden ist. Im Folgenden werden die verschiedenen Komponenten kurz erläutert:

- Das **PENMAN Upper Model** ist ein *top-level* Netzwerk von ca. 200 Knoten, das Teil des PENMAN-Systems ist zur Generierung englischer Sprache. PENMAN stützt sich weitgehend auf Beziehungen zwischen Syntax und Semantik. Wenn ein Konzept unter einem bestimmten Knoten im Upper Model klassifiziert ist, dann übernimmt ein englisches Wort, das sich auf dieses Konzept bezieht, eine bestimmte Standardgrammatik. Ausnahmen sind im Lexikon vermerkt.

- **ONTOS** ist auch eine *top-level* Ontologie mit ähnlichem Umfang wie das PENMAN Upper Model, jedoch entwickelt für die MÜ. Die Struktur entstand aus Erkenntnissen aus Studien der allgemeinen Sprachwissenschaft über Verben. Kasusstrukturen sowie Beschränkungen sind sprachneutral repräsentiert. ONTOS beinhaltet auch Hierarchien, Gradierungen und komplexe Begebenheiten.

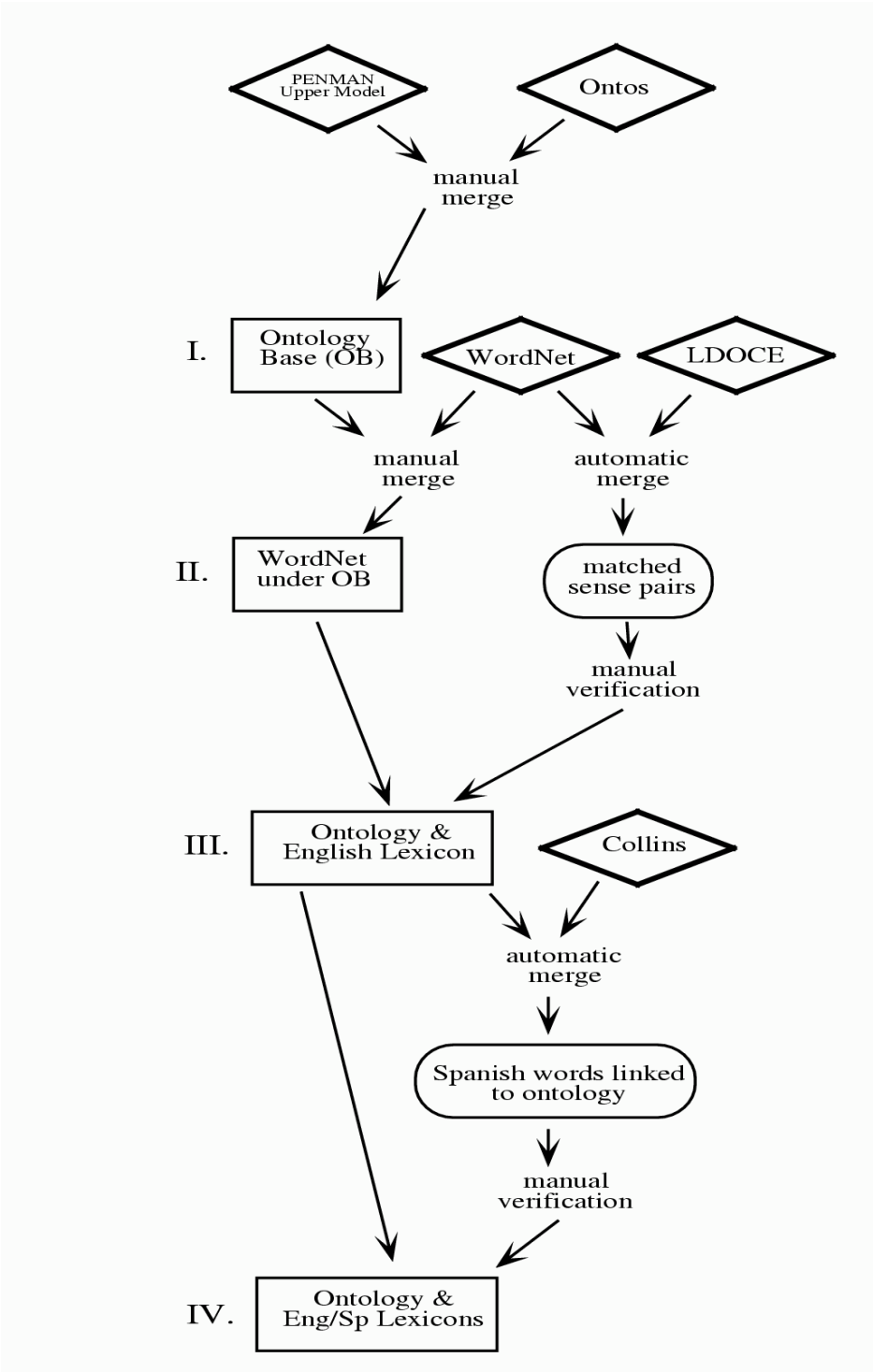

Abbildung 1: SENSUS (Knight und Luk, 1993)

- Das **LDOCE** (Longman Dictionary of Contemporary English) ist ein Englisch Wörterbuch für Lernende mit ca. 28'000 Wörtern und 75'000 Wortbedeutungen. Jede Wortbedeutung hat

  - Eine kurze Definition, wobei sich die Definitionen auf ein Vokabular von 2'000 Wörtern beschränken. Das hat wesentliche Vorteile für den Parser.
  - Anwendungsbeispiele
  - einen oder mehrere der 81 syntaktischen Codes (z.B. [B3]: adj. followed by *to*)
  - für Nomen einen von den 33 semantischen Codes (z.B. [H]: human)
  - für Nomen einen von den 124 pragmatischen Codes (z.B. [ECZB]: economics/business)

