KoMuX – Der Kompositamuster-Explorer

Brunner, Annelen
brunner@ids-mannheim.de
Leibniz-Institut für Deutsche Sprache, Deutschland

Katrin, Hein
hein@ids-mannheim.de
Leibniz-Institut für Deutsche Sprache, Deutschland

KoMuX, der Kompositamuster-Explorer, (www.owid.de/plus/komux) ist eine Webanwendung, die es ermöglicht, mehr als 50.000 nominale Komposita des Deutschen gezielt nach abstrakten oder lexikalisch-teilspezifizierten Mustern zu durchsuchen. Unterschiedliche Visualisierungen helfen dabei, Strukturen und Zusammenhänge innerhalb der Ergebnismenge zu erfassen.

Mit KoMuX machen wir einen Teil der Datengrundlage frei verfügbar, auf der unsere empirischen Forschungen zur Wortbildung basieren und integrieren Analysen und Visualisierungen aus unseren Arbeiten. Der Explorer ist damit auch ein Beitrag zu OpenScience, indem er es ermöglicht, unsere Forschungsergebnisse in Teilen nachzuvollziehen und zu reproduzieren.

Forschungshintergrund

In der Wortbildungsforschung stellt die Einbeziehung authentischen Sprachmaterials nach wie vor ein Desiderat dar (vgl. z.B. Hein ersch. 2023; Elsen und Michel 2007). KoMuX ermöglicht es, Untersuchungen zur Komposition auf eine breite empirische Basis zu stellen und einer ‚empirischen Wortbildungsforschung' somit ein Stück weit näher zu kommen. Der Explorer basiert auf einer systematischen Datenerhebung, bei der alle nominalen Komposita automatisch aus dem KoGra-Untersuchungskorpus (KoGra 2022), einem Ausschnitt des Deutschen Referenzkorpus DeReKo (Kupietz u. a. 2010) , extrahiert wurden. Diese Datengrundlage ist unseres Wissens nach die erste ihrer Art.

Mit KoMuX wird erstmals eine Untermenge dieses Komposita-Inventars des Deutschen frei zugänglich und systematisch durchsuchbar gemacht, und zwar aus einer Muster-Perspektive (vgl. Stein und Stumpf 2019) heraus: Wir betrachten Komposita als konkrete sprachliche Realisierungen von zugrundeliegenden abstrakten oder lexikalisch-teilspezifizierten Mustern. Diese Muster aus spezifischen Paarungen von Erst- und Zweitgliedern wiederum können z.B. zur Erklärung von beobachtbaren Produktivitätsunterschieden herangezogen (vgl. Hein und Brunner 2020; Brunner u. a. 2021) oder – ganz allgemein – als Grundprinzip verstanden werden, das erklärt, wie die Komposition funktioniert bzw. wie sich das Inventar von Komposita grundsätzlich systematisieren lässt (vgl. Hein ersch. 2023) . Der Musteransatz bietet darüber hinaus eine direkte Anschlussfähigkeit an Grammatiktheorien wie die Konstruktionsgrammatik bzw. die Construction Morphology (Booij 2010) .

Datengrundlage

Das KoGra-Untersuchungskorpus umfasst ca. 7 Milliarden Tokens und besteht zum größten Teil (~90%) aus Pressetexten (zur genauen Zusammensetzung vgl. KoGra 2022; Bubenhofer, Konopka und Schneider 2014) . Es wurde mit einem automatischen Werkzeug annotiert, welches die Canoo Language Tools adaptiert , und für jedes Token detaillierte morphologische Informationen liefert, auf deren Basis nominale Komposita extrahiert wurden. KoMuX basiert auf einer Untermenge von 100.000 Komposita-Tokens, die zufällig aus der Gesamtmenge von ca. 489 Millionen Komposita-Tokens gezogen wurden. Daraus ergibt sich die Frequenzliste mit ca. 50.000 Komposita-Types, die durchsucht werden kann.

