



Wortschatzquantifizierung der französischen Tagespresse 2005: Wieviele Wörter braucht der Mensch ?

Bernd Sebastian Kamps und Stephan Kamps

ABSTRACT

Steinhäuser Internet (St.K) und
Flying Publisher (B.K.)

Wörtermühle 2005; P2.

Copyright © 2005, Flying Publisher.

HINTERGRUND

Die Sprache der Presse ist durch Fernsehen und Internet ubiquitär und für jeden leicht erreichbar. Der Umfang des Wortschatzes, den ein Student im Jahre 2005 beherrschen muß, um französische Tages- und Wochenzeitungen zu verstehen, ist nicht bekannt.

METHODE

Vom 19. August bis zum 26. September wurden 47 Textproben aus der französischsprachigen Tagespresse analysiert und unbekannte Wörter in einer Wörterbuchtafel erfaßt. Immer dann, wenn 500 neue Wörter aufgenommen waren, wurden jeweils drei Texte von den Internet-Titelseiten von Le Monde, Libération und Le Figaro auf die Zahl der bekannten Wörter untersucht

ERGEBNIS

Bei 1000, 2000 und 3000 Wörterbucheinträgen betrug der Prozentsatz der bekannten Wörter 24,8, 37,6 und 48,1%.

SCHLUSSFOLGERUNG

Der zum Verständnis der Tagespresse notwendige Wortschatz scheint umfangreicher zu sein als vielfach angenommen wird.

Computertechnologie und Internet haben in den letzten Jahren den Zugang zu fremdsprachlichen Texten vereinfacht. Tageszeitungen, Wochenzeitschriften und Fernsehstationen bieten auf ihren Webseiten täglich eine Fülle von aktuellen Texten an. Die Texte können leicht kopiert und weiterverarbeitet werden können, und so eröffnet das Internet zahlreiche Anwendungsmöglichkeiten für den Fremdsprachenunterricht.

Die Sprache der Presse ist zwar nur eine der möglichen Sprachebenen und Sprachstile, im Vergleich zu anderen Stilvarianten ist sie in der heutigen Zeit aber präsender. Sie könnte daher ein wichtiges Hilfsmittel für den Einstieg in eine Fremdsprache sein.

Es ist wenig bekannt darüber, wieviele Wörter notwendig sind, um die Tagespresse in einer Fremdsprache zu verstehen. In den 70er Jahren versprach eine Grundwortschatzserie ein Verständnis von 80% eines durchschnittlichen Textes mit 2000 Wörtern und 95% mit 4000 Wörtern. Manche Sprachlehrer veranschlagen heute noch niedrigere Werte, doch sind diese Aussagen zuweilen dürftig belegt. Wir beschlossenen daher, eine Datenbank mit Texten der französischen Tages- und Wochenpresse aufzubauen und zu prüfen, mit wievielen Wörtern welcher Prozentsatz eines Textwortschatzes erkannt werden.

Methoden

Herkunft der Texte

Die Texte entnahmen wir den Internetnet-Titelseiten von Tages- und Wochenzeitungen aus Frankreich, Belgien und der Schweiz (Tabelle 1). Das wichtigste Auswahlkriterium war die Bekanntheit des Inhalts für den deutschen Leser. Aus diesem Grund wurden vorzugsweise Texte zu Deutschland oder zu überregionalen Ereignissen ausgewählt.

Tabelle 1. Tages- und Wochenzeitschriften, deren Texte in die Auswertung einbezogen wurden

La Libre Belgique
Le Figaro
Le Monde
Le Soir
Libération
Nouvel Observateur
Tribune de Genève

Datentabellen

Wir erfaßten die Wörter in zwei MySQL-Tabellen, eine Worttabelle und eine Wörterbuchtafel. Die erste Tabelle – die **Worttabelle** – enthielt Einträge für die Wörter so wie sie im Text erschienen, also auch in deklinierter oder konjugierter Form. Pro Datensatz wurden zwei weitere Felder erfaßt: *notshown* und *dic_id*. Notshown erhielt den Wert 1, wenn das Wort ein sogenanntes Nullwort war, dessen Übersetzung nicht angezeigt werden sollte. **Nullwörter** waren im wesentlichen Personalpronomen, Possessivpronomen, die Präsenzformen der Hilfsverben, Zahlen und Eigennamen (siehe Tabelle 2). Notshown hatte den Wert 0, wenn das Wort ein **Lernwort** war. Dieses wurde dann über das Feld *dic_id* mit der Wörterbuchtafel verbunden.

