Originalveröffentlichung in: Lobin, Henning/Witt, Andreas/Wöllstein, Angelika (Hrsg.): Deutsch in Europa. Sprachpolitisch, grammatisch, methodisch. – Berlin [u.a.]: de Gruyter, 2021. S. 323-328. (Jahrbuch des Instituts für Deutsche Sprache 2020)
DOI: *https://doi.org/10.1515/9783110731514-019*

Erhard Hinrichs (Mannheim/Tübingen)/Patricia Fischer (Tübingen)/Yana Strakatova (Tübingen)

Rover und TüNDRA: Such- und Visualisierungsplattformen für Wortnetze und Baumbanken

Abstract: Geeignete Such- und Visualisierungswerkzeuge, idealiter in Form von Webapplikationen, sind für den benutzerfreundlichen Zugang zu Sprachressourcen von großer Bedeutung. In diesem Beitrag stellen wir die Webapplikationen Rover und TüNDRA vor, die am CLARIN-D Zentrum Tübingen im Rahmen des BMBF-Projekts CLARIN-D entwickelt wurden.

1 GermaNet mit Rover

Rover[1] bietet einen benutzerfreundlichen web-basierten Zugang für GermaNet an. GermaNet (Hamp/Feldweg 1997; Henrich/Hinrichs 2010) ist ein von der Universität Tübingen entwickeltes, maschinenlesbares lexikalisch-semantisches Wortnetz der deutschen Sprache. In GermaNet werden Nomen, Verben und Adjektive modelliert, indem synonymische lexikalische Einheiten in *Synsets* gruppiert werden. Ein Synset steht für ein Konzept in der Sprache, das mit unterschiedlichen lexikalischen Einheiten ausgedrückt werden kann. Zwischen den Synsets werden *konzeptuelle* Relationen definiert, z. B. Hyperonymie/Hyponymie (*Gebäck–Brezel*) und Meronymie (*Hand–Finger*). Hyperonymie/Hyponymie ist die grundlegende Relation in GermaNet, auf der die hierarchische Struktur des Wortnetzes basiert. Die *lexikalischen* Relationen in GermaNet werden zwischen lexikalischen Einheiten definiert, z. B. Synonymie (*Karotte–Möhre*) und Antonymie (*klein–groß*). GermaNet wird jährlich erweitert, die aktuelle Version (15.0) beinhaltet 144.113 Synsets mit 185.000 lexikalischen Einheiten.

Mit Hilfe von Rover können Nutzende die in GermaNet modellierten Lesarten von Einzelwörtern, semantische Relationen zwischen Wortbedeutungen und semantische Ähnlichkeit zwischen Lesarten in einer grafischen Benutzeroberfläche recherchieren. Zurzeit bietet Rover zwei Funktionen: Synsetsuche und

1 Online verfügbar unter https://weblicht.sfs.uni-tuebingen.de/rover (Stand: 27.5.2020).

https://doi.org/10.1515/9783110731514-019

semantische Ähnlichkeit. Bei der Suchfunktion wird sowohl die hierarchische Struktur der einzelnen Synsets grafisch angezeigt als auch alle konzeptuellen und lexikalischen Relationen, die für dieses Synset definiert sind. Die Synsetsuche bietet verschiedene Möglichkeiten, die Suchergebnisse zu steuern: durch Auswahl der Wortart, der semantischen Klasse und/oder der orthografischen Variante des Wortes. Außerdem ermöglicht Rover die Suche mit regulären Ausdrücken. Die Vielfalt an Suchoptionen macht Rover bereits für komplexere linguistische Studien anwendbar. Ein Beispiel dafür wäre die Analyse der Produktivität deutscher Komposita, bei der untersucht werden kann, in wie vielen Nominalkomposita ein gewähltes Substantiv als Kopf des Kompositums auftritt. Abbildung 1 a) veranschaulicht die Ergebnisse für die Anfrage in Rover: Im Suchfeld wird der reguläre Ausdruck [.+kuchen] angegeben und die Suche nur auf Substantive beschränkt. Damit werden alle nominalen Synsets aufgelistet, die auf „-kuchen“ enden (*Käsekuchen*, *Mohnkuchen*, *Krümelkuchen usw.*), insgesamt 47 Ergebnisse, dabei kann man sich jedes Synset genauer anschauen. Eine vergleichende Suche nach den Komposita mit dem Kopf „-torte“ ergibt nur 19 Treffer. Daraus wird deutlich, dass in der Bildung der deutschen Komposita das Substantiv „Kuchen“ produktiver als „Torte“ ist.

