Entwurf eines (semi-)automatischen semantischen Annotationssystems für deutschsprachige Korpora

60. Jahrestagung
des Leibniz-Instituts für Deutsche Sprache

Dieses Forschungsprojekt wird vom Programm *Formación de Profesorado Universitario* vom spanischen Bundesministerium für Universitäten unter dem Kennzeichen **FPU21/00188** gefördert

Projekt ESMAS-ES+ (PID2022-137170OB-I00), gefördert von MCIN/AEI//FEDER

Iván Arias-Arias
Universidade de Santiago de Compostela
Instituto da Lingua Galega
Forschungsgruppe Humboldt (GI-1920)
ivanarias.arias@usc.gal

Forschungsstand und theoretische Grundlagen

Zur Beschreibung von nominalen Argumentstrukturen erweist sich **die Erhebung und Annotation korpusgestützter semantischer Daten** als unerlässlich (vgl. Domínguez, 2022 oder Valcárcel und Pino, 2023).

Es bestehen **semantische Ansätze zur Beschreibung des Lexikons und der Syntax-Semantik-Schnittstelle**: *FrameNet*, *CPA-PDEV*, *Verbario*, *ADESSE*, *Portlex*, Sprachgeneratoren wie *Xera* u.a. Keiner von denen kann direkt in einen semantischen Tagger umgewandelt werden.

Im Bereich der **künstlichen Intelligenz** und der **natürlichen Sprachverarbeitung** stehen uns Tools zum morphologischen Tagging zur Verfügung. Algorithmen, wie *word2vec* oder *BERT*, erfassen die Bedeutungen von Wörtern anhand ihres Kontexts.

Es besteht dennoch ein **Mangel an automatischen semantischen Annotationssystemen mit lexikographischer Anwendung**. Eine vorläufige Studie (Arias-Arias, 2022) weist auf die Durchführbarkeit eines derartigen Systems. Die Methode hat sich als zufriedenstellend für die Annotation von nominalen Argumentstrukturen im Deutschen herausgestellt.

Zwei methodologische Aspekte zur Entwicklung des Annotationssystems sind von grundlegender Bedeutung sind:

1. **die Kompilation eines Korpus**
2. **eine lexikalische Ontologie mit definierten semantischen Klassen**

Korpusbildung und -annotation: *wikidumps*

Zur **Datenerhebung und zur Annotation linguistisch-semantischer Informationen** stellt sich die Arbeit mit annotierten Korpora als wesentlich heraus.

Warum werden deutschsprachige *wikidumps* als Fokuskorpus verwendet?

- Das Korpus enthält hauptsächlich **aktuelle deutschsprachige Artikel aus Wikipedia** zu verschiedenen Themen.
- Es handelt sich um **wissenschaftlich informative Texte**, die im Prinzip von Laien verfasst werden und die **im Internet zum Herunterladen** bereitstehen.
- Da die Wikipedia **eine mehrsprachige Ressource** ist, kann **das Korpus einfach erweitert werden** unter Einbeziehung anderer Sprachen.
- Das **Web dient zur Erstellung von Korpora:** Wikipedia-Artikel sind bereits in verbreiteten Korpora wie *DeReKo* oder *DeTenTen20* vorhanden.
- Das Korpus ist **repräsentativ nur zu dem Zweck des Forschungsprojekts**: der Annotation und Extraktion semantischer Informationen. Zu einer höheren Repräsentativität können Referenzkorpora verwendet werden.
- Es werden *wikidumps* zusammengestellt, die bis **September 2023** veröffentlicht wurden.

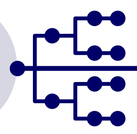
Inwiefern ist die Annotation von Korpora wichtig?

- Das Korpus wurde mithilfe des **TreeTaggers** (Schmidt, 1994) tokenisiert, mit Wortarten-Informationen annotiert und lemmatisiert. Das **Stuttgart-Tübingen-Tagset (STTS)** (Schiller, Teufel und Thielen, 1995) wurde angewandt.
- In Anlehnung an Stefanowitsch (2020, S. 122) wird unter linguistischer Annotation Folgendes verstanden: „a comprehensive operational definition for a particular variable, with detailed instructions as to how the values of this variable should be assigned to linguistic data".