- **WordNet** ist ein englischer Thesaurus mit ca. 118'000 Wortformen, 90'000 Wortbedeutungen und 166'000 Worten (d.h. (*f,s*)-Paare).[4] WortNet unterteilt die Worte in fünf syntaktische Kategorien und beinhaltet weit verbreitete, leicht verständliche semantische Relationen wie Synomymie, Antonymie, Hyperonymie, Meronomie und Troponomie.

- **Collins Spanisch/Englisch Wörterbuch**. Wie im LDOCE beinhalten die Wortdefinitionen keine semantischen Informationen. Sie sind aber manchmal versehen mit Themencodes wie [MIL] für Militär, oder [COM] für Handel.

Das PENMAN Upper Model und Ontos liefern die *Ontology Base* (OB), die oberste Ebene und somit die Interlingua. In Kapitel 4 wird dieser Begriff noch weiter untersucht. Sie besteht aus 400 Konzepten, die mit dem Parser und Generator funktionieren müssen. Unter diese OB-Struktur wurden ca. 100 Bäume vom WordNet manuell eingefügt. Dies ergab die zweite Ebene (WordNet under OB, Abbildung 1), die auch *Ontology Body* genannt wird. Sie umfasst ca. 50'000 Konzepte, bildet den Rahmen für ein generisches Weltmodell und repräsentiert viele Bedeutungen von englischen Wörtern. Um die dritte Ebene vorzubereiten wurden WordNet und die Einträge des LDOCE mit einem Algorithmus daraufhin untersucht, ob Definitionen von Wörtern übereinstimmen. Die sich entsprechenden Definitionen wurden integriert und ergaben die dritte Ebene, ein englisches Lexikon mit einer ausgedehnten Ontologie. Von hier aus konnte dann das Spanisch-Englisch Lexikon integriert werden. Im Weiteren wird vor allem auf die ersten beiden Ebenen, OB und *Ontology Body*, eingegangen.

