Affixerkennung in deutschen Wortformen

Ein nicht-lexikalisches Segmentierungsverfahren nach N. D. Andreev

Oliver Cromm
Universität Göttingen
e-mail: ocromm@gwdg.de

24. März 1997

A non-lexical statistical method for affix recognition within a corpus — developed by N. D. Andreev ([1], [2]) in the early 60's — is applied to German texts. The algorithm was designed to identify inflectional as well as derivational paradigms in any language, using both statistical overrepresentation of letter sequences and combinatorial systematicity as a measure for the reliability of segmentation.

By testing this method on smaller samples, the influence of several parameters is investigated and the most suitable values are selected. The resulting algorithm is then applied to a bigger corpus, the insuing results are analysed both quantitatively and qualitatively. These display quite good recall rates but reveal some problems due to certain characteristics of German.

1 Überblick

1.1 Grundlegender Ansatz

Das Verfahren von Andreev sucht Affixe am Beginn und Ende von Wörtern. In seiner reinsten Form, wie sie hier untersucht wird, benutzt es keinerlei syntaktische oder semantische Information, sondern berücksichtigt ausschließlich Abweichungen von der Gleichverteilung der Buchstaben, nach Harris eine der wichtigsten Grundlage einer Sprachtheorie [5]. Der Erkennungsprozeß beginnt damit, nach überrepräsentierten Buchstaben im Anfangs- und Endbereich der längeren Wörter eines Korpus zu suchen, im Deutschen z. B. *e* an vorletzter Wortposition. Deren Nachbarbuchstaben werden daraufhin untersucht, ob gewisse Kombinationen auffällig häufiger auftreten als im Falle ihrer Unabhängigkeit zu erwarten, im Beispiel etwa die Endungen *-en, -er,* aber auch *-ten*. Diese werden zu mutmaßlichen Affixen erklärt, sofern sie gewisse formale und statistische Bedingungen erfüllen. Diese Affixe wiederum werden, wenn möglich, zu Paradigmen von Stämmen und Affixen erweitert. Nur wenn ein Paradigma gefunden wird, werden die Morphemgrenzen akzeptiert.

Es werden Kriterien gegeben, um Muster agglutinierenden Typs zu erkennen (wobei dann auch nach weiteren Affixen am selben Wort gesucht wird) und zwischen Flexion und Derivation zu unterscheiden.

Wie alle Methoden, die auf der Kookkurrenz benachbarter Buchstaben beruhen, steht das Andreev-Verfahren in der Tradition von Harris' Ansatz zur Morphgrenzenerkennung [4].

1.2 Einschränkungen dieses Ansatzes

Die Methode setzt Input mit gegebenen Wortgrenzen voraus. Um Sprachen ohne Markierung von Wortgrenzen oder gesprochenen Input zu behandeln, muß sie modifiziert oder mit einem eigenen Algorithmus zur Wortgrenzenerkennung kombiniert werden. Die Identifikation beschränkt sich auf identische Muster nur an Wortgrenzen. Daraus resultieren Schwierigkeiten beim Erkennen von Infixen,

mehrfachen gegenseitig abhängigen Affixen, Reduplikation, nicht-zusammenhängenden Morphemen wie z. B. bestimmten Vokalfolgen in den semitischen Sprachen oder nicht-konkatenativer Morphologie wie bei Umlautphänomenen.

Das Verfahren funktioniert am besten beim Auffinden von *item-and-arrangement*-Strukturen, wie sie für agglutinierende Konstruktionen besonders typisch sind.

2 Mögliche Anwendung

Die Arbeit des Algorithmus beschränkt sich auf die Erkennung von Flexions- und Derivationsaffixen. Darüber hinaus sind, wie bei jeder statistischen Methode, die Recall- und Precision-Raten begrenzt; daher ist das System nicht direkt verwendbar, um alle Morphemgrenzen eines bestimmten Typs aufzufinden.

Andererseits könnte es Hinweise darauf geben, ob bestimmte Modelle der menschlichen Sprachverarbeitung und des Spracherwerbs realistisch sind (vgl. etwa [3], [6]), und es könnte bei der Segmentierung und dadurch bei der Dechiffrierung einer unbekannten Sprache Anwendung finden.