Auswahl aus den *wikidumps* mit linguistischen Annotationen

BM_wiki_00.txt	Der	ART	die	article
BM_wiki_00.txt	Film	NN	Film	noun
BM_wiki_00.txt	beschreibt	VVFIN	beschreiben	verb
BM_wiki_00.txt	das	ART	die	article
BM_wiki_00.txt	Leben	NN	Leben	noun
BM_wiki_00.txt	der	ART	die	article
BM_wiki_00.txt	beiden	PIAT	beide	pronoun
BM_wiki_00.txt	Freunde	NN	Freund	noun
BM_wiki_00.txt	Blue	NN	<unknown>	noun
BM_wiki_00.txt	und	KON	und	conjunction
BM_wiki_00.txt	Eli	NE	Eli	name
BM_wiki_00.txt	:	$.	:	fullstop
BM_wiki_00.txt	Blue	NE	<unknown>	name
BM_wiki_00.txt	,	$,	,	comma
BM_wiki_00.txt	ein	ART	eine	article
BM_wiki_00.txt	nur	ADV	nur	adverb
BM_wiki_00.txt	wenig	ADV	wenig	adverb
BM_wiki_00.txt	erfolgreicher	ADJA	erfolgreich	adjective
BM_wiki_00.txt	Maler	NN	Maler	noun
BM_wiki_00.txt	erotischer	ADJA	erotisch	adjective
BM_wiki_00.txt	Bilder	NN	Bild	noun
BM_wiki_00.txt	,	$,	,	comma
BM_wiki_00.txt	wurde	VAFIN	werden	verb
BM_wiki_00.txt	von	APPR	von	preposition
BM_wiki_00.txt	seiner	PPOSAT	sein	pronoun
BM_wiki_00.txt	Freundin	NN	Freundin	noun
BM_wiki_00.txt	verlassen	VVPP	verlassen	verb

Lexikalische Ontologie: Definition der Klasse {Mensch}



Zur **Erstellung eines semantischen Tagsets** geht man von einer **lexikalischen Ontologie** (vgl. Domínguez, Valcárcel und Bardanca, 2021) aus.

Als semantische Klasse wird hier die **Klasse {Mensch}** ausgewählt: "Within the class of first-order entities persons occupy a privileged position; and the distinction between persons and non-personal entities is lexicalized or grammaticalized, in various ways, in many, and perhaps all, languages." (Lyons, 1977, S. 244)

{Mensch}
- {Allgemein}
- {Herkunft}
- {Verwandschaft}
- {menschlicher Umstand}: {negativer Umstand}, {Erziehung}, {Medizin} usw.
- {Agens}
- {Eigennamen}

Definition der Klassen ***per negationem***

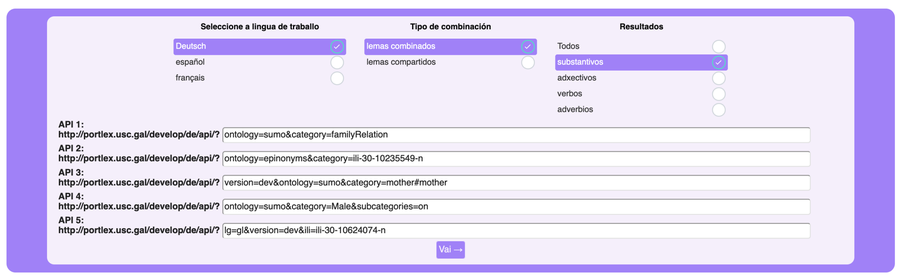
Festlegung der Unterklassen durch Beantwortung von Fragen wie bspw.:

1. Handelt es sich um ein Lebewesen?
2. Weist das Substantiv eine Verwandschaftsbeziehung auf?
3. Führt das Nomen eine negative Handlung aus?