Tabelle 2. "Nullwörter"

1.	a, au, auquel, aurait, autre, autres, aux, avait, avec, avons, à, ce, cela, celle, celles, celui, ces, cet, cette, ceux, ça, dès, de, des, deux, dont, du, elle, elles, en, est, et, eux, été, était, il, ils, je, la, laquelle, le, lequel, les, leur, leurs, lui, me, ne, ni, non, nos, notre, nous, où, on, ont, ou, par, pas, plus, que, quel, qui, sa, se, sera, serait, ses, si, son, sont, tous, tout, toute, toutes, un, une, vous, y
2.	Eigennamen
3.	Zahlen

Die Worttabelle wird mit dem folgenden SQL-Befehl kreiert:

Wieviele Wörter braucht der Mensch ?

```
CREATE TABLE `words` (  
  `notshown` tinyint(2) NOT NULL default '0',  
  `dic_id` double default NULL,  
  `word` varchar(254) default NULL,  
  PRIMARY KEY (`word`)  
) TYPE=MyISAM;
```

Die zweite Tabelle, die **Wörterbuchtafel**, hat ebenfalls drei Felder: id, source, language1. Über das erste Feld ist die Tabelle mit dic_id der ersten Tabelle verbunden; in den beiden anderen Feldern werden die Grundform des französischen Wortes (source) sowie die deutsche Entsprechung (language1) gespeichert. Der SQL-Befehl zur Anlage der Wörterbuchtafel lautet:

```
CREATE TABLE `dictionary` (  
  `id` double NOT NULL auto_increment,  
  `source` varchar(254) NOT NULL default "",  
  `language1` text,  
  PRIMARY KEY (`id`)  
) TYPE=MyISAM AUTO_INCREMENT=0;
```

Datenerfassung

Für die Datenerfassung nutzen wir eine PHP-Software, die wir im Herbst 2004 für die arabische Sprache entwickelt hatten (www.Arabic4Europeans.com). Erfasst wurden die Daten durch ein Programm, das die einzelnen Wörter auf dem Bildschirm einblendete. Schon bekannte Wörter, die in den ersten vier Buchstaben mit dem neuen Wort übereinstimmten, zeigte das Programm ebenfalls an (Abb. 1).

Für jedes Wort bestanden drei Optionen:

1. Das Wort war ein Nullwort (Pronomen, Zahlwörter, Eigennamen etc.) und war nicht für die Aufnahme in die Wörterbuchtafel vorgesehen. In diesem Fall wurde nur die Checkbox angeklickt (siehe Abbildung 1).
2. Das Wort war dem System bekannt (Beispiel: prochain in Abb. 1), aber unter einer anderen Form (z. B. prochains, prochaine, prochaines). In diesem Fall war es ausrei-

chend, die ID-Nummer von prochain – 1276 – einzugeben.

3. Das Wort war dem System nicht bekannt. In diesem Fall wurden die Grundform des französischen Wortes und die deutsche Entsprechung erfasst.

Umfang der Eingaben

In der ersten Ausbauphase des Projekts erfaßten wir keine Redewendungen. Das Layout der deutschen Entsprechungen ist schicht gehalten. Wann und in welchem Maße die Wörterbuchdatei ausgebaut wird, hängt davon ab, wie intensiv das System im Internet genutzt wird.

importance	<input type="checkbox"/>	1189 important wichtig, bedeutend; umfangreich
		981 imposer durchsetzen, druchdrücken; aufbürden, auferlegen; besteuern; durcksetzen, durchdrücken; imponieren; verzinsen; vorschreiben
		1268 impossible unmöglich, ausgeschlossen
prochain	<input type="checkbox"/>	1276 prochain 1. m: Mitmensch; der Nächste - 2. adj: nächster

Abbildung 1: Eingabemaske für neue Wörter

Datenerfasser

Alle Texte wurden von einem der beiden Autoren bearbeitet (BSK), um eine größtmögliche Konsistenz der Daten zu gewährleisten.

