Projekt Deutscher Wortschatz

Uwe Quasthoff
Matthias Richter
Leipzig

Ce projet ambitieux du lexique allemand a pour but d'établir une vaste récolte de données linguistiques - telles qu'elles sont contenues dans des textes allemands de tout genre -, de les soumettre à une analyse statistique pour saisir l'utilisation du lexique tant en mots séparés qu'en groupes et de rendre les résultats accessibles au public. Chacun peut s'en servir comme simple dictionnaire d'orthographe ou de synonymes, mais également comme source d'inspiration pour lettres et autres textes à rédiger, grâce aux multiples phrases d'exemples. Finalement, enseignants et étudiants de langue à tous niveaux peuvent y trouver un outil précieux, puisque tous les mots sont présentés dans leur contexte ce qui facilite évidemment la compréhension et l'assimilation du vocabulaire et des formulations typiques d'une langue. Les méthodes d'analyse développées par la Section de Traitement automatique de langue à l'Institut d'Informatique de l'Université de Leipzig s'appliquent à toute langue et ont déjà généré un corpus pour des langues aussi différentes que l'anglais, le français, le hollandais, l'islandais ou le sorabe (idiome slave de l'Allemagne orientale). (réd.)

Das Projekt Deutscher Wortschatz baut zu einzelnen Sprachen umfangreiche Textkorpora auf, wertet diese statistisch aus und stellt die Ergebnisse unter der Web-Adresse http://wortschatz.uni-leipzig.de/ mit Hilfe eines Vollformenwörterbuchs durchsuchbar zur Verfügung. Für das Deutsche sind zusätzlich syntaktische und semantische Angaben wie Grundform zur Vollform, Grammatikangaben zur Grundform, Sachgebiet und Synonyme erschlossen. Im Vordergrund bei der Entwicklung stehen jedoch automatische Verfahren zur Verarbeitung von Text, welche unabhängig von der untersuchten Sprache funktionieren. Dabei werden zum Einen die Vorkommen einzelner Wörter und Wortgruppen gezählt. Zum Anderen werden statistisch signifikant miteinander auftretende Wörter und Gruppen ermittelt, sowohl als direkte Nachbarn, wie auch innerhalb eines Satzes. Die Ergebnisse dieser Analyse werden zusätzlich ähnlich einem Assoziationsnetz visualisiert. Gegenwärtig werden im Internet an Wochentagen mehr als 40.000 Einträge täglich nachgeschlagen. Die Interessen dieser Nutzer sind sehr breit gefächert und reichen von der Verwendung als Rechtschreibwörterbuch bis hin zur allgemeinen Informationsbeschaffung über die Beispielsätze. Von besonderer Bedeutung ist die Anwendung als ein erweitertes Synonymwörterbuch: Mittels Kookkurrenzen und Beispielsätzen werden umfangreiche Formulierungshilfen angeboten.

1. Geschichte und Einordnung

Seit Anfang der 90er Jahre werden an der Abteilung Automatische Sprachverarbeitung am Institut für Informatik der Universität Leipzig Ressourcen im Bereich der geschriebenen Sprache aufgebaut. Zu Beginn der Samm-



lung stand das Ziel, frei verfügbare Daten zur deutschen Sprache zu sammeln und im Lauf der Zeit für die weitere Nutzung zu erschliessen. Schnell wichen Sammeln einzelner Wörter und Speicherung von Beispielsätzen dem Aufbau einer umfassenden Textdatenbank. Zu deren Erschliessung wurde schliesslich Software entwickelt, mit deren Hilfe beliebige Sprachdaten in grossem Massstab statistisch untersucht und präsentiert werden können.

2. Daten und untersuchte Sprachen

Die Datengrundlage bilden überwiegend Sätze aus der aktuellen überregionalen Tagespresse, in besonderem Masse ab ca. 1995. Zum Zeitpunkt des Erscheinens dieses Artikels umfasst das Korpus Deutsch ca. 38 Millionen Sätze mit 750 Millionen laufenden Wörtern und über 7 Millionen Einträgen zu unterschiedlichen Wortgruppen und Vollformen. Für jeden Eintrag sind die Anzahl im Korpus sowie Beispielsätze verfügbar; zudem wird, sofern diese vorliegt, weitere Information bereit gehalten. Zunächst einmal sind dies hinlänglich aus Wörterbüchern bekannte Angaben wie Beschreibung, Grammatik, Morphologie usw. Eine Reduktion auf Lemmata wie in üblichen Wörterbüchern findet nicht statt. Statt dessen sind Lemmata und die darunter zu fassenden Vollformen über Links miteinander verbunden. Weitere Links enthalten Angaben zur Pragmatik, zu Teilwörtern und Wortgruppen und vor allem auch zu semantisch verwandten Einträgen wie Synonymen und Antonymen, Ober- und Unterbegriffen usw. Die Bedeutungsgruppen des anhand des Wortschatz-Korpus überarbeiteten onomasiologischen Wörterbuchs "Dornseiff – Der Deutsche Wortschatz nach Sachgruppen" verdienen hierbei besondere Erwähnung.