[4] *f* ist Wortform, *s* ist Wortbedeutung

3.2 Veranschaulichung von SENSUS anhand von Beispielen

Bei der Suche im Internet nach Beispielen für SENSUS kommt man an verschiedenen Stellen zu folgendem (Tabelle 1), welches im Original von Swartout et al. stammt (Swartout et al.;1996)[5]:

Tabelle 1: Beispiel 'strut' als Nomen in SENSUS (Swartout et al., 1996)

Sense 1	"strut, swagger" | swagger, strut ==>"manner of walking" | walk, manner_of_walking ===>"bearing, carriage" | carriage, bearing
Sense 2	"strut/brace" | strut =>"brace, bracing" | brace, bracing ==>"structural member" | structural_member ===>"support/supporting structure" | support ====>"supporting structure" | supporting_structure =====>"construction, structure" | structure, construction ======>"instrumentality/artefact" | instrumentality =======>"artifact" | artifact, article, artefact ========>"inanimate object" | object, inanimate_object, physical_object, thing =========>entity
Sense 3	"prance/gait" | strut, prance, swagger

Da weder aus dieser Darstellung noch aus dem Text hervorgeht, warum nur Bedeutung 2 bis an das oberste Konzept ‚entity' reicht, wohinein eigentlich alle Konzepte führen sollten, ging ich der Sache auf den Grund. Im Internet gibt es den *SENSUS Ontosaurus*, einen Browser für diese Ontologie, womit man nach Wortbedeutungen und Konzepten suchen kann. Der Zugang ist allerdings nicht öffentlich.[6] Richard Whitney, Programmierer am ISI und schon seit der Konstruktion vom PENMAN Upper Model in den 80er Jahren dabei, hat freundlicherweise per E-Mail vermittelt. Er erhielt bei SENSUS auf die Eingabe von ‚strut' das in Tabelle 2 dargestellte Resultat. Die erste Zeile, eingeleitet mit „USER" ist jeweils die Anfrage an das System. Die erste Anfrage (0) zeigt, dass SENSUS die Möglichkeit bietet, sie auf der lexikalischen Ebene mit der Wortart (hier Nomen) einzuschränken. Die erste Antwort ergibt zwei Bedeutungen (und nicht drei wie bei Swartout et al.). Die beiden Bedeutungen sind jeweils mit ‚<' verbundenen, unterschiedlichen lexikalischen Wörtern differenziert. Jede

[5] Dieses Beispiel ist in folgenden Websites wiedergegeben:
http://www.aifb.uni-karlsruhe.de/Lehrangebot/Sommer1999/IWM/IWM/kap3_4.pdf
http//www.arches.uga.edu/~abhijitp/SemWeb/LargeOntology.ppt [University of Georgia]

[6] Es ist ohne weiteres möglich vom ISI die Lizenz für die Benutzung des Browsers zu erhalten. Ich verzichtete allerdings für diese Arbeit auf die Installation des Systems.

Einheit ist mit ‚||' eingegrenzt. Eine nächste Anfrage (sense 1) verlangt nach der *Is*-a-Hierarchie für Bedeutung 1 mit dem Term „superclass". In der Antwort folgen fünf Einheiten (davon die erste gemäss der Anfrage) in Klein- und vier in Grossbuchstaben. Eine Einheit in Grossbuchstaben, die verschiedene Wörter beinhaltet, ist mit Bindestrichen zusammengehalten. Die nächste Anfrage (sense 2) ergibt fünf Einheiten in Klein- und fünf in Grossbuchstaben. Beide Bedeutungen münden in „OB-THING", also in das oberste Konzept der *Ontology Base*.

Tabelle 2: Beispiel 'strut' in SENSUS vermittelt von Whitney

0	USER> (sensus-ask '(lex "strut" noun ? ?)) ((1 |strut<gait|) (2 |strut<brace|))
sense 1	USER> (superclass? '|strut<gait|) (|strut<gait| |gait<walk| |walking,walk| |locomotion<move| |change of location,move| MOTION-PROCESS NONDIRECTED-ACTION MATERIAL-PROCESS PROCESS OB-THING)
sense 2	USER> (superclass? '|strut<brace|) (|strut<brace| |bracing,brace| |structural member| |support<supporting structure| |supporting structure| |structure<artifact| |artifact| INANIMATE-OBJECT SEPARABLE-ENTITY NON-CONSCIOUS-THING OB-THING)

Es gibt grundsätzlich zwei Unterschiede zwischen Swartout et al.s Beispiel für ‚strut' in SENSUS und Withneys Resultate. Tabelle 3 zeigt eine Gegenüberstellung mit nummerierten Hierarchie-Ebenen. Bis zur Ebene 5 stimmen die lexikalischen Konzepte überein. Die Ebenen 6 und 7 bei Swartout et al. sind im Resultat von Whitney nicht vorhanden. Wichtig ist jedoch, dass in Swartout et al.s Beispiel der Übergang in die *Upper Level*, die *Ontology Base* (OB), also in die Interlingua nicht markiert ist. Diese OB-Konzepte sind in Whitneys Resultat mit Grossbuchstaben von den unteren Ebenen unterschieden. Im Kapitel 4 unten werden wir sehen, dass diese Markierung für Interlingua-Ausdrücke charakteristisch ist.