Datenanalyse

Immer dann, wenn 500 neue Wörter in die Wörterbuchtafel aufgenommen waren, wurden jeweils drei Texte von den Internet-

Titelseiten von Le Monde, Libération und Le Figaro auf die Zahl der bekannten Wörter untersucht (Auswertung A). Bei jeder folgenden Untersuchung wurden auch die Texte der vorherigen Tests einbezogen (also 9 Texte bei 500 Wörtern, 18 Texte bei 1000 Wörtern, 27 Texte bei 1500 Wörter usw.).

Die Ergebnisse von Auswertung B sind in Abbildung 3 dargestellt.

Diskussion

(folgt)

Zudem wurden vor der Bearbeitung eines neuen Textes die folgenden Daten erhoben (Auswertung B):

- Gesamtzahl der Wörter
- Zahl der Nullwörter und der Lernwörter
- Zahl der bekannten Wörtern unter den Lernwörtern (den „Nicht-Nullwörtern“)

Publikation im Internet

Die bearbeiteten Texte wurden im Internet unter der Adresse www.Woertermuehle.de publiziert.

Ergebnisse

Die Daten von Auswertung A sind in Tabelle 3 und Abbildung 2 dargestellt.

Der Anteil der Nullwörter blieb über die Zeit relativ konstant. Unter den verbleibenden Lernwörtern (56 bis 58% der Texte) stieg der Anteil der bekannten Wörter in Abhängigkeit von der Zahl der Wörterbucheinträge stetig an. Bei 1000, 2000 und 3000 Wörterbucheinträgen betrug der Prozentsatz der bekannten Wörter 24,8, 37,6 und 48,1%.

Tabelle 3. Erkannte Wörter in Abhängigkeit von den Einträgen in der Wörterbuchtabelle *

Wörter in Tabelle	Geprüfte Wörter	Nullwörter (%)	Erkannte Wörter (%)
500	6312	n.b.	14
1000	11945	n.b.	24,6
1500	19143	41,9	31,7
2000	26414	42,6	37,6
2500	32403	43,1	43,0
3000**	38746	43,3	48,1

* Auswertung mit beliebigen Texten aus drei Tageszeitungen – n.b. = nicht bestimmt

** geschätzte Werte nach der 5. Auswertung

Wieviele Wörter braucht der Mensch ?

Abbildung 2: Bekannte Wörter in Abhängigkeit von der Zahl der Wörterbucheinträge - Auswertung in 500-Wort-Abständen

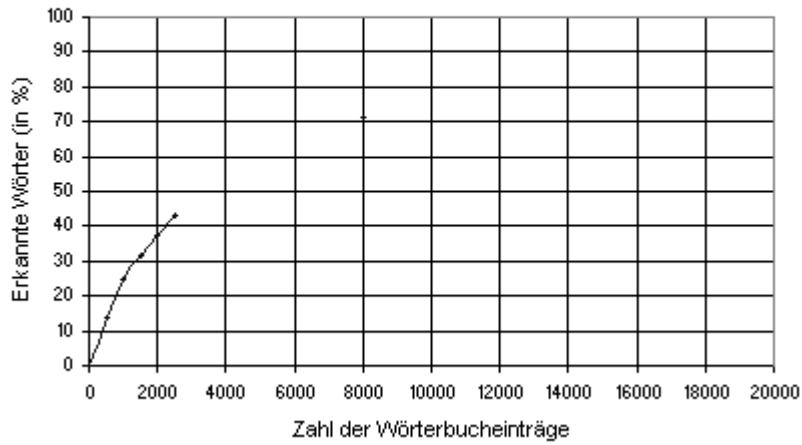


Abbildung 3: Bekannte Wörter in Abhängigkeit von der Zahl der Wörterbucheinträge - Auswertung vor Aufnahme neuer Texte

