Ein elektronisches Lexikon im OLIF-Format für die Erzählanalyse

Marc Luder
Simon Clematide
Universität Zürich

Bernhard Distl
Eidgenössische Technische Hochschule Zürich

We present the JAKOB lexicon, a semantically rich German lexical resource, and its migration to the OLIF format (Open Lexicon Interchange Format). This lexicon is part of a web-based text and narrative analysis application. The JAKOB narrative analysis is a qualitative research tool to systematically analyze patient's narratives. It conceptualizes narratives as dramaturgically-constructed linguistic productions and interprets them with regard to the un-conscious conflicts of the narrator contained there in. In this process, narratives are extracted from transcripts, then a linguistic analysis is performed, and after that the vocabulary is encoded according to predetermined psychological conceptual categories incorporated in the JAKOB lexicon. The need for the proper treatment of multi-word units in the JAKOB project made OLIF a reasonable target format. OLIF is word-sense oriented and allows a broad linguistic description-syntactical, morphological, and semantic-for each lexical entry. The OLIF data categories and attributes are well defined in the case of German but it turned out that the data-category labels in OLIF aren't specified very clearly sometimes. In addition to that, there are few resources that prove their practical use. In a corporate project, the lexicon was half-automatically reassessed and finally migrated. OLIF is an open XML-based standard for structuring lexical data and provides a rich choice of linguistic categories and predefined values. Multi-word entries represent an essential improvement for the JAKOB application. The narrative texts represent spoken language; therefore the utterances aren't well formed, in most cases, and not eligible for a standard syntactic analysis. We use a construction-grammar approach to gather the sense of multi-word expressions in the text and to match them to lexicon entries with their corresponding conceptual categories. We use multi-word entries as containers for constructions-form-meaning units-like idioms and collocations. Further investigations will show to which extent more general constructions can be lexicalized. Our project goal is to improve precision in coding the JAKOB narratives. We decided to create an OLIF database, using the XML schema as the basis for the database structure. Thus, import and export of OLIF data is straight-forward. The implementation is object-oriented and solely based on open source software using PHP / MySQL.

Einleitung

Im vorliegenden Beitrag wird die Implementierung eines elektronischen Lexikons in deutscher Sprache im OLIF-Format (OLIF – Open Lexicon Interchange Format)[1] vorgestellt. Dieses Lexikon ist Teil einer internetbasierten Text- und Erzählanalyseapplikation, die am Psychologischen Institut der Universität Zürich (Lehrstuhl Klinische Psychologie, Psychotherapie und Psychoanalyse) entwickelt wurde (Boothe 2004; Boothe, Grimmer, Luder, Luif, Neukom & Spiegel 2002). Die *Erzählanalyse JAKOB* ermöglicht die Erfassung, Kodierung und Auswertung von *Alltagserzählungen* aus Psychotherapiegesprächen im Dienste klinischer Diagnostik. Das Lexikon im OLIF-Format stellt dafür das syntaktische, semantische und pragmatische Hintergrundwissen zur Verfügung und enthält neben den lexikographischen Angaben zusätzlich applikationsspezifische Daten für die Analyse und Kodierung der Texte.

[1] Siehe *http://www.olif.net*.

OLIF ist ein offener Standard für die Strukturierung lexikalischer Daten und stellt ein reichhaltiges Arsenal an linguistischen Kategorien zur Verfügung, die die computergestützte Analyse der Erzähltexte ermöglichen. Die vorgegebenen Datenkategorien wurden weitgehend übernommen.

Die Erzählanalyse JAKOB

Die Erzählanalyse JAKOB ist ein qualitatives Untersuchungsinstrument für *Alltagserzählungen*. Diese offenbaren in kompakter Form emotional bedeutsames Selbst- und Beziehungserleben und sind dazu prädestiniert, konflikthaftes Material zu transportieren, das in der Dramaturgie der Erzählungen zur Darstellung kommt. Die Erzählanalyse JAKOB ermöglicht narrative Einzelfallanalysen, die —ausgehend vom Erleben und von der subjektiven Sicht des Patienten— einen Forschungsbeitrag zur Konflikt-, Beziehungs- und Prozessdiagnostik auf dem Hintergrund von psychodynamischen und psychoanalytischen Theorien leisten.

