Maschinelle morphologische Analyse für das Deutsche. Ein Überblick. – Exposé zur Studienarbeit –

Peter Adolphs, 1. Juli 2005

Zusammenfassung: Ziel meiner Studienarbeit ist es, einen Überblick über bestehende Verfahren und Systeme zur maschinellen morphologischen Analyse deutscher Wörter zu geben. Diese ist insbesonders in den im Deutschen sehr produktiven Wortbildungskomponenten Derivation und Komposition noch nicht zufriedenstellend gelöst. In der Studienarbeit werde ich die Stärken und Schwächen der verfügbaren Systeme diskutieren.

1. Motivation

Für viele computerlinguistische Anwendungen und andere sprachverarbeitende Systeme ist eine Analyse von Wortformen erforderlich. Oftmals muss eine Lexikonkomponente[1] die grammatischen Merkmale einer Wortform bestimmen oder sie auf ihren Wortstamm bzw. auf einen zugrundeliegenden Lexikoneintrag zurückführen (Lemmatisierung). Für die meisten Anwendungen ist es notwendig, dass die Komponente auch die Wörter in unbekanntem Text, mit einem nicht abgegrenztem Vokabular, richtig analysieren kann.

Wie wahrscheinlich ist es, dass eine Lexikonkomponente auf ein Wort stößt, das sie zuvor nicht gesehen hat? Statistische Untersuchungen zeigen, dass natürlichsprachliche Texte bzw. Textsammlungen eine große Anzahl von Wörtern mit sehr niedriger relativer Vorkommenshäufigkeit aufweisen. Häufigkeitsverteilungen mit einer solchen Charakteristik werden auch LNRE-Verteilungen genannt (Large Number of Rare Events; siehe Baayen 2001). Anders als bei anderen Verteilungen kann eine LNRE-verteilte Stichprobe nicht repräsentativ für die Grundgesamtheit stehen, da die Stichprobe nur einen kleinen Teil der tatsächlich möglichen Elemente aufweist. Bei Häufigkeitsverteilungen von Wörtern in Texten gilt diese Eigenschaft für alle gängigen Größen von Texten[2]. Folglich ist die Wahrscheinlichkeit, in unbekanntem Text auf ein vorher nicht gesehenes Wort zu stoßen, hoch. Es ist daher nicht möglich, ein auch nur annähernd vollständiges Lexikon allein durch Auflistung der in einem Textkorpus vorkommenden Wörter zu erstellen.

Die unbekannten Wörter lassen sich in verschiedene Typen klassifizieren. Ein bedeutender Typ sind die Wörter, die mit regulären morphologischen Prozessen, also durch produktive Wortbildungsmittel, gebildet werden[3]. Diese hohe Produktivität deutscher Wortbildung lässt sich ohne eine maschinelle morphologische Analysekomponente nicht handhaben.

---

1 Unter "Lexikon" verstehe ich ganz allgemein diejenige Komponente, die zu jeder Wortform zugehörige Eigenschaften bestimmt, z.B. Wortart, Flexionseigenschaften, phonologische Form, Semantik, etc. Jedes Lexikon verfügt über einen deklarativen Teil, dessen Einträge "Lexikoneinträge" heißen.

2 Baayen (2001:51): "Even large corpora with tens of millions of words are located in the LNRE zone."

3 Weitere Typen sind nach ten Hacken & Lüdeling (2002) unter anderem Eigennamen, Rechtschreibfehler, fremdsprachliche Zitate und Neologismen.

2. Linguistische Grundlagen

Die Morphologie beschäftigt sich mit dem formalen Aufbau von Wörtern. Für das Deutsche werden dabei gängigerweise drei reguläre Phänomenbereiche unterschieden: Flexion, Derivation und Komposition. Derivation und Komposition werden auch als Wortbildung bezeichnet[4].

Die Flexion behandelt die systematische Anpassung von Wortstämmen an einen grammatischen Kontext. Die grammatischen Kategorien, die dabei eine Rolle spielen, sind beispielsweise Numerus, Genus und Kasus für Nomen, Tempus für Verben, etc. Die daraus resultierenden Wortformen bilden das Flexionsparadigma und werden in einem abstrakten Lexikoneintrag, dem sogenannten Lexem[5], zusammengefasst.

