In proceedings from *the XVI:th Scandinavian Conference of Linguistics* , Department of Linguistics, University of Turku. 1998

Some Frequency based Differences between Spoken and Written Swedish

Jens Allwood

1. Introduction

This is a report on the differences in word frequency found between two Swedish corpora, - a transcribed spoken language corpus of 276,391 words and a written language corpus of 271,216 words. The spoken language corpus contains material from 14 activity types while the written language corpus contains material from novels (40%) and newspapers (60%). The report expands and continues earlier work on differences between spoken and written language, e.g. Jörgensen (1976) or Biber (1988).

The word frequencies of the two corpora are described and more fully presented in Allwood 1996. Below I will now present some of the main differences between the corpora.

2. Word types and word occurrences

2.1 Word types on different frequency levels

A first observation is that the ratios between types and tokens in the two corpora are very different. In the spoken corpus there are 276.391 tokens and 18.406 types while there are 271.216 tokens and 39.638 types in the written language corpus. There are, thus, twice as many types in written as in spoken language. It seems that in speech we only use around 50% of vocabulary we use in writing.

In Table 1, the distribution of word types on different levels is shown for the corpora. The rows in the table show the relative corpus share (% of the total number of word tokens) for the types

given in rank order, starting with the ten most frequent types in the first row.

Table 1 The corpus share of word types on different frequency levels

Frequency based rank	Speech	Writing
1-10	23.3%	19.4%
11-50	28.7%	18.8%
51-100	10.3%	6.8%
101-1000	22.2%	20.7%
1001-10000	12.5%	21.3%
10001-18406 (speech)	3.0%	
10001-39638 (writing)		13.0%

The table shows that the ten most common types make up 23.3% of the spoken corpus and 19.4% of the written corpus. The fifty most common types make up 52% of the spoken corpus and 38.2% of the written corpus. The hundred most common types make up 67.3% of the spoken corpus and 45% of the written corpus. The thousand most common word types make up 84.5% of the spoken and 65.7% of the written corpus.

One way of summarizing the above is to say that the most common words are more common in speech than in writing. Using the 50 most common words, one can understand 52% of all words that are uttered but only 38% of all words that are written. However, this does not mean that 50 words would be sufficient to understand corresponding shares of spoken or written language in real use, since this would, in addition, require that one understands how the frequent words are related to other, perhaps not as frequent, words. This will usually be impossible without also understanding these other words. It is, thus, clear that there are limits for how far one can go with the 50 most common words. However, it should be clear that it is a good idea to include the 50 or 100 most common words in different practical circumstances such as language teaching or speech recognition.

2.2 From Speech to Writing

In our transcriptions we have used **M**odified **S**tandard **O**rthography (MSO), a standard for the transcription of spoken Swedish which is somewhat closer to real spoken language pronunciation than the standard orthograpy of Swedish. However, it is not as detailed as a phonetic or phonematic transcription would be. In MSO, standard orthography is used unless there are several spoken language variants corresponding to a given standard form. According to this principle, the standard word form **jag** (I) is written both as **jag** (I) and **ja** (I) and the standard form **det** (it) both as **det** and **de** since both pronunciations exist in spoken language. The number of spoken language variants that are distinguished is, to a certain extent, arbitrary and has therefore ultimately been decided stipulatively. We have for example not distinguished the forms [**de**] and [**de**:] from [**d**_], [**d**_:], which would have been possible.

Often several written language words correspond to one spoken language form. Thus, the spoken form **ja** corresponds both to **jag** (I) and to **ja** (yes) in written language. When this happens, the spoken language forms have been disambiguated with numerical indexes so that the degree of disambiguation which is present in written language can be preserved in spoken language. We have, however, not separated any homonyms over and above the level of standard orthography. It is therefore not possible to know whether the Swedish word **springa** is a noun or a verb. In some cases, however, spoken language introduces a disambiguation not present in written language, eg between **att** (that) as a subordinating complementizer (mostly pronounced **att**) and **att** as an infinitival marker (mostly pronounced **å**).

Let us now, in tabular form, consider the relation between some spoken language words transcribed in MSO without disambiguation and the written language words corresponding to them. Since the words are often polysemous, the translations given in all tables below only represent some of the main functions of the words. The cases where we have not been able to decide what the corresponding form is have been marked as ”unclear”.