3 Das Verfahren im einzelnen

Unter Auslassung technischer Details funktioniert der Algorithmus folgendermaßen:

1. Zunächst wird fortlaufender Text in eine Liste von Wörtern und ihren Häufigkeiten transformiert.
2. Nur Wörter ab einer bestimmten Mindestlänge werden als möglicherweise affixbehaftet in Betracht gezogen.
3. Für eine gewisse Anzahl initialer und finaler Wortpositionen werden die höherfrequenten Buchstaben untersucht, beginnend bei den am stärksten überrepräsentierten.
4. Diese Buchstaben werden sukzessive mit geeigneten Nachbarn verkettet, bis die Kette den Wortrand und die erwartete minimale Affixlänge erreicht.

5. Das „Affix“ wird probeweise abgeschnitten, es wird nach Kombinationen der resultierenden „Stämme“ mit anderen „Affixen“ gesucht.
6. Die letzten Schritte werden wiederholt, bis keine weiteren Affixe akzeptiert werden können.
7. Schließlich wird getestet, ob die Morphemgrenze nach links oder rechts verschoben werden sollte und ob es sich um ein Paradigma agglutinierenden Typs handelt.

Bei jedem dieser Schritte müssen gewisse statistische Kriterien erfüllt sein, damit die Arbeit des Algorithmus fortgesetzt wird.

Wenn ein Paradigma, ein *morphologischer Typ* gefunden wurde, werden seine Mitglieder aus der Liste aller Wörter entfernt (im agglutinierenden Fall nur die affixbehafteten Formen).

In der Originalarbeit von Andreev wird die weitere Bearbeitung des Textes nur mit den Nachbarn der vorher erkannten Wörter fortgesetzt, das heißt, syntaktische Information wird eingebracht. In der vorliegenden Arbeit beginnt die nächste Runde wieder mit der reinen, nur um die bisher analysierten Formen verminderten Wortliste.

Man kann, wenn alle morphologischen Typen gefunden sind, diejenigen mit ähnlichen (syntaktischen) Eigenschaften zu Wortklassen, wie Substantiven, Adjektiven und Verben, zusammenfassen. Da hier keine syntaktische Information benutzt wird, kommt eine Zusammenfassung nicht in Betracht.

4 Einstellung der Parameterwerte

Die Methode benutzt zahlreiche Parameter, deren Werte in den vorliegenden Arbeiten von Andreev nicht näher motiviert werden. Daher wurden Testläufe mit unterschiedlichen Parameter-Einstellungen auf Beispieltexten aus der Bibel und aus Computerforen durchgeführt. Die Bibelpassagen lieferten weit bessere Resultate, da sie sprachlich einheitlicher sind und viel weniger Tippfehler und ähnliche Quellen von Rauschen enthalten.

Der Einfluß von Parametern wurde untersucht, darunter

- die Anzahl der untersuchten Wortpositionen,
- die minimale Häufigkeit von Buchstaben, die in Betracht gezogen werden,
- deren minimaler Grad der Überrepräsentation,
- die minimale Rate der Kookkurrenz benachbarter Buchstaben, damit sie als Teil eines Affixes betrachtet werden.

Es stellte sich heraus, daß einige Parameter weit großzügiger gewählt werden können als von Andreev vorgeschlagen, d. h. so, daß wesentlich mehr Buchstaben und -kombinationen in Betracht gezogen werden. Dies führt zu größerem Recall ohne großen Verlust von Precision. Daß Andreev die Anforderungen eher hoch wählte, mag darin begründet liegen, daß zu seiner Zeit die Computer-Ressourcen sehr beschränkt waren, und die großzügigere Wahl der Parameter zu weit höherem Rechenaufwand führt. Tatsächlich wurden viele Ergebnisse in Andreevs Gruppe von Hand ermittelt. Im Gegensatz dazu wurden die hier vorgestellten Ergebnisse weitgehend automatisch auf einem kleinen Rechner erreicht.

Darüber hinaus dienen viele dieser Parameter dazu, mit einer hohen Wahrscheinlichkeit sicherzustellen, daß die in der Stichprobe gefundenen Abweichungen von der Gleichverteilung nicht zufällig sind, sondern Ausdruck einer Gesetzmäßigkeit. Diese Wahrscheinlichkeit steigt entsprechend dem statistischen Gesetz der großen Zahl bei gleicher Parametereinstellung auch mit der Stichprobengröße, daher kann man in größeren Stichproben wiederum geringere Abweichungen bereits als signifikant betrachten.