Der oberste Begriff ‚entity' der Hierarchie in Tabelle 1 lässt vermuten, dass es sich bei diesem Beispiel um Ausgaben aus dem WordNet handelt. Zum Vergleich gibt Tabelle 4 eine Aufstellung der Hyperonyme von 'strut' im WordNet. Das System gibt für die syntaktische

Tabelle 3: Vergleich 'strut' in SENSUS gemäss Swartout et al. und Whitney

Swartout et al.s Beispiel ‚strut' in SENSUS	Whitneys Resultat von ‚strut' in SENSUS
1. "strut/brace" | strut 2. "brace, bracing" | brace, bracing 3. "structural member" | structural_member 4. "support/supporting structure" | support 5. "supporting structure" | supporting_structure 6. "construction, structure" | structure, construction 7. "instrumentality/artefact" | instrumentality 8. "artifact" | artifact, article, artefact 9. "inanimate object" | object, inanimate_object, physical_object, thing 10. entity	1. (|strut<brace| 2. |bracing,brace| 3. |structural member| 4. |support<supporting structure| 5. |supporting structure| 6. |structure<artifact| 7. |artifact| 8. INANIMATE-OBJECT 9. SEPARABLE-ENTITY 10. NON-CONSCIOUS-THING 11. OB-THING)

Tabelle 4: 'strut' im WordNet

Sense 1	strut, prance, swagger -- (a proud stiff pompous gait) => gait -- (a person's manner of walking) => walk, walking -- (the act of traveling by foot; "walking is a healthy form of exercise") => locomotion, travel -- (self-propelled movement) => motion, movement, move – (…)[7] => change -- (…) => action -- (…) => act, human action, human activity -- (…)
Sense 2	strut -- (brace consisting of a bar or rod used to resist longitudinal compression) => brace, bracing -- (a structural member used to stiffen a framework) => structural member -- (support that is a constituent part of any structure or building) => support -- (…) => device -- (…) => instrumentality, instrumentation -- (…) => artifact, artefact -- (…) => object, physical object -- (…) => entity, physical thing -- (…) => whole, whole thing, unit -- (…) => object, physical object -- (…) => entity, physical thing -- (…) => strengthener, reinforcement -- (…) => device -- (…) => instrumentality, instrumentation -- (…) => artifact, artefact -- (…) => object, physical object -- (…) => entity, physical thing -- (…) => whole, whole thing, unit -- (…) => object, physical object -- (…) => entity, physical thing -- (…)

[7] Die semantischen Umschreibungen in Klammern habe ich der Einfachheit der Darstellung halber zum Teil ausgelassen.

Kategorie Nomen zwei Bedeutungen, die denjenigen von Tabelle 2 entsprechen. Swartout et al.s Darstellung (Tabelle 1) unterscheidet sich aber nicht nur in der Menge der Bedeutungen. Die erste *Is-a*-Beziehung von Bedeutung 1 - „manner of walking“ - kommt weder in Whitneys Resultat noch im WordNet vor. Im Letzteren ist dieser Ausdruck Teil der in Klammern natürlichsprachig umschriebenen (also nicht konzeptuellen) Bedeutung „a person’s manner of walking“.

Im Weiteren werde ich nicht mehr auf die Darstellung in Tabelle 1 eingehen, weil sie wenig Sinn macht. Anstelle davon werde ich Whitneys Resultate mit denjenigen von WordNet näher vergleichen.