Die Erzählanalyse JAKOB wird hauptsächlich für die Analyse von Erzähltexten aus *Psychotherapiegesprächen* verwendet. Die Therapiestunden werden mit dem Einverständnis der Patienten aufgezeichnet und für die Verwendung in der Erzählanalyse transkribiert. Aus den Transkripten werden mit Hilfe eines Regelkataloges die Erzählungen extrahiert, der Text wird mit einem vorgegebenen, auf psychoanalytischen Konzepten beruhenden Kategoriensystem kodiert. Die Annotation der Texte und die Verwaltung der Codes erfolgte ursprünglich manuell, wurde aber im Laufe der Zeit in die webbasierte Computeranwendung AutoJAKOB[2] (Luder 1999) übergeführt; die Annotation der Texte erfolgt nun mit Computerunterstützung, das Kodiersystem ist im elektronischen JAKOB-Lexikon abgebildet. Die linguistische Analyse erfolgt in mehreren Schritten und verarbeitet morphologische, syntaktische und semantische Merkmale der Textsegmente mit dem Ziel, den Bedeutungsgehalt der Segmente möglichst genau zu bestimmen, so dass im Lexikon die passende Kodierung gefunden wird.

In der weiteren Auswertung versuchen wir einerseits, die besondere Dramaturgie der Erzählung und ihre Vermittlung durch den Erzähler durch präzise lexikalische Analyse zu ergründen, andererseits Wunsch-, Angst- und Abwehrbewegungen sichtbar zu machen und dadurch zentrale Konflikte des Erzählers zu erschliessen.

Implementierung des Lexikons im OLIF-Format

In Rahmen eines von der Universität Zürich unterstützten interdisziplinären Projektes wurden im Laufe des Jahres 2007 eine Neuimplementierung des bestehenden Lexikons und die Migration zum OLIF-Format vorgenommen. Es folgt eine kurze Übersicht über die technischen Aspekte der Migration.

Zur Repräsentation eines Lexikoneintrags in OLIF wird ein XML-Format benutzt, dessen Struktur durch das OLIF2-Konsortium festgelegt wurde. In der OLIF-Version 2 liegt diese Strukturbeschreibung als XML-Dokumenttypdefinition vor, für die Version 2.1 (und aktuell 2.1.1) wurde die Strukturdefinition auf ein XML(W3C)-Schema umgestellt[3].

Unsere Implementierung des Lexikons sollte einerseits Lexikoneinträge im OLIF-Format importieren und exportieren können, um später den Datenaustausch mit weiteren lexikalischen Ressourcen zu ermöglichen, andererseits sollte das Lexikon möglichst gut auf die Anforderungen der JAKOB-Kodierung vorbereitet sein. Um die Lexikonabfrage nach diversen Kriterien möglichst flexibel und schnell abwickeln zu können, haben wir uns dazu entschlossen, für die Speicherung der Lexikondaten eine Datenbank einzusetzen. Da die bereits bestehende JAKOB-Applikation MySQL und PHP einsetzt, wurde diese Technologie beibehalten. Damit setzen wir auf eine Lösung, die ausschliesslich mit Open Source-Komponenten arbeitet.

[2] Siehe *http://www.jakob.uzh.ch*.

[3] *http://www.olif.net/documentation.htm*.

Die Implementierung gliedert sich in drei Teile. Der erste Teil ist die Datenbankstruktur, die auf Basis des OLIF 2.1.1 XML Schemas entworfen wurde und die Speicherung eines gesamten OLIF Eintrags ermöglicht. Der zweite Teil ist die objektorientierte Implementierung der OLIF-Datenstruktur. Dabei wurde darauf geachtet, möglichst viel Funktionalität generisch zu gestalten. Kern dieser Komponente ist die Klasse „OLIFObject“ von der sich ein OLIF-entry und alle anderen Subkomponenten ableiten. Der dritte Teil besteht aus einer objektorientierten Schnittstelle zwischen den OLIF-Objekten und der Datenbank, wodurch die OLIF Objekte unabhängig von der Datenbankstruktur sind.