Die Derivation bezeichnet die Bildung eines Lexems durch die Abwandlung des Wortstamms eines anderen Lexems. Dies geschieht üblicherweise durch Suffigierung (z.B.[6] *Arbeiter*$_N$ aus dem Verbstamm *arbeit*$_V$) oder Präfigierung (z.B. *verarbeit*$_V$ aus *arbeit*$_V$). Wie aus den Beispielen hervorgeht, wird die Bedeutung des Basislexems durch ein Affix mehr oder weniger systematisch verändert. Dies kann auch mit einem Wortartwechsel verbunden sein. So erzeugt das Suffix *-er* in der Regel aus einem verbalen Lexem ein nominales Lexem, das das Agens oder Instrument aus der durch das verbale Lexem denotierten Tätigkeit bezeichnet.

Die Komposition bezeichnet die Bildung eines Lexems durch die Kombination zweier Lexeme (z.B. *Arbeitsamt*$_N$ aus *Arbeit*$_N$ und *Amt*$_N$). Die Wortarten beider Kompositionsglieder können dabei verschieden sein (z.B. *Fernglas*$_N$, *ofenwarm*$_A$). Die Wortart und auch die Flexionseigenschaften des Kompositums werden durch das letztstehende Kompositumsglied bestimmt.

Wortbildungsbasen können selbst komplex sein. So lassen sich Derivationen von derivierten Wörtern bilden (*verarbeitbar*$_A$, *unverarbeitbar*$_A$, *Unverarbeitbarkeit*$_N$), Komposita von Komposita (*Arbeitsamtsleiter*$_N$), Komposita von derivierten Wörtern (*Verarbeitbarkeitskriterium*$_N$) sowie Derivationen von Komposita (*jungfräulich*$_A$, *Verhochschulung*$_N$, *Schildbürgerei*$_N$). Die Anwendung von Wortbildungsregeln unterliegt zwar vielfältigen individuellen Beschränkungen. Doch insbesonders Nominalkomposita zeigen, dass die Wortbildung im Deutschen eine unbeschränkte Anzahl von Neubildungen zulässt (*Dampfschifffahrtskapitänsmützeninnenfutterhersteller*$_N$). Da zudem viele dieser Wortbildungsmittel hoch produktiv[7] und für einen großen Teil des Wortschatzes verantwortlich sind[8], ist eine morphologische Komponente für ein maschinelles Lexikon, welches eine hohe Abdeckung erreichen soll, unverzichtbar[9].

4 Auf weitere Wortbildungsmittel wie Konversion, Kurzwortbildung und Wortkreuzung gehe ich hier nicht ein.

5 In der Computerlinguistik ist auch der Begriff "Lemma" üblich.

6 Abkürzungen: N=Nomen, A=Adjektiv, V=Verb, Fem=Femininum, Nom=Nominativ, Gen=Genitiv, Dat=Dativ, Akk=Akkusativ, Sg=Singular, Präf=Präfix, Suff=Suffix

7 Zur Produktivität von einzelnen deutschen Wortbildungsprozessen siehe z.B. Lüdeling & Evert (2003) und Evert & Lüdeling (2001). Zur Produktivität im Niederländischen siehe z.B. Baayen (2001:203-10).

8 Laut Ortner et al (1991:3) besteht der dt. Wortschatz zu etwa zwei Dritteln aus Nominalkomposita.

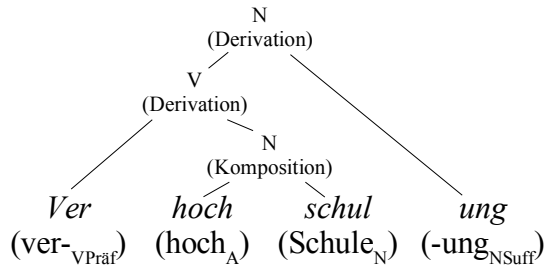
9 Vgl. auch Fitschen (2004) sowie ten Hacken & Lüdeling (2002).

3. Maschinelle morphologische Analyse

Eine maschinelle morphologische Analyse sollte verschiedene Anforderungen erfüllen. Ihre Analysen sollten generell zutreffend sein, spezifisch genug für den Anwendungszweck sein und eine hohe Abdeckung erreichen. Außerdem ist eine effiziente Verarbeitung und Robustheit wünschenswert.