3.3 WordNet und SENSUS: ein Vergleich

WordNet ist ein Thesaurus der englischen Sprache, der mit syntaktischen Kategorien und semantischen Relationen zwischen den Wortbedeutungen einen beachtlichen Wortschatz online verfügbar macht. Unter den semantischen Relationen entspricht die Hyperonymie der *Is-a*-Beziehung, welche auf höheren Ebenen in abstraktere Begriffe überführt und so in eine konzeptuelle Repräsentation. Der Vergleich in Tabelle 5 zwischen ‚strut’ im WordNet und in SENSUS soll vor allem die Struktur der *Ontology Base* hervorheben.

Tabelle 5 : Vergleich von 'strut' in WordNet und SENSUS

	WordNet	**SENSUS**
Sense 1	strut, prance, swagger gait walk, walking locomotion, travel motion, movement, move change action act, human action, human activity	|strut<gait| |gait<walk| |walking,walk| |locomotion<move| |change of location,move| MOTION-PROCESS NONDIRECTED-ACTION MATERIAL-PROCESS PROCESS OB-THING
Sense 2	strut brace, bracing structural member support device instrumentality, instrumentation artifact, artefact object, physical object entity, physical thing	(|strut<brace| |bracing,brace| |structural member| |support<supporting structure| |supporting structure| |structure<artifact| |artifact| INANIMATE-OBJECT SEPARABLE-ENTITY NON-CONSCIOUS-THING OB-THING)

Die Darstellung der Beispiele aus dem WordNet in der Tabelle 5 ist stark vereinfacht. Bei der Bedeutung 2 führen gemäss Tabelle 4 vier Hyperonym-Äste zum obersten Begriff. Dieser ist für diese Bedeutung „entity, physical thing“. Im Gegensatz dazu endet ‚strut’ mit der Bedeutung 1 im WordNet bei „act, human action, human activity“. Dies zeigt, dass es bei WordNet nicht nur einen einzelnen obersten Knoten gibt wie bei SENSUS. Miller (1993) begründet dies folgendermassen: „[T]hese abstract generic concepts carry little semantic information; it is doubtful that people could even agree on appropriate words to express them“. Auf die Schwierigkeit der Wortwahl für die abstrakten Konzepte wird im nächsten Kapitel eingegangen. Wichtig an dieser Stelle ist festzuhalten, dass WordNet und SENSUS unterschiedliche Theorien verfolgen. Sie sind dementsprechend auch für ganz verschiedene Zwecke bestimmt. In WordNet werden die Konzepte mit Synonym Sets voneinander unterschieden. Miller erklärt (1993): „These synonym sets (synsets) do not explain what the concepts are; they merely signify that the concepts exist. People who know English are assumed to have already acquired the concepts, and are expected to recognize them from the words listed in the synset.” WordNet ist somit auf die Anwendung von Personen ausgerichtet, und zwar Personen die englisch sprechen. Das heisst, das System hat einen niedrigen Formalisierungsgrad und ist sprachabhängig. SENSUS hingegen, welches die Ontologie für

eine Übersetzungsmaschine liefert, muss maschinell verarbeitbar sein, was einen relativ hohen Formalisierungsgrad voraussetzt. Dazu kommt, dass die *Ontology Base*, also die abstraktesten, obersten Konzepte in der Hierarchie, sprachunabhängig sein soll. Der Vergleich in Tabelle 5 zwischen WordNet ohne diesen Anspruch und SENSUS macht diesen Sachverhalt nicht klar. Es lässt sich nicht nachvollziehen warum „NONDIRECTED-ACTION“ bzw. „INANIMATE-OBJECT“ in SENSUS sprachunabhängiger sein soll als „action“ bzw. „physical object“ im WordNet. Dies soll im nächsten Kapitel näher durchsucht werden.