Das Datenbankdesign sowie die PHP-Klassen für die OLIF-Objekte sollen unter einer freien Lizenz veröffentlich werden, damit das Lexikon auch in andern Projekten verwendet werden kann.

Migration bestehender Daten ins OLIF-Format

Die Orientierung von OLIF am *Wortsinn* (McCormick, Lieske & Culum 2004) entspricht den Bedürfnissen einer konzeptuellen Klassifikation, welche das JAKOB-Lexikon kodiert. Anders als bei WordNet-artigen (Miller, Beckwidth, Fellbaum, Gross & Miller, 1993), semantisch-lexikalischen Ressourcen steht bei uns nicht eine durchgängige paradigmatische Organisation der lexikalischen Beziehungen (Hyperonymie, Meronymie, etc.) im Vordergrund, sondern die konsistente Annotation von Syntagmen wie bei FrameNET (Baker, Fillmore & Lowe 1998). Weiter erlaubt OLIF eine breite linguistische Beschreibung und stellt einen brauchbaren Kompromiss dar zwischen konkreten Applikationsbedürfnissen und komplexen Standardisierungsanstrengungen, wie sie etwa vom ISO/TC 37 vorangetrieben werden. Die linguistische Abdeckung für Deutsch mit inhaltlich interpretierten und vorgefertigten Ausprägungen von Datenkategorien vereinfacht den Import der bestehenden Daten. Das Institut für Computerlinguistik stellt morphologische Ressourcen in OLIF-kompatibler Auszeichnung der Flexionsklassen bereit. Weiter hat uns die freie Verfügbarkeit, die Flexibilität und Anpassbarkeit von OLIF überzeugt.

Die OLIF-Migration bedeutet auch Schwierigkeiten. Aus einer einfachen Relation mit einer Reihe von Attributen entstand ein komplexes Datenmodell. Die inhaltliche Interpretation der Kürzel von Datenkategorien ist nicht immer so klar spezifiziert, insbesondere weil OLIF mehrere Kodierungstraditionen nebeneinander anbietet. Trotz der freien Verfügbarkeit von OLIF selbst sind kaum allgemeine Referenzdaten zugänglich.

Das ursprüngliche Format des verb-lastigen Lexikons mit 5’000 Einträgen (2688 Verben, 1522 Nomen, 637 Adjektive, 200 Übrige) umfasste folgende linguistischen Datenkategorien: Grundform (nur Einzelwörter); Ergänzung (obligatorisch mitauftretende Wörter oder Wortgruppen); Wortart; Genus, semantische Kategorie (Substantive); Subkategorisierung, Hilfsverb, Reflexivität, Partizip Perfekt (Verben); Sinn (unrestringierte Klärungen); JAKOB-Code (konzeptuelle Kodierung).

Vorbereitend für die Konversion waren automatische und aufbauend darauf intellektuelle lexikalische Verifikationen der Einträge notwendig, da deren Konsistenz und Vollständigkeit aus der Entwicklung dieser Ressource heraus nur bedingt gegeben war. In der ursprünglichen Datenkategorie Ergänzung fanden sich (a) Präpositionen „[antreten] gegen”[4], „[Fernsehen] im“, (b) Präpositionalphrasen „[kehren] unter den Tisch”, (c) Objekte „[haben] engen Kontakt“, (d) adverbiale Adjektive „[machen] fertig“, (e) Adverbien „[sein] ziemlich unten“ sowie Information, welche eigentlich im Subkategorisierungsrahmen zu kodieren wäre „[melden] sich“ (Reflexivität) oder „[vorsprechen] etwas“ (Transitivität). Die Wichtigkeit zur systematischen Erfassung und Unterstützung von Mehrwortausdrücken für eine optimale Anwendung des JAKOB-Lexikon zur halbautomatischen Annotation war schon früher erkannt worden. OLIF stellt dafür eine brauchbare Strukturierung zur Verfügung: Mit der Datenkategorie „prep“ können Präpositionen wie in (a) abgelegt werden. Eine Kasusangabe (z. B. bei Dativ/Akkusativ-Mehrdeutigkeiten) oder alternative Präpositionen sind jedoch nicht