Für das Deutsche sind bereits etliche Morphologiesysteme verfügbar[10]. Diese Systeme basieren auf verschiedenen formalen Modellen, wie z.B. endlichen Automaten, Merkmalstrukturen mit Unifikation oder linksassoziative Grammatiken. Dabei beherrschen fast alle Systeme die Flexionsanalyse, einige unterstützen die Analyse ausgewählter Kompositionsmuster oder bieten eine (meist unvollständige) Derivationsanalyse[11]. Es gibt derzeit kein frei verfügbares System, welches eine Wortformanalyse mit hoher Korrektheit, hoher Spezifizität und hoher Abdeckung erstellt.

Die Wortformanalyse kann in mehreren Detailstufen und damit in verschiedener Spezifität erfolgen. In der einfachsten Variante wird die Eingabezeichenkette in die Wortbestandteile zerlegt (z.B. *Ver· hoch·schul·ung*). Diese Wortbestandteile können in einer weiteren Stufe mit Lexikoneinträgen in Verbindung gebracht werden (z.B. *Ver*(ver-$_{VPräf}$) · *hoch*(hoch$_{A}$) · *schul*(Schule$_{N}$) ·*ung*(-ung$_{NSuff}$)). Die detaillierteste Analyse verdeutlicht die Wortbildung anhand eines Strukturbaums über den segmentierten Wortbestandteilen:



Das Hauptproblem bei jeder dieser Analysestufen besteht in der hohen Ambiguität. So lassen das Lexikon der Wortbestandteile in Zusammenarbeit mit den Wortbildungsregeln häufig eine Vielzahl von Analysen zu, die einem Menschen zwar unsinnig erscheinen, doch ohne weitere Kriterien völlig gleichberechtigt neben den gewollten stehen.

Bei der Segmentierung in Wortbestandteile lässt sich etwa ohne weiteres nicht bestimmen, ob das Wort *Druckerei* in *Druck·erei* (Wortstamm *druck*, Ableitungssuffix *-erei*) oder *Druck·er·ei* (Wortstamm *druck*, Ableitungssuffix *-er*, Wortstamm *ei*) zerlegt werden soll. Ebenso ist bei *Staubecken* eine Segmentierung in *Stau·becken* oder in *Staub·ecke·n* möglich.

10 Einige werden in Fitschen 2004 vorgestellt.
11 vgl. Fitschen 2004:45f und ten Hacken & Lüdeling 2002

Bei der Assoziation von Segmenten mit Lexikoneinträgen kann es ebenfalls zu Ambiguitäten kommen. Schon bei einer einfachen Wortform wie *Kunde* kann ohne weiteres nicht entschieden werden, zu welchem Lexikoneintrag sie gehört (*der Kunde* vs *die Kunde*). Ebenso bei komplexen Wortformen: *Platz* als Kompositumsbestandteil kann eine verbale Lesart (von *platzen*) haben oder eine nominale (von *Platz* im Sinne von *Ort*). Dementsprechend gibt es auch zu *Platz·regen* und *Platz·karte* jeweils zwei verschiedene strukturelle Analysen, wovon jeweils nur eine die richtige ist.

Bei der Bestimmung der Struktur zu den segmentierten Wortformen *Bundes·presse·amt* sowie *Bundes·kanzler·amt* sind jeweils zwei Analysen möglich, aber nur je eine beabsichtigt[12]: [*Bundes·*[*presse·amt*]] und #[[*Bundes·presse*]·*amt*] sowie #[*Bundes·*[*kanzler·amt*]] und [[*Bundes·kanzler*]·*amt*].

Unabhängig vom Aufbau komplexer Wörter müssen die Flexionsmerkmale einer Wortform bestimmt werden. Auch hier gibt es ambige Analysen, die auf die Ambiguität von Wortformen zurückgehen. So sind die Formen für Nominativ, Genitiv, Dativ und Akkusativ Singular von *Verhochschulung* alle gleich. Wenn lediglich einzelne Wortformen analysiert werden sollen, so ist dies das gewünschte Ergebnis. Wenn die zu analysierenden Wortformen aber Teil eines Textes sind, so sind die Flexionsmerkmale durch den grammatischen Kontext genau bestimmt und damit nicht ambig.

Ambige Analysemöglichkeiten stellen ein häufig auftretendes Problem in der Computerlinguistik dar. Für verschiedene Problembereiche (z.B. für die maschinelle Syntax- oder Semantikanalyse von Sätzen) wurden unterschiedliche Lösungsstrategien entwickelt. Mögliche Verfahren sind, die Alternativen einfach aufzulisten, sie durch Unterspezifizierung zusammenzufassen oder sie zu disambiguieren (z.B. durch regelbasierte, probabilistische oder hybride Verfahren). Wie sich diese Ansätze auf die maschinelle morphologische Analyse übertragen lassen, muss noch gezeigt werden.