4. Die Sprachunabhängigkeit der Interlingua

In der maschinellen Übersetzung wird neben der Transfer-Technik das Interlingua-Verfahren in vielen Forschungsprojekten angewendet, so auch in SENSUS. Text in der Ausgangsprache wird analysiert, in eine sprachunabhängige Repräsentation (Interlingua) gebracht, woraus der Text in der Zielsprache generiert wird:

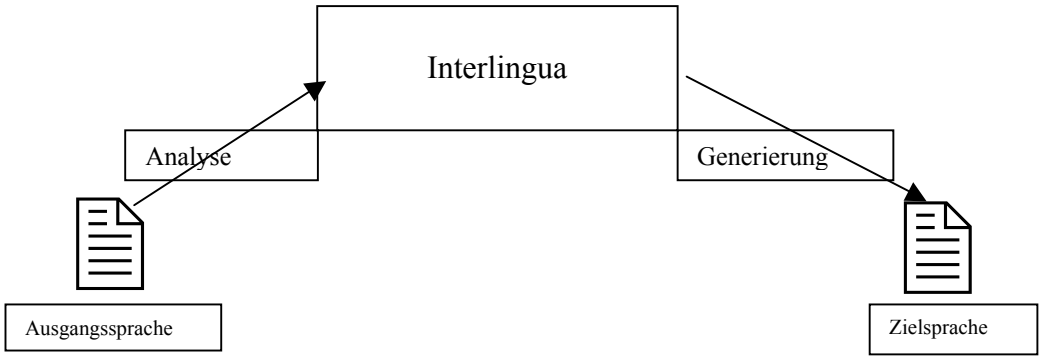

Abbildung 2: Interlingua

Die Frage ist nun, wie diese sprachunabhängige Repräsentation erzeugt wird. Grundsätzlich gibt es zwei verschiedene Ansätze. Zum einen bietet sich die formale Logik an. Diese ist allerdings traditionellerweise weniger für die Textgenerierung geeignet als für die Analyse. Zum anderen hat sich in der Computerlinguistik die konzeptuelle Repräsentation entwickelt.

Mattiessen und Bateman (1991), die sich beim Erstellen des PENMAN Upper Models auf Hallidays *Systemic Grammar* gestützt hatten, meinen dazu:

> The two techniques may often be inter-translatable [...], but they have been used for different purposes: [...] semantic nets are often used to represent the knowledge potential whereas predicate logic is often used to represent instantial knowledge. (233)

Auch Siegel (1996) schreibt in ihrer Darstellung des Interlingua-Verfahrens, dass Logik für die Textgenerierung zu komplex ist .

Die Interlingua von SENSUS ist also eine konzeptuelle Repräsentation. Um einen etwas grösseren Einblick in diese zu erhalten als es das Beispiel ‚strut' oben gibt, sind im Anhang die verschiedenen Bedeutungen von ‚bank' und ‚love' aufgestellt, wie sie in SENSUS dargestellt sind. Die sog. Interlingua-Ausdrücke sind in Abbildung 3 alphabetisch aufgelistet.

ABSTRACTION	LIKING	ORGANIZATION
ABSTRACT-OBJECT	MATERIAL-PROCESS	PERSON
AMBIENT-PROCESS	MENTAL-INACTIVE	PHYSICAL-OBJECT
ANIMAL	MENTAL-PROCESS	PROCESS
ANIMATE-OBJECT	NAMED-OBJECT	REACTION
CONSCIOUS-BEING	NON-CONSCIOUS-THING	SEPARABLE-ENTITY
DIRECTED-ACTION	NONDIRECTED-ACTION	SOCIAL-OBJECT
DISPOSITIVE-MATERIAL-ACTION	OBJECT	VERTEBRATE
INANIMATE-OBJECT	OB-THING	
	ORDERED-OBJECT	

Abbildung 3: Auswahl von Interlingua-Ausdrücken in SENSUS

Die Konzepte sind durch ihre Darstellung in Grossbuchstaben gekennzeichnet. Wie aber werden die Begriffe gewählt? Laut Hovy und Nirenburg (1992) sind sie semantisch und nicht syntaktisch gewählt (z.B. münden Handlungen in PROCESS und nicht in VERB). Weiter müssen die Begriffe so gewählt werden, dass sie einerseits vom Parser und Generator bewältigt werden können und andererseits genug Unterscheidungskraft aufbringen um Ambiguitäten usw. zu steuern. Die Distanz zur Sprache muss also minimal gehalten werden (Hovy/Nirenburg;1992). Einen Hinweis darüber, wie die Terme gewählt werden, gibt auch folgende Definition für Ontologie, die in der Informatik verbreitet ist:

> *An ontology is an explicit specification of a **shared** conceptualization.*

Laut Hovy und Nirenburg (1992) müsste der Konsens in der Konzeptualisierung, der das „shared“ voraussetzt, von der Theorie kommen.

