COLING 82, J. Horecký (ed.)
North-Holland Publishing Company

ZUM WIEDERAUFFINDEN VON INFORMATIONEN IN AUTOMATISCHEN WÖRTERBÜCHERN

Gunter Neubert

Sektion Angewandte Sprachwissenschaft
Technische Universität Dresden
Dresden
DDR

Aus automatischen Wörterbüchern kann oft nicht so viel Information zurückgewonnen werden wie aus herkömmlichen gedruckten Wörterbüchern. Das ist u. a. auf Wortbildungssynonymie zurückzuführen, die mindestens folgende Ursachen hat: unterschiedliche lexikalisch-morphematische Elemente, unterschiedliche Konstruktion und unterschiedliche Motivation von Wortbildungsprodukten. Es erscheint dienlich, Wege zur intellektuellen und programmgesteuerten Synthese synonymer Abfragewörter zu suchen.

EINFÜHRUNG

Unter einem "automatischen Wörterbuch" wollen wir einen technisch beliebig gearteten Speicher lexikalischer (speziell fachlexikalischer/terminologischer) Informationen verstehen, den ein Übersetzer auf mittelbare Weise, d. h. unter Zwischenschaltung eines Suchprogramms, befragt, um Kenntnisse über die für die Übersetzung zu wählenden Äquivalente lexikalischer Einheiten (oder dessen, was er dafür hält) in der Zielsprache zu erlangen. Alle anderen möglichen Verwendungszwecke eines derartigen Speichers, wie maschinelles Übersetzen, Erlernen von Sprachen, Gewinnen systematischer Terminologieübersichten, Aufbau von Informationsrecherchesprachen u. a. m., seien hier unberücksichtigt.

Automatische Wörterbücher sind heute wohl durchweg als Speicher elektronischer Datenverarbeitungsanlagen ausgeführt; sie werden auch "Terminologiedateien" oder "terminologische Datenbanken" genannt. Zu ihrer Befragung stehen Programmpakete unterschiedlichen linguistischen und rechentechnischen Komforts bereit. Der Übersetzer fragt entweder im Dialogverkehr über Schreibmaschinentastatur und Bildschirm oder im Stapelbetrieb mit zeitlich verzögerter Listenausgabe.

SUCHPROBLEME

Der einfachste Suchalgorithmus beruht auf dem zeichentreuen Vergleich des Abfragewortes[1] mit den Wörtern[1] (derselben Sprache) im Speicher. Es leuchtet ein, daß ein solcher Suchalgorithmus eine nur geringe effektive Antwortquote[2] q_e bieten kann, werden doch selbst bei Vergleichspaaren wie verbaler/substantivierter Infinitiv ('sandstrahlen'/'Sandstrahlen') oder Zusammen-/Getrenntschreibung ('slideway'/'slide way') keine Antworten geliefert. Die Unterdrückung von Groß-/Kleinschreibung, Zusammen-/Getrenntschreibung und bestimmter Sonderzeichen wie des Bindestrichs ('Zwei-Richtungs-Verkehr'/'Zweirichtungsverkehr') erhöht q_e nur unwesentlich, denn Vergleichspaare

wie 'Gasaufkohlen'/'Gasaufkohlung', 'Pfeilverzahnung'/'pfeilverzahnt' u. ä. bleiben weiter unerschließbar.

Zur Beseitigung dieses Nachteils des automatischen gegenüber dem gedruckten Wörterbuch wurden verschiedene Verfahren entwickelt, wie die der längstmöglichen Übereinstimmung oder des Abschneidens von Endzeichen(ketten) nach Vergleich mit einer Liste zulässiger bzw. potentieller Endzeichen(ketten). Für ihr befriedigendes Funktionieren sind aber stets empirische Zusatzregeln und/oder -informationen erforderlich. Andere Abfrageverfahren, angewandt z. B. bei der Dresdner Terminologiedatei EWF[3], beziehen den Abfragenden ein, indem sie ihm die zusätzliche Angabe der Anzahl der vergleichsrelevanten Zeichen abverlangen. Alle diese Verfahren vermögen q_e erheblich zu erhöhen, liefern mitunter sogar viel zu viel Information, sie sind jedoch teils schwer theoretisch fundierbar, teils aufwendig in der Anwendung.