4. Vorhaben

In meiner Studienarbeit werde ich den Forschungsstand zur maschinellen morphologischen Analyse des Deutschen zusammenfassen. Dafür möchte ich die grundsätzlichen Problemfelder der deutschen Morphologie sowie die vorhandenen Lösungsansätze für deren maschinellen Analyse, insbesonders in Hinblick auf die Wortbildung, vorstellen und diskutieren. Soweit es anhand der Dokumentation und öffentlichen Verfügbarkeit möglich ist, möchte ich die für das Deutsche entwickelten Morphologiesysteme beschreiben und in Hinblick auf die oben genannten Kriterien (Korrektheit, Spezifizität, Abdeckung, Effizienz und Robustheit) qualitativ untersuchen.

Die Studienarbeit soll die Grundlage für die Diplomarbeit bilden. Das geplante Thema für die Diplomarbeit ist die Entwicklung von Strategien zur Disambiguierung von mehrdeutigen strukturellen Analysen von Komposita.

12 Das Zeichen # kennzeichnet, dass eine Struktur zwar grammatisch wohlgeformt, aber nicht sinnvoll interpretierbar ist.

Literatur

R. Harald Baayen: *Word Frequency Distributions*. Dordrecht / Boston / London 2001.

Laurie Bauer: *Introducing Linguistic Morphology*. Edinburgh 2003: Edinburgh University Press.

Geert E. v. Booij, Christian Lehmann, Joachim Mugdan (Hrsg.): Morphologie / Morphology. 1. Halbband / Volume 1. Berlin, New York 2000: de Gruyter. (Handbücher zur Sprach- und Kommunikationsforschung / HSK 17.1)

Walter Daelemans: *Computational linguistics.* In: Geert E. v. Booij, Christian Lehmann, Joachim Mugdan, Stavros Skopeteas (Hrsg.): Morphologie / Morphology. 2. Halbband / Volume 2. Berlin, New York 2004: de Gruyter. (Handbücher zur Sprach- und Kommunikationswissenschaft / HSK 17.2) S. 1893-1900.

Stefan Evert & Anke Lüdeling: *Measuring morphological productivity: Is automatic preprocessing sufficient?* In: Paul Rayson; Andrew Wilson; Tony McEnery; Andrew Hardie and Shereen Khoja (eds.), Proceedings of the Corpus Linguistics 2001 conference. Lancaster 2001. pp. 167 – 175

Arne Fitschen: Ein Computerlinguistisches Lexikon als komplexes System. Dissertation Universität Stuttgart, 2004.

Ulrich Heid: *Morphologie und Lexikon*. In: Günther Görz (Hrsg.): Handbuch der Künstlichen Intelligenz. München: Oldenbourg, 2000, S. 665-709.

Pius ten Hacken & Anke Lüdeling: *Word Formation in Computational Linguistics*. Proceedings of TALN 2002 [Traitement Automatique du Langage Naturel], Vol. 2. Nancy 2002. pp. 61-87.

Anke Lüdeling & Stefan Evert: *Linguistic experience and productivity: Corpus Evidence for fine-grained distinctions*. In: Proceedings of the 2003 Corpus Linguistics Conference, Lancaster 2003.

Lorelies Ortner, Elgin Bollhagen-Müller, Hanspeter Ortner, Hans Wellmann, Maria Pümpel-Mader, Hildegard Gärtner: *Deutsche Wortbildung. Typen und Tendenzen in der Gegenwartssprache 4: Substantivkomposita*. Berlin 1991: Walter de Gruyter.

Richard Sproat: *Lexical Analysis*. In: Robert Dale, Hermann Moisl, Harold Somers (eds.), Handbook of Natural Language Processing. New York, Basel 2000: Marcel Dekker. S. 37-57.

Jochen Trommer: *Morphologie*. In: Kai-Uwe Carstensen, Christian Ebert, Cornelia Endriss, Susanne Jekat, Ralf Klabunde und Hagen Langer (Hrsg.): Computerlinguistik und Sprachtechnologie. München 2004: Elsevier.

Harald Trost: *Morphology*. In: Ruslan Mitkov (Hrsg.): The Oxford Handbook of Computational Linguistics. Oxford 2003: Oxford University Press.