Wie aus der Auflistung in Abbildung 3 erkennbar ist, beinhalten die Ausdrücke keine semantischen Informationen, wie z.B. die im WordNet. Sie sind terminologisch und die Bedeutung wird durch die Taxonomie, die bestimmte Anordnung und Vernetzung der Ausdrücke, repräsentiert. Auf die wird hier nicht weiter eingegangen. Was hier untersucht werden soll, ist die Frage, wie diese Terme zu ihrer Sprachunabhängigkeit kommen. Wenn wir nämlich die letzten Ausdrücke der *Ontology Body* (in Kleinbuchstaben) mit denen der *Ontology Base* (in Grossbuchstaben) vergleichen, wie bei den folgenden Beispielen,

- |natural object| vs. INANIMATE-OBJECT
- |organization<social group| vs. ORGANIZATION
- |psychological feature| vs. ABSTRACT-OBJECT

dann erkennt man doch, dass es sich bei den oberen Ebenen der WordNet-Ausdrücke schon um Konzepte handelt. Matthiessen und Bateman (1991) erklären, was es mit dieser Steigerung in der Abstraktion auf sich hat:

> We can [...] relate this general increase in abstraction to a central area of concern in text generation development: i.e., the need to provide *control* of the language generation resources. The more abstract and ‘semantic’ the level of interaction between a text generation system and its supporting environment becomes, the more manageable those resources are. (231)

Die Sprachunabhängigkeit hat also nicht unbedingt mit der Abstraktion zu tun. Sowohl das Penman Upper Model, das für die Generierung englischer Texte erstellt wurde, wie auch das WordNet sind sprachabhängig.

Auf die Frage warum z.B. das Konzept „NONDIRECTED-ACTION” sprachunabhängig sein soll, hat Hovy in einer unveröffentlichten Stellungnahme zwischen zwei Auffassungen, was man unter Interlingua verstehen kann, unterschieden:

> (1) An Interlingua is a set of totally language-independent concepts, shared potentially by all people in the world
>
> (2) An Interlingua is a set of concepts that generalize linguistic constructs to the point where cross-language correlations become apparent.

Dazu meint er:

> I think (1) is the traditional conception. And I think (1) is impossible to get in practise: how will you prove it is 100% language-independent? What does "language-independent" mean? So I think (2) is what we build when we really try to make an Interlingua.

Richtlinien, wie eine solche Interlingua erstellt werden soll, gibt er in seiner Arbeit mit Sergei Nirenburg “Approximating an Interlingua in a Principled Way” (1992). Der Titel lässt darauf schliessen, dass es dabei zu eher intuitiven Ad-hoc-Formationen kommen kann. Das verhindert dann oft die nötige Distanz des Programmierers zur eigenen Sprache, die es zur Erstellung einer sprachunabhängigen Ontologie braucht. Die Arbeit zeigt aber auch auf, dass der Anspruch der Sprachunabhängigkeit ein idealistischer ist. Hovy und Nirenburg plädieren dafür, die Debatte darüber den Philosophen zu überlassen. Wie oben schon erwähnt (S. 12), muss die Distanz zur Sprache minimal gehalten werden, damit die Interlingua von Parser und Generator bewältigt werden können. Der praxisorientierte Ansatz dieser Arbeit befürwortet anstelle von Sprachunabhängigkeit eine **sprachneutrale** Lösung. Diese Sprachneutralität wird erarbeitet, indem der Programmierer von einer Ausgangssprache ausgeht. Von dieser Sprache her werden linguistische Abstraktionen erarbeitet, die für die Analyse und Generierung relevant sind. Für Englisch müssten z.B. Ausdrücke wie

- UNCOUNTABLE-OBJECT, und
- COUNTABLE-OBJECT

unter OBJECT angeordnet werden um die Nomen in den Griff zu bekommen. Nachdem diese Ontologie, ausgerichtet auf die Ausgangssprache aufgestellt ist, geht es darum eine zweite Sprache hinzuzunehmen. Dann kann es vorkommen, dass weitere Ausdrücke hinzugefügt werden müssen, um Unterscheidungen, die in der neuen Sprache wichtig sind, die jedoch nicht in der Ausgangssprache vorkommen, handhaben zu können. Die obigen Ausdrücke müssten dann beispielsweise auf

- UNCOUNTABLE-OBJECT1
- COUNTABLE-OBJECT1
- UNCOUNTABLE-OBJECT2
- COUNTABLE-OBJECT 2

erweitert werden.