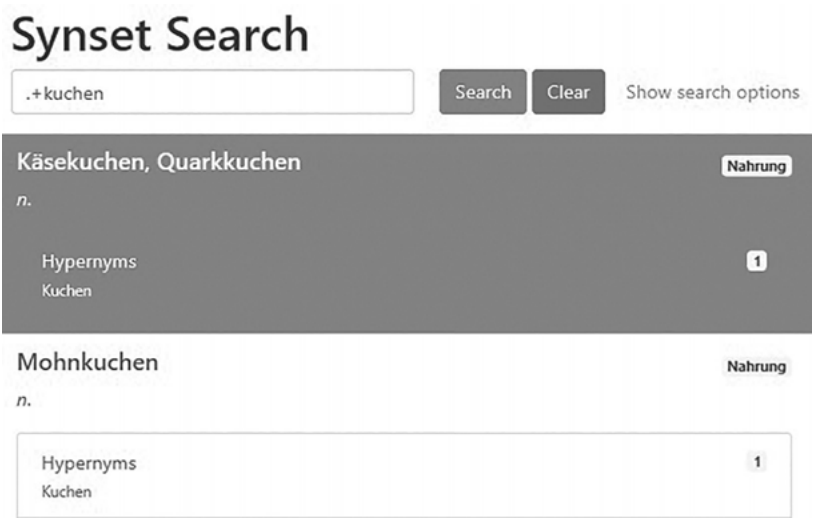

Abb. 1: a) Rover Synsetsuche nach allen Substantiven, die auf „-kuchen“ enden

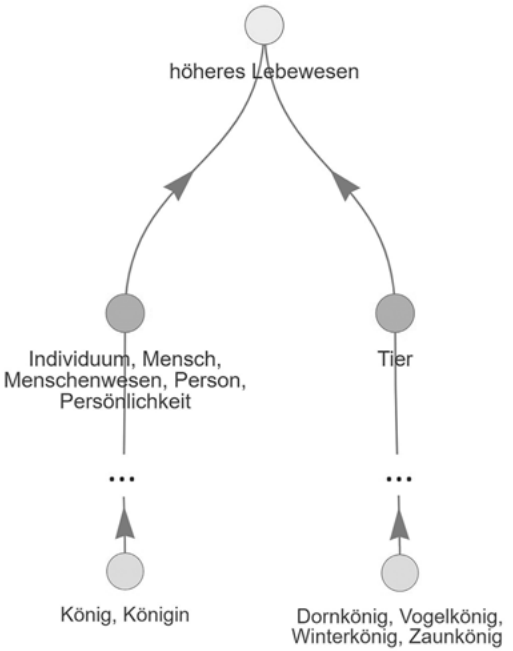

Abb. 1: b) Visualisierung der semantischen Nähe in Rover

Die zweite Funktion in Rover ermöglicht es den Nutzenden, semantische Ähnlichkeit zwischen zwei ausgewählten Synsets zu visualisieren und zu messen. So kann beispielsweise die Kompositionalität der Komposita untersucht werden, indem die semantische Ähnlichkeit zwischen einem Kompositum und seinem Kopf berechnet wird (Jana et al. 2019). Abbildung 1 b) zeigt die Ergebnisse für das Substantiv „Zaunkönig“ und seinen Kopf „König“. Die Grafik zeigt, dass die zwei Suchwörter den zwei unterschiedlichen Pfaden zugehörig sind. Das deutet an, dass das Kompositum „Zaunkönig“ seinem Kopf nicht nahe und deswegen nicht kompositionell ist.

Die obigen Suchanfragen sind nur beispielhaft für die Möglichkeiten, die Rover bereits bietet. Darüber hinaus sind weitere Funktionen in naher Zukunft geplant, die u. a. auf Rückmeldungen der Nutzenden basieren werden.