Der Vollformenansatz kann Sprachlernenden besonders dann eine Orientierungshilfe sein, wenn bestimmte Umfelder eine spezifische Form bedingen oder bevorzugen.

Die im Umfeld des Projekts entwickelten Verfahren arbeiten unabhängig von konkreten Sprachen und wurden bereits erfolgreich auf Korpora derart unterschiedlicher Sprachen wie Französisch, Finnisch, Koreanisch oder Japanisch angewandt. Über den eingangs genannten Website sind neben dem beschriebenen Deutschen derzeit Korpora von Englisch, Französisch, Niederländisch, Isländisch, Estnisch und Sorbisch abfragbar. Für diese Sprachen steht jeweils das Vollformenwörterbuch mit Beispielsätzen sowie die Kookkurrenzen zur Verfügung. Der Umfang der Korpora schwankt je nach Verfügbarkeit der Daten von 300.000 Sätzen (Sorbisch) bis ca. 10 Millionen Sätze (Englisch).

3. Statistische Auswertung

Wörter, Wortformen und deren Frequenzen zu zählen und in Ranglisten zu ordnen, hat im Deutschen eine über 150 jährige Tradition, die mit Namen wie Wilhelm Kaeding oder Helmut Meier verknüpft ist. Die sichere Kenntnis der Einträge einer solchen Rangliste ermöglicht es Lernenden bereits mit sehr beschränktem Vokabular die Bausteine der Gerüste beliebiger Texte zu erkennen und den Kopf frei zu halten für die sinntragenden Elemente. Natürlich werden also beim Projekt Deutscher Wortschatz auch Frequenzen und dem korpusübergreifenden Vergleich dienende Frequenzklassen berechnet und angezeigt.
Während der Vollformenansatz die lexikographische Auswertung eher erschwert, kann er Sprachlernenden besonders dann eine Orientierungshilfe sein, wenn bestimmte Umfelder eine spezifische Form bedingen oder bevorzugen. Statistik über das auffällige gemeinsame Auftreten von Wörtern kann auf solche Umstände aufmerksam machen, die belegen und auch dabei helfen, diese Gewinn bringend anzuwenden. Zu den Korpora des Wortschatz-Projekts werden so genannte Kookkurrenzen berechnet, das sind nichts anders als statistisch signifikant gemeinsam vorkommende Wörter. Dies geschieht zum Einen für linke und rechte Nachbarn, zum Anderen auf der Basis ganzer Sätze.
Da im Korpus schriftlicher Sprachgebrauch und Wissen übergangslos miteinander verbunden sind, umfassen Kookkurrenzen zweierlei: Kollokationen, also Fügungen mit sprachlichem Hintergrund, und eher semantisch bedingte typische Kontexte. Je nachdem, ob nun Sprachproduktion oder -rezeption im Vordergrund steht, ist der eine oder aber der andere Teil wichtiger. Kookkurrenzen auf der Basis von Sätzen fassen offensichtlich in erhöhtem Maße die mehr vom Inhalt als von der Sprache bestimmten Zusammenhänge.
Sehr deutlich wird das unterschiedliche Umfeld verschiedener Flexionsformen am Beispiel der jeweils 24 stärksten Kookkurrenzen der Vollformen *Schwein* und *Schweine*.

Satz-Kookkurrenzen zu Schwein: *Rind, Mensch, gehabt, Du, Huhn, Schaf, Kuh, Spieß, Geflügel, Kalb, armes, geschlachtet, intrigantes, Lamm, Kein, ruft, Sau, Mücker, mußt, arme, du, Ferkel, Fisch, Tier*

Satz-Kookkurrenzen zu Schweine: *Rinder, Schafe, Hühner, Kühe, Geflügel, Ziegen, verfüttert, geschlachtet, Tiermehl, Tiere, Pferde, Kälber, arme, Hunde, Futter, Stall, Gänse, Ferkel, Schweinepest, schlachten, Verfütterung, getötet werden, Bauern, Enten*

Sprachlich fällt auf, dass Singular und Plural des Ausgangsworts sich jeweils auch wieder bei den Kookkurrenzen finden lassen. Inhaltlich lässt sich zu-

sammenfassen, dass sich in der Pluralform nahezu ausschliesslich der Kontext des Nutztiers, das auf dem Bauernhof lebt und geschlachtet wird, wiederspiegelt, während die Singularform zusätzlich noch auf eine Reihe von Redewendungen hinweist.