[4] Die Grundform ist zwischen eckigen Klammern notiert.

vorgesehen. Die Fälle (b-e) werden durch echte Mehrworteinträge gelöst. Auf Grund der Regeln zur kanonischen Form ergibt sich aus einem Eintrag „geraten“ mit der Ergänzung „Rand und Band“ neu „geraten ausser Rand und Band“.

Lexikonsysteme für Mehrwortausdrücke und Redewendungen wie der PhraseManager (Pedrazzini, 1994) enthielten bereits eine Beschreibungssprache, welche optionale Erweiterungen, syntaktische Modifizierbarkeit, morphologische Eigenschaften und Einschränkungen, Platzhalter für syntaktische Konstrukte, strikte Invarianz und Kongruenzbedingungen kodieren konnte. Ob sich der Aufwand einer detaillierteren linguistischen Beschreibung lohnt und auch hinreichend zuverlässig von Personen gemacht werden kann, denen die linguistische Beschreibung nicht das primäre Anliegen ist, ist offen.

Die Relevanz von Verbgefügen, Kollokationen und idiomatischen Wendungen (Fellbaum 2007) ist im (schweizerischen) Textkorpus sicher hoch: Verbgefüge mit direkten Objekten („und habe fast Krach bekommen mit ihr”), Präpositionalphrasen („es kam mir jetzt gerade diesen Moment in den Sinn”), Verbkombinationen („ich musste alles sausen lassen”) und Kollokationen („dann würde ich mich also dumm und dämlich hintersinnen”) werden häufig verwendet.

Für die *Subkategorisierung der Verben* brauchte das bestehende Lexikon die numerischen Kodes von Wahrig (1997)[5]. OLIF selbst hat dafür den Slot-Grammar-Ansatz für Englisch adaptiert, wobei bei Infinitivkonstruktionen die Abdeckung für Deutsch nicht befriedigt. Auch sind keine Verben mit mehr als einer fixierten Präposition vorgesehen („mit etwas an jmdn. herantreten“) —wie auch nur ein einziges „prep“-Element zur Deklaration von Präpositionen vorgesehen ist. Um Informationsverluste durch eine Konversion der Wahrig-Kodes zu vermeiden, haben wir die in OLIF vorgesehene Ersetzbarkeit von Datenkategorien benutzt. Die Abstützung auf das ausgereifte Standardwerk Wahrig vereinfacht und vereinheitlicht die lexikographische Arbeit.

Perspektiven für die Erzählanalyse JAKOB

Die neue OLIF-Lexikonstruktur erlaubt *Mehrworteinträge* und eröffnet deshalb für die JAKOB-Anwendung neue Perspektiven. Ein wichtiger Teilschritt der Erzählanalyse JAKOB besteht in der Kodierung des verwendeten Vokabulars mit dem a priori festgelegten konzeptuellen Kategoriensystem und in der genauen Untersuchung des verwendeten sprachlichen Materials sowohl in Bezug auf Inhalt als auch auf Form. Angewendet werden verschiedene Methoden zur Untersuchung von für die Erzählungen typischen individuellen, aber auch lexikalisierten Sprachstrukturen. Da sich die klassischen linguistischen Kategorien der Grammatik, Syntax, Semantik und Pragmatik nur bedingt für die Verarbeitung von gesprochener Sprache eignen, wie sie in den Transkripten und Erzählungen vorliegt, sind wir bestrebt, Erkenntnisse aus Gesprächsforschung und interaktionaler Linguistik für unseren Forschungsansatz zu nutzen. Gemäss diesen Forschungsansätzen entsteht die „Bedeutung“ des Gesprochenen erst in der Interaktion und im Gesprächskontext (*Bedeutungskonstitution*, Deppermann 2006a), d.h. die Einzelwörter haben in der Regel keine kontextfreie, feststehende Bedeutung, sondern die Bedeutung wird grossenteils in mehr oder weniger idiomatischen und mehr oder weniger festen Wortverbindungen konstituiert.