Table 2 The 10 most frequent words in undisambiguated spoken language with corresponding words in standard orthography. All frequencies derive from spoken language

Undisambig. speech		Corresponding standard orthographical form	Freq.	Undisambig. speech		Corresponding standard orthographical form	
1.	d 17207	det (it)	16936	5.	så 6472	så (so)	6463
	e	den (it)	145			sådan (such)	4
		de (they)	117			sådant (such)	3
		dig (you)	3			såg (saw)	1
		unclear	6			unclear	1
2.	ja 10625	jag (I)	5987	6.	att 5579	att	5579
		ja (yes)	4626				
		unclear	12	7.	va 4722	var (where was)	1607
3.	e 9108	är (is)	6690			vad (what)	1293
		eh (eh)	2380			va (what)	1268
		en (an)	15			vara (be)	544
		ett (an)	10			vart (where)	5
		det (it)	9			varje (every)	2
		unclear	4			unclear	8
4.	å 7700	och (and)	6689	8.	på 4043	på	4043
		att (that)	958				
		å (oh))	34	9.	som 3866	som	3866
		ja (yes)	6				
		på (on)	4	10.	man 3794	man	3974
		om (if)	3				
		unclear	5				

2.3 From writing to speech

Also when we consider the ten most frequent words of written language and their correspondents in spoken language, we find a great deal of variation.

Table 3. The 10 most frequent words in written language with their spoken language correspondents. The frequencies derive separately from the spoken and written corpus.

Written			Spoken		Written			Spoken	
		Frequency		Frequency			Frequency		Frequency
1.	och (and)	8387	å och o	6689 1133 2 7824	6.	en (a, an, one)	4638	en e	3585 1 3586
2.	i (in)	7683	i	3696	7.	på (on)	4161	på å	4056 4 4060
3.	att (that, to)	6638	att å a	5580 958 54 6592	8.	är (is)	3923	e ä är	6693 1066 328 8087
4.	det (it)	5887	de det re dä e d di	16936 381 68 34 9 4 1 17433	9.	med (with)	3243	me me d mä be	1734 477 226 1 2243
5.	som (that, which)		som se	3877 2 3879	10.	av (of by)	3133	av a	1026 33 1059

The variation seems to be of roughly the same size as when going from speech to writing. Many words which are differentiated in writing correspond to only one spoken form and many words which are differentiated in speech correspond to only one written language form. In some cases, semantic pragmatic differences can be connected with the differences between speech and writing. It seems likely, that the fact that both **och** (and) and **att** (that, to) can be pronounced as **å** could be taken to indicate that the two

words have a joint function in speech. In the same way the fact that written **att** (that, to) can be pronounced both as **å** and as **att** could show that written **att** has two functions which are differentiated in speech. However, for the majority of cases such explanations can probably not be found and all we can say is that we are dealing with a number of conventionalized spoken language forms which have developed in order to reduce the burden on articulation in on-line spoken communication.

3. The 10 most frequent words in speech and writing.

Let us now see what we can learn by comparing the 10 most frequent words in speech and writing.

Table 4 The 10 most frequent words in speech and writing

Rank	Speech	Frequency	Writing	Frequency
1.	de (det) (it)	16898	och (and)	8378
2.	e (är) (is)	6666	i (in)	7683
3.	å (och) (and)	6645	att (that, to)	6638
4.	så (so)	6424	det (it)	5887
5.	ja (jag) (I)	5977	som(that,which)	4808
6.	att (that, to)	5579		4638
7.	ja (yes)	4475	en (a, an, one)	4161
8.	på (on)	4043	på (on)	3923
9.	som	3866	är (is)	3243
10.	(that,which) man (one)	3794	med (with) av (of, by)	3133

A similarity between speech and writing is that the 10 most frequent words in both cases are syncategorematic or so called function words, ie words that are used to connect and structure utterances in speech and sentences in writing. As a matter of act 95% of all the most frequent words in both speech and writing are syncategorematic (cf Allwood 1996). We use these words continuously in order to structure what we are saying or writing.

The table further suports the observation made above that highly frequent words in speech have a higher frequency than highly frequent words in writing. Turning to the words in the table, we can see that the words **så** (so), **ja** (I), **ja** (yes) and **man** (one) are among the 10 most frequent in speech but not in writing. Conversely, the words **i** (in), **en (**a, an, one), **med** (with) and **av** (of, by) are among the 10 most frequent in writing but not in speech. The words **de** (det) (it), **e** (är) (is), **å** (och) (and), **på** (on) and **som** (that, which) are among the most frequent both in speech and in writing.

Some further differences between speech and writing are: There are two pronouns **ja** (jag) (I) and **man** (one) among the 10 most frequent in speech but not in writing. Both are, in Swedish spoken language, used to refer to the speaker. See also Dahl 1995. Another highly frequent word of spoken language is the feedback word **ja** (yes). The same word has rank 137 in writing. We also find the words **de** (det) (it) and **så** (so). Both of these words can be used for deictic as well as for well anaphoric reference. Of these two, only **det** (it) has a high frequency in writing. But it only has 30% of the frequency it has in speech. Among the highly frequent words of written language, we find conjunctions like **och** (and) and **att** (that, to) as well as prepositions such as **i** (in), **på** (on), **med** (with) and **av** (of, by). The reason they are more common in writing than in speech, is probably that they are needed to construct a complex phrase and sentence structure which is more typical of written than of spoken language.