D ie automatischen morphologischen Analysen wurden manuell und semi-automatisch verbessert. Dies umfasste v.a. das Entfernen von falschen Einträgen (Tokens ohne Komposita-Status) sowie Korrekturen von falschen Zerlegungen und fehlerhaften Zuweisungen von Wortbildungstyp- und/oder Wortart-Klassifikationen für die Komposita-Konstituenten.

Technische Details

Die Komposita-Daten werden in einer MySQL-Datenbank verwaltet, die Anwendung selbst ist in JavaScript (node.js) programmiert. Für das Frontend wurden die Frameworks Vue.js und Vuetify.js verwendet, sowie die Bibliothek Apache ECharts für die Grafiken.

KoMuX-Tabellen können im CSV-Format heruntergeladen werden. Zusätzlich verfügt die Anwendung auch über eine API, über die direkte Anfragen an die Datenbank gestellt und die Ergebnisse im JSON-Format heruntergeladen werden können (vgl. http://www.owid.de/plus/komux/komux_api_doku/KoMuX.html).

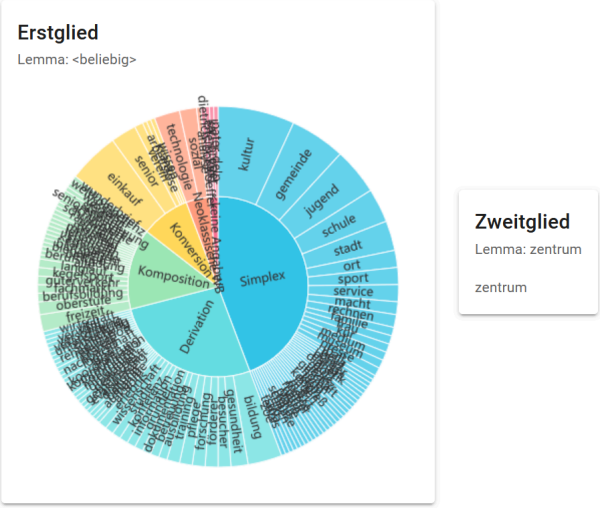
Funktionalitäten und Anwendungsbeispiele

Die musterbasierte Suche in KoMuX beruht darauf, dass grammatische Merkmale (Wortbildungstyp oder Wortart) oder lexikalische Eigenschaften (konkretes Lemma) für das Erst- und Zweitglied spezifiziert werden. Dies erlaubt es beispielsweise, gezielt alle Adjektiv+Nomen-Komposita (z.B. *Kleinkind*) oder Nomen+Nomen-Komposita (z.B. *Zeitpunkt*) zu extrahieren sowie systematisch nach Phrasenkomposita (z.B. *Nacht-und-Nebel-Aktion*) zu recherchieren. KoMuX ermöglicht zudem Untersuchungen zur Rekursivität von Komposita, indem gezielt nach Komposita gesucht werden kann, deren Erst- und/oder Zweitglied ebenfalls ein Kompositum ist (z.B. *Autobahnraststätte*) . Neben diesen formalen Suchebenen lassen sich auch konkrete Lemmata in Erst-

oder Zweitgliedposition festlegen und somit Muster mit lexikalischen Ankern definieren, z.B. Komposita mit Farbwörtern (z.B. *Dunkelrot, Abendrot*).

Visualisierungen helfen dabei, die Ergebnismenge näher zu analysieren. Auch hier steht die musterhafte Betrachtung von Erst- und Zweitgliedposition im Mittelpunkt.

Quantitative Verteilungen in Hinblick auf Wortart, Wortbildungstyp und Lemma werden mit Hilfe von mehrstufigen Tortendiagrammen sichtbar gemacht.