4. Visualisierung

Die Satz-Kookkurrenzen lassen sich ähnlich wie Assoziationsnetze graphisch darstellen. Zu einem Startwort werden dafür solche Kookkurrenzen ausgewählt, die auch untereinander wieder Kookkurrenzen sind. Damit wird ein minimaler Zusammenhang zwischen den ausgewählten Kookkurrenzen erzwungen. Die Stärke der Verbindungslinie repräsentiert den Wert des verwendeten statistischen Masses. Die Anordnung erledigt ein Programm so, dass sie möglichst übersichtlich wird.

Wörtern. Mehrdeutige Wortformen, bei denen für die einzelnen Bedeutungen jeweils genügend viele Vorkommen im Korpus belegt sind, zeigen eine Trennung in diese so genannten Cluster besonders deutlich. Im abgebildeten Beispiel ist Faust das Startwort und die ermittelten Kookkurrenzen zerfallen klar in zwei Bedeutungen, welche auch in den Angaben zum Eintrag Faust auftauchen: einmal die *geballte Hand* mit *ballte, geballter, geballten, schlug, ins Gesicht, auf den Tisch*, dann das *Schauspiel von Goethe* mit *Goethe, Goethes, Gretchen, Mephisto, Mephistopheles* und *Peter Stein* (der den Faust inszeniert hat).

In der Webpräsentation sind die Listen und Graphen von Kookkurrenzen als Links auf die entsprechenden Einträge ausgeführt. Dadurch wird unter anderem eine interessengeleitete spielerische Nutzung entlang von Kontexten nahegelegt. Gerade Sprachlehrende und -lernende können dies immer wieder neu zur Inspiration und als Wissensquelle nutzen:

- Wörter werden im Sinnzusammenhang angeboten und können so einfacher gelernt werden.
- Dieser Sinnzusammenhang wird für nahezu alle Wörter angeboten, nicht nur für einige wenige wie im Lehrbuch.
- Bei der Textproduktion eignen sich Kookkurrenzen und Beispielsätze für Sprachlerner mit fortgeschrittenen Kenntnissen, um typische Formulierungen wiederzuerkennen
- Für Übersetzer hilfreich ist das Nachschlagen solcher Wörter, die in sonst keinem Wörterbuch gefunden werden.

Auch wenn das Wortschatz-Wörterbuch nicht vorrangig für Sprachlerner konzipiert wurde, zeigen die Ergebnisse einer Nutzerumfrage doch den grossen Anteil solcher Anwender. Für Anregungen zur weiteren Verbesserung der Funktionalität sind die Autoren stets dankbar.

Graph v.1.5 für Faust

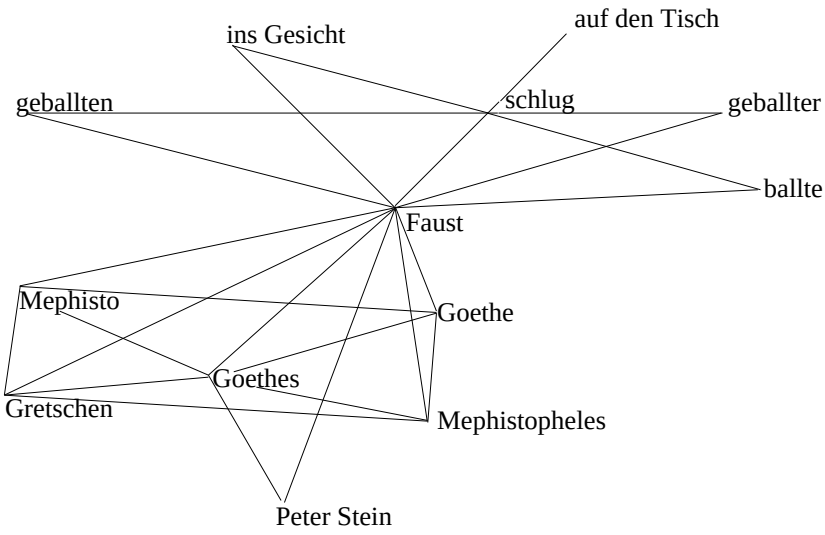

Abbildung 1: Kookkurrenzgraph zu "Faust"

Da im Kontext häufig mehrere Wörter gemeinsam auftreten, zeigen die Graphen auch häufig ein oder mehrere eng verbundene Grüppchen von

Uwe Quasthoff
Prof. Dr., ist Mathematiker und Informatiker und leitet das Projekt Deutscher Wortschatz seit seinem Beginn. Sein Interesse gilt sprachunabhängigen Verfahren zur inhaltlichen Analyse von Texten.

Matthias Richter
M.A., studierte Kommunikations- und Medienwissenschaft und Informatik. Er beteiligt sich seit Ende 2000 am Projekt Deutscher Wortschatz und beschäftigt sich mit der Ermittlung von Trends aus zeitlichen Abfolgen von Texten.