Zur Auswahl von Kandidaten und zur Erweiterung der Klassen

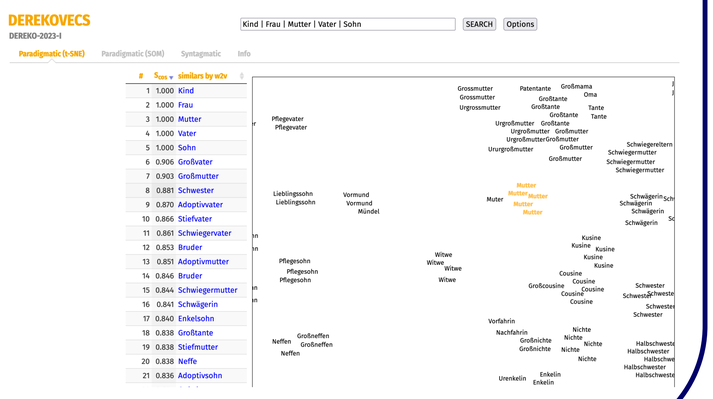
1. Eine **Frequenzliste der häufigsten deutschen Substantive** (vgl. Wolfer, Koplenig *et al.*, 2023) wird zu Rate gezogen und für jede Unterklasse werden 5 Nomina manuell extrahiert.
Am Beispiel der Unterklasse {Verwandschaft}: *Kind*, *Frau*, *Mutter*, *Vater*, *Sohn*.

2. Es werden **APIs** erstellt und es wird dank der Umsetzung von den **MultiTools** auf **WordNet** zurückgegriffen, um die Liste der Kandidaten zu erweitern.



Mögliche Kandidaten: *Tochter*, *Vater*, *Vatertier*, *Verbindung*, *Verwandter*, *Verwandte väterlicherseits*, *Verwandtschaft*, *Verwandtschaftsverhältnis*, *Großeltern*, *Großelternteil*, *große Schwester*, *Großmama*, *Großvater*, *Großvater mütterlicherseits*, *Großvater väterlicherseits*

3. Es werden gleichzeitig die 5 häufigsten Nomina in **DeReKoVecs** nachgeschlagen und andere **Kandidaten** können **herausgefiltert** und **ausgewählt** werden.



Vor der **Validierung der Kandidaten** ist eine **manuelle Überprüfung** der Substantive erforderlich. Verfahren, wie die **Inter-Coder-Reliabilität**, können zur Erstellung der Listen beitragen.

Ausgewählte Literatur:

Arias-Arias, I. (2022): Anotação semântica (semi)automática de corpora: a frase nominal em alemão. Braga: RepositoriUM. https://hdl.handle.net/1822/79649

Domínguez Vázquez, M. J. (2022): Contribución de la semántica combinatoria al desarrollo de herramientas digitales multilingües. In: *Círculo de Lingüística Aplicada a la Comunicación* 90, S. 171-188. DOI: 10.5209/clac.73849

Europäischer Master für Lexikographie (2018). Villa Vigoni Thesen. url: https://www.emlex.phil.fau.de/ueberuns/publikationen/andere-publikationen/

Lyons, J. (1977). *Semantics*. Cambridge: Cambridge University Press.

Schiller, A., S. Teufel und C. Thielen (1995): *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Universität Stuttgart, Universität Tübingen.

Schmid, Helmut (1994): Probabilistic part-of-speech tagging using decision trees. In: *Proceedings of International Conference on New Methods in Language Processing* 12.

Stefanowitsch, A. (2020). *Corpus linguistics. A guide to the methodology*. Berlin: Language Science Press. doi: 10.5281/zenodo.3735822

Valcárcel Riveiro, C. und L. Pino Serrano (2023). "Application d'une méthodologie d'analyse des prédicats nominaux: l'exemple du lexème MORT1". In: *Çédille, Revista De Estudios Franceses* 24, S. 557–579. doi: 10.25145/j.cedille.2023.24.27

Wolfer, S., A. Koplenig, M. Kupietz und C. Müller-Spitzer (2023). "Introducing DeReKoGram: A novel frequency dataset with lemma and part-of-speech information for German". In: *Research Square*. doi: 10.21203/rs.3.rs-3139640/v1

Online Ressourcen:

ADESSE = http://adesse.uvigo.es/
CPA-PDEV = https://pdev.org.uk/
DeReKoVecs = https://corpora.ids-mannheim.de/openlab/derekovecs
FrameNet = http://framenet.icsi.berkeley.edu/
MultiTools (darunter auch **Xera**) = http://portlex.usc.gal/combinatoria/
Ontologie = Domínguez V., M. J., Valcárcel R., C., Bardanca O., D. (2021). Ontología léxica. http://portlex.usc.gal/ontologia/.
Portlex = http://portlex.usc.gal/diccionario/
Verbario = http://www.tecling.com/verbario/