Eine der Ursachen für geringe Antwortquote kann darin bestehen, daß die Arbeitsvorschriften für die lexiko-/terminographische Kompilation nicht ausreichend an den Suchalgorithmus angepaßt sind, der für das zu schaffende automatische Wörterbuch vorgesehen ist. Die übliche Einsparung sprachlich regulärer Abwandlungsformen beispielsweise, die für die Benutzer des gedruckten Wörterbuchs kaum Schwierigkeiten mit sich bringt, darf bei programmgesteuerter Abfrage trotz Erhöhung des Kompilations- und Speicheraufwands nur so weit getrieben werden, wie es der Algorithmus zuläßt.

Eine weitere, weitaus weniger leicht zu beherrschende Ursache liegt in der Synonymie, die trotz aller Bemühungen um terminologische Standardisierung nicht beseitigt ist und unseres Erachtens auch nicht beseitigt werden kann, da sie zum Teil prinzipielle, sprachimmanente Gründe hat. Bei der Erörterung von Wegen, die terminologische Synonymie in den Griff von Algorithmen zu bekommen, gehen wir von der realen Situation aus, daß als "Adresse" für die Abfrage von Wortstellen vom Übersetzer im allgemeinen keine metasprachliche Begriffsbeschreibung benutzt werden kann, sondern die im Quellentext vorgefundene (oder eine ähnliche) Benennung (meist lemmatisiert) herangezogen werden muß. Damit folgt die prinzipielle Nichterschließbarkeit lexikalischer Synonymie wie 'Akrolein'/'Propenal', 'Saflor'/'Färberdistel'[4] oder 'Schraube'/'Bolzen'. Als erschließbar könnten jedoch Synonymiefälle ins Auge gefaßt werden, die durch Anwendung unterschiedlicher Wortbildungsregeln auf ein und denselben Begriff entstanden sind. Um das Ausmaß dieses Problems der "Wortbildungssynonymie" anzudeuten, sei ein Beispiel angeführt, bei dem die Wortbildungsverfahren zu einer besonders hohen Anzahl von Varianten führen. Für ((ein Mittel, das die Bildung eines die Durchsichtigkeit beeinträchtigenden Films z. B. auf einer Glasscheibe verhindert bzw. vermindert)) wurden verschiedenen Wörterbüchern und der Fachliteratur folgende Benennungen entnommen und durch einige weitere regelgerechte Varianten ([c]) ergänzt; Vollzähligkeit, falls überhaupt möglich, wurde nicht angestrebt, es könnten z. B. noch Fremdwörter benutzt werden:

Antibeschlagmittel
[c]Antischleiermittel
Beschlagfreiausrüstung
[c]Beschlagverhinderungsmittel
Beschlagschutzmittel
[c]Entschleierer
Entschleierungsmittel
Klarsichtmittel
Mittel gegen Verschleierung
schleierdämpfendes Mittel
Schleiergegenmittel
Schleiermittel
Schleierschutzmittel
schleierverhinderndes Mittel
[c]Schleierverhinderungsmittel
schleierverhütendes Mittel
schleierwidriges Mittel
Verschleierungsschutzmittel

WEGE ZUR WORTBILDUNGSSYNONYMIE

Zur Grundlegung von Algorithmen, mit denen Synonyme dieser Art als zusätzliche Abfrageadressen gewonnen werden können, müssen die Wege untersucht werden, die zur Wortbildungssynonymie führen.

Die Synonymenvielfalt läßt sich auf mehrerlei Weise ordnen. Ein erstes Ordnungskriterium ergibt sich daraus, daß die einzelnen Sprachen häufig mehrere gleichberechtigte Konstruktionsweisen bereitstellen. Das beginnt bei der Behandlung der Fuge im Deutschen: Zwischen (älterem) 'Aschfall', 'Aschenfall' und (neuerem) 'Aschefall' bestehen semantisch keine Unterschiede. Oft spielt es keine Rolle, ob der Verbalstamm oder das Verbalsubstantiv auf '-ung' zur Wortzusammensetzung herangezogen wird ('Wiederholgenauigkeit' = 'Wiederholungsgenauigkeit'). Am anderen Rande des Bandes konstruktiver Variabilität stehen Konstruktionssynonymenpaare aus Zusammensetzung o. ä. und Wortgruppe ('Handschleifen'/'Schleifen von Hand', 'Finite-Elemente-Methode'/'Methode der finiten Elemente'). Ohne feinere Unterschiede zu berücksichtigen, ordnen wir die Benennungen des Beispiels:

- K1. Zusammensetzung o. ä.
  - Beschlagschutzmittel
  - Beschlagverhinderungsmittel
  - Klarsichtmittel
  - Schleiermittel
  - Schleierschutzmittel
  - Schleierverhinderungsmittel
  - Verschleierungsschutzmittel
- K2. Verwendung eines Präfixes o. ä.
  - Antibeschlagmittel
  - Antischleiermittel
  - Entschleierer
  - Entschleierungsmittel
  - Schleiergegenmittel
- K3. Verwendung eines Suffixes o. ä.
  - Beschlagfreiausrüstung
  - schleierwidriges Mittel
- K4. Wortgruppe (auch mit zusammengesetzten Teilen)
  - Mittel gegen Verschleierung
  - schleierdämpfendes Mittel
  - schleierverhinderndes Mittel
  - schleierverhütendes Mittel

Eine zweite Möglichkeit der Ordnung folgt daraus, daß vom Benennenden verschiedene lexikalisch-morphematische Elemente benutzt werden, ohne daß er an Bedeutungsdifferenzierung des Ganzen denkt:

- L1. für die Erscheinung, gegen die das Mittel wirken soll
  - Beschlag
  - Schleier
- L2. für die Wirkungsrichtung des Mittels

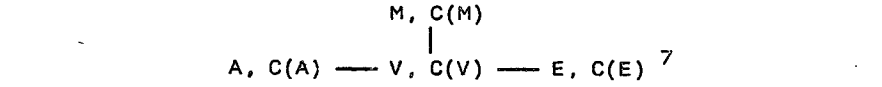
anti-	gegen-
dämpfen	schützen
ent-	verhindern
-frei	verhüten
gegen	-widrig

- L3. für das Mittel
  - Ausrüstung
  - Mittel

Bei der Durchsicht der Varianten fallen zwei Bildungen - 'Klarsichtmittel' und 'Schleiermittel' - auf, für deren Deutung als reguläre Wortbildungsmöglichkeiten ein drittes Ordnungskriterium herangezogen werden muß: Welche der Elemente des Begriffs sind in die Benennung aufgenommen worden? Das ((Mittel)) wird spezifiziert durch

- M1. den Vorgang bzw. das Ergebnis, der bzw. das verhindert werden soll + Wirkungsrichtung des Mittels
  Antibeschlagmittel
  Antischleiermittel
  ...
  Verschleierungsschutzmittel
- M2. den Vorgang bzw. das Ergebnis, der bzw. das verhindert werden soll (ohne Ausdruck der Wirkungsrichtung des Mittels)
  Schleiermittel
- M3. den Vorgang bzw. das Ergebnis, der bzw. das herbeigeführt werden soll
  Klarsichtmittel

Obwohl alle Synonyme im linguistischen Sinne motiviert sind, d. h. ihre Bestandteile etwas über die Bedeutung des Ganzen aussagen, sind Maß und Art dieser Aussage verschieden. Die Motivation als Selektions- bzw. Eliminierungsprinzip für Elemente des Begriffs läßt sich verallgemeinert fassen, da sich einige allgemeingültige Regeln herausgebildet haben[5], die vom allgemeinen Betrachtungsstandpunkt der Vertreter der einzelnen Fachgebiete bedingt sind. Für das Fachgebiet der Technik lassen sie sich aus dem Wesen der Technik ableiten:[6] Bei einem technischen Vorgang V handelt es sich um eine Tätigkeit, die vom Menschen mit Hilfe eines technischen Mittels M ausgeführt wird, um ein Ausgangsobjekt A so zu verändern, daß ein verwertbares Ergebnisobjekt E entsteht. Die Tätigkeit des Menschen ist begleitet von der Erfassung und Beschreibung der Eigenschaften C der Vorgänge und Gegenstände. Die durch Großbuchstaben gekennzeichneten Kategorien verkörpern gleichsam die begrifflichen Beziehungen, die der Motivation bei fachlexikalischen Wortbildungsprodukten zugrunde liegen. Sie lassen sich wie folgt als System darstellen:



[7]

das die funktionsorientierte Sicht des Technikers wiedergibt. Nun schließt die Regularität der Motivation aber durchaus ein, daß unterschiedliche Benennungen entstehen können, denn die Selektion begrifflicher Elemente wird über den fachlich allgemeinen Standpunkt hinaus von den speziellen Betrachtungsstandpunkten der verschiedenen Vertreter eines Fachs bedingt. Wir wollen das an einem Beispiel zeigen, das gleichzeitig andeuten soll, daß die Motivationssynonymie auch einzelsprachliche Aspekte haben und Unterschiedlichkeit der Standpunkte auch unterdrückt werden kann. Ein englischsprachiger Fachmann der Hydraulik/Pneumatik, der sich mit der Untersuchung von Filtern beschäftigt, wird für ((das im Filter befindliche Element, das das Filtern bewirkt)) wahrscheinlich 'filtering element' bilden; der Verwender, der die Bauteile des Filters bei der Zerlegung zur Säuberung kennenlernt, bildet 'filter element'. Im Deutschen entstehen infolge der Gleichheit der wortbildungswirksamen Formen von 'filtern' und 'Filter' keine verschiedenen Benennungen, sondern jeweils 'Filterelement', und die speziellen Betrachtungsstandpunkte werden (im vorliegenden Fall) unterdrückt.[8]

Übrigens zeigt die Durchsicht von Fachwortschätzen, daß der von

einem Mittel herbeizuführende bzw. positiv zu beeinflussende Vorgang den Normalfall der Motivation darstellt und die Synonymenanzahl in unserem Beispiel nur deshalb so groß ist, weil sich für negative Wirkungsrichtung keine Vorzugswortbildungsregel ausgeprägt hat.

BEDINGUNGEN DES ERSCHLIESSENS VON WORTBILDUNGSSYNONYMIE

Die Synthese von Wortbildungsprodukten muß, gleichgültig ob sie intellektuell durch den Abfragenden oder programmgesteuert erfolgt, die oben aufgezeigten Wege nachvollziehen, die die originäre Wortbildung beschreitet. Dafür müssen folgende Voraussetzungen vorhanden sein:

- Es müssen die Regelwerke der Motivation in den verschiedenen Fachgebieten und der Formverarbeitung bei der Wortbildung, und zwar bezogen auf die auszudrückenden begrifflichen Elemente, für die einzelnen Sprachen algorithmierbar bereitstehen.
- Es müssen die lexikalisch-morphematischen Bausteine der Wortbildung verwendungsäquivalent gruppiert sein.

Um diese Voraussetzungen zu schaffen, braucht die linguistische Forschung nicht etwa grundsätzlich neue Forderungen zu erfüllen; sie muß jedoch ihre Erkenntnisse zur Lösung der Aufgaben der automatischen Verarbeitung von Sprache aufbereiten. Es zeigt sich allerdings rasch, daß diese Aufbereitung sowohl von Kenntnislücken als auch von sprachlichen Problemen behindert wird. Insbesondere sind die Motivationsprinzipien der einzelnen Fachgebiete einschließlich der potentiellen Betrachtungsstandpunkte weitgehend unerforscht, aber auch die konstruktiven Wortbildungsregeln sind nicht ausreichend den einzelnen Ausdrucksbedürfnissen zugeordnet. Die Gruppierung der lexikalisch-morphematischen Elemente macht eine spezielle Synonymieauffassung erforderlich. Zwischen den verschiedenen Wortbildungsregeln und -mitteln bestehen zudem zahlreiche wechselseitige Bedingtheiten, die empirisch zusammengetragen werden müssen.

Andererseits offenbaren sich Probleme, deren Lösung wohl vorwiegend oder ausschließlich auf der Ebene der fachlichen Inhalte erfolgen kann. Dazu rechnen wir viele Fälle des Definitionsstandpunktswechsels, z. B. auch den Übergang zum Ausdrücken des herbeizuführenden anstelle des zu verhindernden Vorgangs wie im Beispiel. Außerdem werden durch fachinterne Gepflogenheiten z. B. beim Lexemgebrauch Variationsmöglichkeiten eröffnet bzw. ausgeschlossen, was sich einer Fassung in Regeln und damit einer Algorithmierung zu entziehen scheint.