Die Sprachneutralität kann auf diese Weise mit zusätzlichen Sprachen grundsätzlich weiter entwickelt werden. Die Autoren glauben, „that a true language-neutral ontology can only be approached asymptotically […]“. Der in der Forschungsliteratur immer erwähnte Vorteil der Interlingua gegenüber den Transferregeln, dass beim Hinzufügen von neuen Sprachen nur der

Parser und der Generator neu erstellt werden müssen, weil die Interlingua sprachunabhängig ist, ist somit in Frage gestellt. Wie Siegel darstellt „sind die meisten maschinellen Übersetzungssysteme mit einer Interlingua für zwei, höchstens drei Sprachen entwickelt worden“ (53). So auch im Fall vom PANGLOSS-System, das ursprünglich für Spanisch und Englisch entwickelt und später für Japanisch erweitert worden ist.

5. Zusammenfassung

SENSUS ist einerseits ein Beispiel für eine gross angelegte Wissensrepräsentation für die maschinelle Übersetzung. Andererseits lässt sich damit darstellen, wie Projekte mit unterschiedlichen zugrunde liegenden Theorien erfolgreich vereint werden können. Die Untersuchungen dieser Arbeit anhand von Beispielen und Vergleichen mit dem WordNet illustrieren, was unter den Begriffen konzeptuelle Wissensrepräsentation, *Ontology Base* und Interlingua gemeint sein kann. Die unfassbare Sprachunabhängigkeit, die allgemein bei einer Interlingua vorausgesetzt wird, ist mit einem pragmatischen Ansatz zur realistischeren Sprachneutralität neu definiert worden.

6. Bibliografie

Egger, Carole. 2002. "Cyc - Computerizing 'common sense'". Handout Seminar: Aspekte der Wissensrepräsentation in der Computerlinguistik. Dr. Kai-Uwe Carstensen. Zürich: SS 2002.

Hovy, Eduard; Kevin Knight. 1993. "Motivating Shared Knowledge Resources: An Example from the Pangloss Collaboration." <www.isi.edu/natural-language/resources/shared-ontologies.ps> [März 2002].

Hovy, Eduard; Sergei Nirenburg. 1992. "Approximating an Interlingua in a Principled Way." <www.isi.edu/natural-language/mt/jp-lex-ontol-bridge.ps].

Knight, Kevin; Steve K. Luk. 1994. "Building a Large-Scale Knowledge Base for Machine Translation." <http://www.isi.edu/natural-language/people/knight.html> [März 2002].

Matthiessen, Christian; John A. Bateman. 1991. *Text Generation and Systemic-Functional Linguistics: Experiences from English and Japanese*. Communication in Artificial Intelligence Series. Eds. Robin P. Fawcett; Erich H. Steiner. London: Pinter Publishers.

Matthiessen, Christian; Michael A.K. Halliday. 1997. "Systemic functional Grammar: A First Step into the Theory." <http://minerva.ling.mq.edu.au/Resources/VirtuallLibrary/Publications/sfg_firststep/SFG%20intro%20New.html> [Juni 2002].

Miller, George A. 1993. "Nouns in WordNet: A Lexical Inheritance System." <ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.pdf> [Juli 2002].

Siegel, Melanie. 1996. *Die maschinelle Übersetzung aufgabenorientierter japanisch-deutscher Dialoge. Lösungen für Translation Mismatches*. Diss. Linguistik und Literaturwissenschaft der Universität Bielefeld. <http://www.dfki.de/~siegel/diss.ps.gz>

Swartout, Bill, et al. 1996. "Toward Distributed Use of Large-Scale Ontologies." <http://ksi.cpsc.ucalgary.ca/KAW/KAW96/swartout/Banff_96_final_2.html> [Juni 2002].