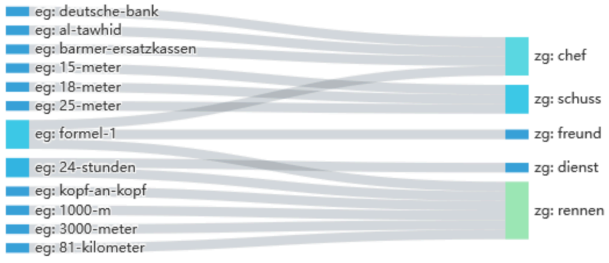


Quantitative Verteilungen (Suchebenen ‚Wortbildungstyp' und ‚Lemma') beim Erstglied für Komposita mit dem Zweitglied zentrum

.

Die Konstituenten-Ansicht zeigt alle Erst- und Zweitglied-Lemmata der Ergebnismenge, sowie deren Vorkommenshäufigkeiten in den jeweiligen Positionen. So lässt sich untersuchen, in welcher Position die lexikalische Vielfalt größer ist und welche Lemmata starke Tendenzen zu einer der beiden Positionen aufweisen.

Die Verknüpfungsansicht zeigt Komposita, deren Erst- oder Zweitglied-Lemma in mindestens einem weiteren Kompositum der Ergebnismenge auftritt und weist so auf produktive Bildungsmuster hin.



Ausschnitt aus dem Verknüpfungs-Diagramm für die Ergebnismenge mit dem Muster [PHRASE Erstglied^{+NOMEN} Zweitglied].

Ausblick

Wir streben an, in späteren Versionen auch semantische Suchebenen zu integrieren, z.B. die semantisch-thematische Klasse der unmittelbaren Konstituenten (z.B. ARTEFAKT, GEFÜHL) (vgl.Brunner u. a. 2021) und die Visualisierungen auszubauen . Perspektivisch wäre es wünschenswert, unser gesamtes Komposita-Inventar über den Explorer verfügbar zu machen.

Leider ist die Seite canoo.net nicht mehr online verfügbar. Teile der Inhalte wurden von LEO übernommen, allerdings nicht der morphologische Analysierer (vgl. https://dict.leo.org/pages/about/ende/canoonet_de.html).

Bibliographie

Booij, Geert E.2010. *Construction morphology*. Oxford: Oxford University Press.

Brunner, Annelen, Stefan Engelberg und Katrin Hein. 2021. „The distribution of constituent words in nominal compounds and its impact on semantic interpretation: an empirical study". *Journal of Word Formation*1: 7–36.

Bubenhofer, Noah, Marek Konopka und Roman Schneider.2014. *Präliminarien einer Korpusgrammatik*. *Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache* 4. Tübingen: Narr.

Elsen, Hilke und Sascha Michel. 2007. „Wortbildung im Sprachgebrauch. Desiderate und Perspektiven einer etablierten Forschungsrichtung". *Muttersprache* 117: 1–16.

Hein, Katrin. ersch. 2023. „Auf dem Weg zu einem Komposita-Konstruktikon? – ein empirischer Anwendungsversuch der Construction Morphology auf die Nominalkomposition im Deutschen". In *Konstruktionsfamilien im Deutschen*, hg. von Fabio Mollica und Sören Stumpf. Tübingen: Stauffenburg.

Hein, Katrin und Annelen Brunner. 2020.„Why do some lexemes combine more frequently than others? – An empirical approach to productivity in German compound formation". In *Rules, patterns, schemas and analogy. Online Proceedings of the 12th Mediterranean Morphology Meetings (MMM12)*12: 28–41.

KoGra. 2022. „Korpus des Projekts Korpusgrammatik". In *Leibniz-Institut für Deutsche Sprache: „Korpusgrammatik". Grammatisches Informationssystem grammis.* https://grammis.ids-mannheim.de/korpusgrammatik/6615.

Kupietz, Marc, Cyril Belica, Holger Keibel und Andreas Witt. 2010. „The German Reference Corpus DeReKo: A primordial sample for Linguistic Research". In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, herausgegeben von Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner und Daniel Tapias: 1848–54. Malta: European Language Resources Association (ELRA).

Stein, Stephan und Sören Stumpf. 2019. *Muster in Sprache und Kommunikation. Eine Einführung in Konzepte sprachlicher Vorgeformtheit*. Berlin: Erich Schmidt.