2 Baumbanken mit TüNDRA

Die Tübingen aNnotated Data Retrieval Application (TüNDRA; Martens 2013)[2] ermöglicht es Nutzenden, Baumbanken und andere linguistisch annotierte Korpora systematisch nach sprachlichen Mustern zu durchsuchen, die Suchergebnisse zu visualisieren und statistische Auswertungen zu Einzelwörtern, Phrasen und linguistischen Strukturen vorzunehmen. Baumbanken im Allgemeinen bezeichnen Korpora, in denen Beziehungen zwischen einzelnen Wörtern, Phrasen oder anderen linguistischen Ebenen annotiert sind. Es gibt zwei Annotationsarten für Baumbanken: Konstituentenstrukturen schließen syntaktische Beziehungen zwischen Phrasen ein, während Dependenzstrukturen nur Relationen zwischen Wörtern herstellen. Abbildung 2 zeigt den ersten Satz der TüBa-D/Z Baumbank (Telljohann et al. 2017) in Konstituenten- (2 a) und Dependenzstruktur (2 b). Neben syntaktischen Annotationen bietet TüNDRA detaillierte Informationen auf Wortebene. Dazu gehören Wortform, Lemma, Wortart sowie Kasus, Numerus und Genus, Person, Tempus und Modus.

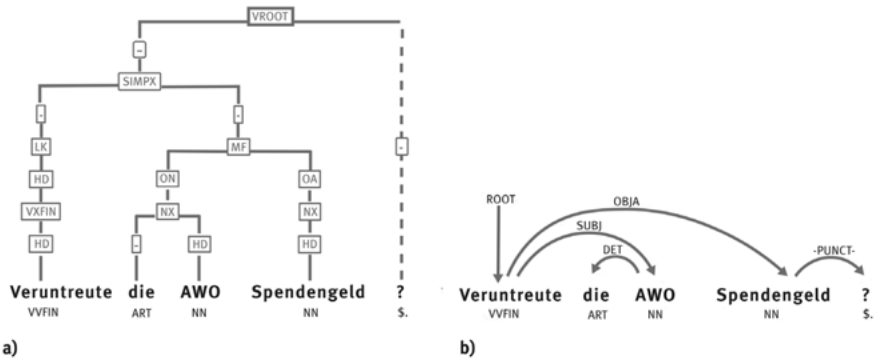

Abb. 2: Vergleich von **a)** Konstituenten- und **b)** Dependenzstruktur anhand des ersten Satzes der TüBa-D/Z Baumbank

2 Online verfügbar unter https://weblicht.sfs.uni-tuebingen.de/Tundra (Stand: 27.5.2020).

TüNDRA enthält 462 sowohl automatisch als auch von Experten händisch annotierte Baumbanken in verschiedenen Sprachen, darunter die folgenden in Tübingen erstellten Baumbanken des Deutschen:

- TüBa-D/Z beinhaltet 104.787 manuell annotierte Sätze aus der Berliner Tageszeitung taz.
- TüBa-D/W (De Kok 2014) beinhaltet 36 Millionen automatisch annotierte Sätze aus der deutschen Wikipedia.
- TüBa-D/S (Stegmann/Telljohann/Hinrichs 2000) beinhaltet rund 38.000 manuell transkribierte und annotierte Äußerungen spontaner Dialogsprache.

Über TüNDRA sind darüber hinaus ein Großteil der Baumbanken des Universal Dependencies Projekts (De Marneffe et al. 2014) sowie weitere Baumbanken verfügbar.

Für Suchanfragen wird die Syntax der TIGERsearch Suchsprache (Lezius 2002) verwendet. Diese erlaubt es, neben konkreten Wortformen auch auf alle anderen verfügbaren Annotationsebenen wie bspw. die Wortart zuzugreifen. Darüber hinaus können allgemeine Suchanfragen in Form von regulären Ausdrücken formuliert werden. Für komplexere Suchen lassen sich Suchkriterien außerdem kombinieren. So kann eine anfängliche Suche nach Sätzen mit der Wortform „Mannheim“ mit der Suchanfrage „Mannheim“ beginnen. In einem zweiten Schritt lässt sich die Suche auf alle Wörter mit der Endung „-heim” durch den regulären Ausdruck [lemma=/.+heim/] erweitern. Um die Suche auf Eigennamen einzuschränken, kann schließlich der Suchbegriff entsprechend zu [lemma=/.+heim/ & pos=“NE”] ausgebaut werden, wobei pos=“NE” hier für die Wortart *Eigenname* (engl. named entity) steht.