Ansätze der *Konstruktionsgrammatik* (*construction grammar*) (Fillmore, Kay & O’Connor 2003; Deppermann 2006b) kommen diesen Anforderungen entgegen, indem sie postulieren, dass gesprochene Äusserungen als eine Reihe von Konstruktionen aufgefasst werden können. Konstruktionen sind Form-Bedeutungseinheiten (form-meaning units) (Croft & Cruse 2004) mit syntaktischen, semantischen und pragmatischen Merkmalen und bilden einen umfassenden Beschreibungsrahmen für sprachliches Wissen (Deppermann 2006b). Die linguistische Analyse stützt sich (nicht nur) auf das Erkennen von möglichen wohlgeformten Phrasen und Teilphrasen, sondern ebenso auf das Erkennen von Konstruktionen als syntaktische Ganzheiten (Deppermann 2006b), die in den Erzähltexten verwendet werden. Konstruktionen sind z.B. Phraseologismen

[5] *Der Kleine Wahrig.*

(Idiome, Metaphern und Kollokationen), die teilweise lexikalisiert werden können, teilweise aber auch sprecherspezifisch sind.

Für die Erweiterung und zukünftige Verwendung des OLIF-Lexikons gilt es nun zu klären, welche der neu erfassten Mehrwortstrukturen als Konstruktionen im Lexikon adäquat beschrieben werden können. Die entsprechenden Lexikoneinträge (form-meaning units) müssen dafür mit den passenden formalen (Morphologie, Syntax), semantischen und pragmatischen Informationen versehen werden (semantische Rollen, Kontexthinweise, Themen, Domänen). Beispiele sind idiomatische Wendungen wie *typisch Frau*, *solches Zeug ausbeinen*, Verbgefüge wie *ein Zeugs machen*, aber auch konventionalisierte Redewendungen wie *Ich bin der Typ*…

Konstruktionen, die nicht lexikalisiert werden können, müssen mit den Kodier- und Analyseprozeduren erfasst werden. Beispiele dafür sind etwa die Verbspitzenstellung, narratives Präsens (*steige aus*, *gehe rüber*…) oder Infinitkonstruktionen (*ich sofort runter*, *kein Essen bekommen*…) (Günthner 2006).

Für die Erzählanalyse JAKOB bedeutet die Implementierung des neuen OLIF-Lexikons, dass die Kodierung der Erzähltexte mit entsprechend angepassten Prozeduren präziser durchgeführt werden kann. Beim aktuellen Stand des Projektes (Ende 2007) ist das Lexikon implementiert, die Daten aus dem ursprünglichen JAKOB-Lexikon wurden in das neue Format übertragen, die AutoJAKOB-Applikation arbeitet bereits mit OLIF-Daten. Die neu geschaffenen Möglichkeiten müssen nun durch lexikographische Bearbeitung des Lexikons und durch neue Analyse- und Kodierprozeduren in der Applikation umgesetzt werden.

Literatur

Baker, C. F.; Fillmore, C. J.; Lowe, J. B. (1998). "The Berkeley FrameNet Project". In Proceedings of COLING-ACL'98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. Montreal.

Boothe, B. (2004). *Der Patient als Erzähler in der Psychotherapie*. 2. Aufl. Giessen: Psychosozial-Verlag.

Boothe, B. *et al*. (2002). *Manual der Erzählanalyse JAKOB: Version 10/02*. Zürich: Universität Zürich, Psychologisches Institut, Klinische Psychologie I.

Croft, W.; Cruse, D. A. (2004). *Cognitive linguistics.* 3. Aufl. Cambridge: Cambridge University Press.