5 Ergebnisse in der Praxis

Die Methode mit den ermittelten optimalen Parametereinstellungen wurde auf den gesamten Text der Bibel angewendet. Die Bibel ist ein relativ freundliches Korpus, da sie ein begrenztes Vokabular mit wenigen Fremd- und sogar Lehnwörtern benutzt. Das Korpus umfaßt insgesamt 4.652.726 Bytes.

Die morphologischen Grenzen, die der Algorithmus ermittelte, wurden mit Morphemgrenzen verglichen, die vom Autor für jeden Wort-Type, ohne Berück-

sichtigung des Kontextes, intellektuell markiert wurden. Im Falle von Homonymie wurde ein Wort als flektiert betrachtet. Diese Vorgehensweise erscheint gerechtfertigt, da einerseits die statistische Methode ebensowenig zwischen homonymen Wörtern unterscheiden kann, andererseits auf diese Weise die Recall-Werte nur schlechter werden können, nicht besser.

Detaillierte quantitative Ergebnisse enthält die Tabelle 1, einen Überblick die Tabelle 2.

Das Ziel war, Flexionsparadigmen zu finden, nicht Derivationen. Daher bezeichnet der Recall den Anteil der intellektuell markierten Flexionsaffixe, die auch vom Algorithmus markiert wurden, die Precision den Anteil der vom Computer ausgegebenen Affixe, die nach menschlichem Urteil Flexionsaffixe darstellen.

Der Recall beträgt, in Types ausgedrückt, 42,4%, die Precision 89,6%. In Tokens gerechnet haben wir einen Recall von 73,8% und eine Precision von 96,4%.

5.2 Auswertung der Ergebnisse

Die statistische Natur des Algorithmus ergibt bessere Recall- und Precision-Werte bei hochfrequenten Wörtern. Dies ist der Grund für den großen Unterschied zwischen den Ergebnissen in Types und Tokens. Man vergleiche dazu bei den Wörtern mit Flexionsaffix die mittlere Häufigkeit der vom Algorithmus richtig identifizierten (25,4) mit derjenigen der nicht identifizierten (5,6), ersichtlich aus Tabelle 1.

Einige Besonderheiten der deutschen Sprache führen zu Fehlern. Diese sollen qualitativ analysiert werden.

1. Einige hochfrequente Stämme sind zu kurz, um vom Algorithmus in Betracht gezogen zu werden. Alleine das Wort *ein-e* ist für 17% aller fälschlich als nicht affixbehaftete Form (also als mutmaßlicher Stamm) erkannten Tokens verantwortlich.
2. Besonders Verben erscheinen oft mit verschiedenen Wortbildungspräfixen. Diese Kombinationen sind so vielfältig, daß sie leicht als Flexion mißinterpretiert werden können. So findet ein Testlauf des Algorithmus als 16. Paradigma die Stämme

Sorte	*Types*	*Tokens*
Alle Wörter	24255	708249
zu kurz (bis 3 Bst.)	348	294570
bleiben	23907	413679
als Angeh. v. Typen		
aussortiert	7224	192242
unanalysiert	16674	221437
aussortiert	7224	192242
als affixlos	378	30480
als affixbehaftet	6846	161762
affixlos	378	30480
korrekt	211	20636
eigtl. flektiert	167	9844
affixbehaftet	6846	161762
tats. flektiert	6137	155871
Wortbildung	680	5990
Komposition	27	78
Zufall	2	21
unanalysiert	16674	221437
davon mit Affix	8170	45444
unflektiert	8504	175993

Tabelle 1: *Detaillierte Auswertung für das Korpus Bibel*

Tokens	*insges.*	*richtig erkannt*	*nicht erkannt*	*falsch erkannt*
flekt.	14474	6137	8337	167
unflekt.	9424	211	8504	709
Tokens	*insges.*	*richtig erkannt*	*nicht erkannt*	*falsch erkannt*
flekt.	211159	155871	55288	9844
unflekt.	202520	20636	175993	5891

Tabelle 2: *Gesamtresultat für das Korpus Bibel*

-fahren, -genommen, -gezogen, -werfen mit den dazugehörigen Affixen *ab-, auf-, aus-, vor-, weg-*, das als Flexion eingestuft wird, und es folgen in kurzer Folge 11 weitere Paradigmen dieses Typs.
Damit nicht genug, werden die so falsch erkannten Wörter dann auch noch aus dem Korpus entfernt, so daß ihre tatsächlichen Flexionsaffixe nicht gefunden werden können.
Bei 95,9% der fälschlich als flektiert identifizierten Wörter wurde ein Wortbildungsaffix fehlinterpretiert. Gar nicht um ein Affix handelte es sich nur bei 0,4% der Types und 0,06% der Tokens.