Die Suchergebnisse können anschließend zur weiteren Verarbeitung im csv-Format gespeichert werden. Statistische Analysen können innerhalb von TüNDRA umgehend durchgeführt werden. Im Falle des obigen Beispiels lässt sich so etwa schnell herausfinden, welche verschiedenen Eigennamen mit der Endung „-heim“ in der ausgewählten Baumbank mit welcher Häufigkeit vorkommen. Es kann dabei nach Wortform, Lemma oder anderen Annotationsebenen wie Kasus unterschieden werden.

Die Diversität an linguistischen Strukturen, Sprachen und Genres in Verbindung mit direkt verfügbaren Analysewerkzeugen macht TüNDRA vielseitig einsetzbar: Angefangen bei empirischer Forschung in den Bereichen Syntax und Grammatik, Lexikografie, Soziolinguistik und historischer Linguistik, über Trainingsdaten für Systeme der automatischen Sprachverarbeitung bis hin zur Ressource in den digitalen Geisteswissenschaften deckt TüNDRA viele zentrale Anwendungsfelder ab.

Literatur

De Kok, Daniël (2014): TüBa-D/W: A large dependency treebank for German. In: Proceedings of the 13th International Workshop on Treebanks and Linguistic Theories. Tübingen, S. 271–278.

De Marneffe, Marie-Catherine/Dozat, Timothy/Silveira, Natalia/Haverinen, Katri/Ginter, Filip/Nivre, Joakim/Manning, Christopher D. (2014): Universal Stanford Dependencies: a cross-linguistic typology. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC' 2014), Reykjavik. Paris: European Language Resources Association (ELRA), S. 4585–4592.

Hamp, Birgit/Feldweg, Helmut (1997): GermaNet – a lexical-semantic net for german. In: Proceedings of the ACL Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP applications.

Henrich, Verena/Hinrichs, Erhard (2010): GernEdiT – The GermaNet editing tool. In: Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC' 2010), Valletta. Tübingen: University of Tübingen, S. 2228–2235. www.lrec-conf.org/proceedings/lrec2010/pdf/264_Paper.pdf (Stand: 7.7.2020).

Jana, Abhik/Puzyrev, Dima/Panchenko, Alexander/Goyal, Pawan/Biemann, Chris/Mukherjee, Animesh (2019): On the compositionality prediction of noun phrases using poincare embeddings. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence: The Association for Computational Linguistics, S. 3263–3274.

Lezius, Wolfgang (2002): TIGERSearch – Ein Suchwerkzeug für Baumbanken. In: Proceedings Konvens 2002. 6. Konferenz zur Verarbeitung natürlicher Sprache. DFKI, Saarbrücken.

Martens, Scott (2013): TüNDRA: a web application for treebank search and visualization. In: Proceedings of the 12th Workshop on Treebanks and Linguistic Theories (TLT12), Sofia, S. 133—144. www.bultreebank.org/bg/twelfth-workshop-treebanks-linguistic-theories-tlt12 (Stand: 23.4.2020).

Stegmann, Rosmary/Telljohann, Heike/Hinrichs, Erhard (2000): Stylebook for the german treebank in VERBMOBIL. In: Verbmobil. Technical Report 239. www.sfs.uni-tuebingen.de/fileadmin/static/ascl/resources/stylebook_vm_ger.pdf (Stand: 7.7.2020).

Telljohann, Heike/Hinrichs, Erhard/Kübler, Sandra/Zinsmeister, Heike/Beck, Kathrin (2017): Stylebook for the Tübingen treebank of written german (TüBa-D/Z). www.sfs.uni-tuebingen.de/fileadmin/static/ascl/resources/tuebadz-stylebook-1707.pdf (Stand: 23.4.2020).