Deppermann, A. (2006b). „Construction Grammar - Eine Grammatik für die Interaktion?". In Deppermann, A.; Fiehler, R.; Spranz-Fogasy, T. (eds.). *Grammatik und Interaktion. Untersuchungen zum Zusammenhang von grammatischen Strukturen und Gesprächsprozessen*. Radolfzell: Verlag für Gesprächsforschung. 43-65.

Deppermann, A. (2006a). „Von der Kognition zur verbalen Interaktion: Bedeutungskonstitution im Kontext aus der Sicht der Kognitionswissenschaften und der Gesprächsforschung". In Deppermann, A.; Spranz-Fogasy, T. (eds.). *Be-deuten. Wie Bedeutung im Gespräch entsteht*. 2. Aufl. Tübingen: Stauffenburg-Verlag. 11-33.

Fellbaum, C. (ed.). (2007). *Idioms and collocations: Corpus-based linguistic and lexicographic studies*. London: Continuum.

Fillmore, C. J.; Kay, P.; O'Connor, M. C. (2003). "Regularity and idiomaticity in grammatical constructions: The case of *let alone*". In Tomasello, M. (ed.). *The new psychology of language. Cognitive and functional approaches to language structure*. Mahwah: Lawrence Erlbaum Associates. Vol. II. 243–270.

Günthner, S. (2006). „Grammatische Analysen der kommunikativen Praxis - Dichte Konstruktionen in der Interaktion". In Deppermann, A.; Fiehler, R.; Spranz-Fogasy, T. (eds.). *Grammatik und Interaktion. Untersuchungen zum Zusammenhang von grammatischen Strukturen und Gesprächsprozessen*. Radolfzell: Verlag für Gesprächsforschung. 95-122.

Luder, M. (1999). Kategorien und Codes: Auf dem Weg zu einer computergestützten Fassung der Erzählanalyse JAKOB. Unveröffentlichte Lizentiatsarbeit. Zürich: Universität Zürich.

McCormick, S. M.; Lieske, C.; Culum, A. (2004). OLIF v.2: A Flexible Language Data Standard [on-line]. *http://www.olif.net/documents/OLIF_Term_Journal.pdf*. [Access date 22.Mar. 2008].

Miller, G. A. et al. (1993). Introduction to WordNet: An On-line Lexical Database (Five Papers on WordNet) [on-line]. *http://wordnet.princeton.edu/5papers.pdf*. [Access date 22.Mar. 2008].

Pedrazzini, S. (1994). *Phrase manager: a system for phrasal and idiomatic dictionaries*. Hildesheim: Olms.

Wahrig, G. (1997). *Der kleine Wahrig: Wörterbuch der deutschen Sprache*. München: Bertelsmann.

Anhang: Beispiel eines Lexikoneintrags *sich nicht ins Bockshorn jagen lassen*

OLIF-Feld	Wert	Beschreibung
canForm	lassen ins Bockshorn jagen	Kanonische Form des Eintrags
ptOfSpeech	verb	Wortart (Head)
head	lassen	Head des Mehrwortausdrucks
verbPart		Verbpartikel
auxType	haben	Hilfsverb
language	de	Sprachversion
subjField	general	subject field, Domäne
definition	sich nicht einschüchtern lassen. Negation obligatorisch.	Definition, Bedeutung, semantische Besonderheiten.
natGender		Natürliches Geschlecht (für Nomen)
semType	ment-act	Semantischer Typ (OLIF-Werteliste)
synType	refl	Marker für Reflexivität
synFrame	570	Satzmuster nach Wahrig (2002)
prep	in	Präposition
morphStruct		Morphologische Struktur (für zusammengesetzte Verben)
geogUsage		z.B. Dialektausdrücke
entryFormation	phr	Marker für Mehrwortausdruck
phraseType	idiom	Typ des Mehrwortausdrucks
entrySource		Quellenangabe (Wahrig, Duden etc.)
originator	ml	Administrative Daten
adminStatus	new	
company	Klipsa	
project	JAKOB	
updater	ml	
modDate	2008-03-22 20:43:41	
JAKOB Code	LAS-FUR	Verbkodierung der Erzählanalyse JAKOB (siehe Kodiersystem)