3. Aus dem eben genannten Grund ist die Unterscheidung von Flexion und Derivation anhand der statistischen Eigenschaften im Deutschen nicht sehr zuverlässig.

4. Im Deutschen werden über alle Wortklassen hinweg homonyme Flexionssuffixe verwendet. 13 Suffixe und 2 Zirkumfixe (die beide eines der 13 Suffixe enthalten) erfüllen unter großer Überschneidung und Synkretismus alle Funktionen der Substantiv-, Adjektiv- und Verbflexion ursprünglich deutscher Wörter. So sind die drei Affixe *-en, -e, -Ø* in Substantiv-, Adjektiv- und

Verbparadigmen gemeinsam anzutreffen. Viele Suffixe bestehen zudem aus hochfrequenten Buchstaben, die auch am Ende von Stämmen häufig sind. Das führt zu einer Vermischung solcher Fälle mit Stämmen, die nur mit unvollständigem Paradigma im Korpus vorkommen, und dadurch zu vereinzelten willkürlichen Morphemgrenzen, deren Zahl im Test allerdings gering bleibt (s. Tabelle 1).

5. Stammumlaut ist ein häufiges Phänomen in deutschen Flexionsparadigmen. Dies führt im besten Fall zu einer Aufspaltung der beteiligten Paradigmen, ansonsten zur Nichterkennung der umgelauteten Formen.

Zieht man diese Tücken in Betracht, so erscheint die Zahl der korrekt identifizierten Wortformen mit Flexionsaffix von rund drei Viertel aller Tokens beeindruckend, wenn auch nicht ausreichend für praktische Anwendungen. Überdies kann man mit wachsender Korpusgröße auch ein weiteres Ansteigen dieser Rate erwarten.

6 Erweiterungsmöglichkeiten

Interessant wäre, das Verfahren auf phonematisch statt graphematisch transkribiertem Text zu testen. Dies könnte eine größere Regelhaftigkeit aufdecken, andererseits könnten aber auch gewisse Verallgemeinerungen verlorengehen, da die deutsche Orthographie teils grammatikalisch motiviert ist.

Die Voraussetzungen des Verfahrens sind sehr restriktiv. Es könnte in Richtung einer flexibleren Mustererkennung erweitert werden, um den Suchbereich auf Infixe und nicht-zusammenhängende Morpheme zu erweitern. Möglicherweise könnten statt buchstäblich übereinstimmenden Mustern ähnliche gesucht werden, obwohl fraglich ist, ob Ähnlichkeit von Morphen angemessen definiert werden kann, zumal einzelsprachunabhängig.

Statt das Verfahren zu ändern, damit es klassischen Definitionen von Morphemgrenzen entspricht, könnte man auch eine empirischere, mehr naturwissenschaftliche Sichtweise von Sprache einnehmen und es selbst als Definition solcher Grenzen ansetzen.

Literatur

[1] Andreev, Nikolaj D. (ed.): Statistiko-kombinatornoe modelirovanie jazykov, Moskau/Leningrad 1965

[2] Andreev, Nikolaj D.: Statistiko-kombinatornye metody v teoretičeskom i prikladnom jazykovedenii, Leningrad 1967

[3] Bybee, Joan L.: Morphology as Lexical Organization. In: Hammond, Michael/Noonan, Michael (ed.): Theoretical Morphology. Approaches in Modern Linguistics, San Diego 1988

[4] Harris, Zellig: From phoneme to morpheme. Language 31 (2), 1955, p. 190—222

[5] Harris, Zellig: A Theory of Language and Information. Oxford 1991

[6] MacWhinney, B. (ed.): Mechanisms of Language Acquisition, Hillsdale, N.J. 1987