Die Sprachwissenschaft sollte sich jedoch vom Bemühen, der sich entwickelnden "Sprachtechnik" zu helfen, durch Unvollkommenheit ihrer Erkenntnisse nicht abhalten lassen, sondern zweierlei in Betracht ziehen: Ein Synthesealgorithmus für programmgesteuerte Wörterbuchabfrage darf von einem Grundsatz ausgehen, der beispielsweise bei der linguistischen Fundierung des Sprachunterrichts geradezu entgegengesetzt gilt - sobald nur die automatische Verarbeitung schnell genug geschieht, spielen etwa erzeugte ungebräuchliche oder fehlerhafte Wortbildungsprodukte keine Rolle, denn der Abfragende bemerkt sie nicht. Zweitens ist der Abfragende in der Lage, die Verarbeitungsergebnisse kritisch zu bewerten, bevor er sie weiterverwendet. Damit wollen wir gleichzeitig anmerken, daß wir beim gegenwärtigen Stand nur von interaktiven Systemen - Systemen, in denen sich Mensch und Maschine wechselseitig ergänzen - brauchbare Ergebnisse erwarten. Das trifft wahrscheinlich noch mehr auf die völlig ausgesparte

Problematik der automatischen Analyse des Abfrageworts hinsichtlich seiner verschiedenen Strukturen zu.

ANMERKUNGEN

1 Wir benutzen 'Wort' mit der pragmatischen Definition als "vom Übersetzer als abfragbar bewertete sprachliche Einheit".

2 Unter Antwortquote q verstehen wir allgemein das Verhältnis der Anzahl der nicht abschlägig beantworteten Anfragen zur Gesamtanzahl der Anfragen. Zu unterscheiden sind: die effektive Antwortquote q_e, in die alle Antworten einbezogen sind, die nützliche Information enthalten; die Wortantwortquote q_w als das Verhältnis der Anzahl der gefundenen zur Gesamtanzahl der abgefragten Wörter; die Begriffsantwortquote q_b analog hinsichtlich der Begriffe. Beim zeichentreuen Vergleich gilt $q_e = q_w$, da genau nur die Wörter erkannt werden, die abgefragt worden sind. Gelingt es, alle Synonyme der Abfragewörter zur Abfrage zu nutzen, wird $q_e = q_b$. q_e kann weiter gesteigert werden, wenn darüber hinaus unterstützende Informationen aus weiteren lexikographischen Einheiten des Speichers bezogen werden können.

3 Diese Datei ist ausführlich beschrieben in Neubert, G., Kukuczka, H., Meyer, E., Das Datenverwaltungssystem für Fachwortschätze EWF, in: Neubert, G. (Hrsg.), Rechnerunterstützung bei der Bearbeitung fachlexikalischer Probleme (VEB Verlag Enzyklopädie Leipzig, 1981).

4 Für den ersten Fall existiert eine metasprachliche Beschreibung in Gestalt der chemischen Symboldarstellung: $CH_2{=}CHCHO$; für den zweiten könnte die botanische Nomenklaturbenennung Carthamus tinctorius L. als solche angesehen werden. Ihre Nutzung als Abfrageadresse setzt voraus, daß sie dem Übersetzer bekannt sind.

5 Ausführlich dazu Reinhardt, W., Produktive Wortbildungsmodelle im technischen Fachwortschatz des Deutschen, Diss. B, Päd. Hochsch. Potsdam, 1973.

6 Vgl. hierzu Neubert, G., Das Motivationsprinzip bei fachlexikalischen Einheiten am Beispiel des Wortschatzes der Hydraulik/Pneumatik, Diss. B, Techn. Univers. Dresden, 1978.

7 Das System ist nicht vollständig dargestellt; s. dazu die Literatur in den Anmerkungen 5 und ausführlicher noch 6.

8 Nur zur Verdeutlichung: Wäre im Deutschen im Fachgebiet der Hydraulik/Pneumatik 'filtrieren' anstelle 'filtern' üblich, ergäben sich 'Filtrierelement' und 'Filtrierereelement'.