4. A Comparison of Parts of Speech in Spoken and Written Language

4.1 Introduction

Investigating the distribution of different parts of speech (p o s) in the two corpora gives us another oportunity to shed light on the nature of the differences and similarities between speech and writing. The results I will present are based on a coding of p o s directly on lists of words from the two corpora without taking

context into consideration in order to disambiguate the words (cf Allwood 1996). The result of such a coding is a set of words with the p o s labels that can potentially be connected with them. Many words are given several p o s labels. Words such as **unga** (young) and **gamla** (odd) are, for example, assigned both the label adjective and the label noun since they, in Swedish, can be used in both ways. As a result, the percentages for the relative shares of different p o s will usually add up to more than 100%.

The parts of speech which were used are relatively traditional with the following changes: Modifying ordinals and cardinals were counted as adjectives, while nonmodifying uses were counted as pronouns. Thus, ordinals and cardinals will potentially be both adjectives and pronouns. Articles were counted as pronouns. Prepositions were also counted as adverbs since they in Swedish can function also as particles for verbs. Two new parts of speech which are characteristic of spoken language were introduced:

(i) FB: Feedback words, ie words which signal contact, perception understanding and reaction to a preceding utterance (cf Allwood 1988 or Allwood, Nivre and Ahlsén 1992).
(ii) OCM: Words used for Own Communicaiton Management, ie words which are used to gain time, for example to hesitate, plan or to indicate that you want to change something you have said (cf Allwood, Nivre and Ahlsén 1990).

The coding is not completely finished. Words with a frequency which is lower than 4 have not been coded for spoken language and words which have a frequency lower than 10 have not been coded for written language. Continued coding will probably primarily increase the shares of nouns and verbs, on the type level and to a lesser extent on the occurrence level.

4.2 Parts of Speech in spoken and written language - an overview

In table 7 below, are indicated the relative shares of the parts of speech in spoken and written language. In relation to what was stated in section 4.1 and presented in Allwood 1996, I have modified the definitions and results further in the direction of the traditional parts of speech, in order to make them easier to understand. This means that words which primarily function as prepositions have not been counted as adverbs and words which primarily function as adjectives have not been counted as nouns. If these changes were not made, adverbs rather than verbs would be second most frequent in spoken language and nouns rather than conjunctions would be fourth most frequent. In the written language corpus, nouns would be second most frequent and adverbs third most frequent.

Table 5 Potential Parts of Speech
(Note: adverbs do not contain words which are primarily prepositions and nouns do not contain words which are primarily adjectives)

Speech		Share of corpus	Share of types		Writing	Share of corpus	Share of type
1	Pron	25.8	2.1	1	Pron	19.5	1.6
2	Verb	20.5	12.4	2	Verb	18.0	6.4
3	Adv	13.6	2.6	3	Prep	13.0	0.2
4	Conj.	11.8	0.3	4	Noun	10.8	9.2
5	Prep	8.5	0.7	5	Conj.	9.4	0.1
6	FB	6.5	0.7	6	Adj	8.4	5.0
7	Adj	6.4	7.6	7	Adv	5.7	0.8
8	Noun	5.9	13.0	8	FB	0.8	0.1
9	OCM	1.7	0.2	9	Interj	0.03	0.03
10	Interj	0.2	0.1	10	OCM	0.01	0.1
11	Not coded	4.8	56.1	11	Not coded	15.4	73.1
No. of word tokens:		276.391			271.216		
No. of word type:		18.406			39.638		

Table 5 shows that the share of uncoded words is 4.8% in the spoken corpus and 15.4% in the written corpus. On the type level, this corresponds to 56.1% and 73.1% uncoded types. This means that in both corpora, but especially in the written language corpus, there is a large number of word types which have low frequencies. Probably these word types are nouns and verbs which, however, already with the present coding, have the largest type shares.

The table, thus, shows that categorematic words, primarily nouns, verbs and adjectives, ie words which are not function words, make up a majority of all coded word types. For speech, they constitute 75% and for writing 76.5% of all types. This should be compared with their shares on the token level which are only 34.5% and 43.9%, respectively. Nouns, verbs and adjectives are apparently more used in writing than in speech in spite of the fact that their share of types is similar in the two corpora. In general, we can note that with only 25% of our word types, we can produce 65% of what we say and 56% of what we write. Turning to the relative size of the different p o s, we can see that pronouns are the most frequent p o s in both speech and writing. We also see that their frequency is higher in speech. Even though the pronouns make up only about 2% of all word types, 25% of what is said and 20% of what is written can be expressed with pronouns. Probably the 6% size difference between the shares of pronouns in speech and writing can be related to the fact that speech is more bound and integrated with context than writing. By using pronouns it is possible to make use of contextual information in speech where one prefers or has to be more explicit and use full nominal descriptions in writing.

Verbs are important in both speech and writing and have the second largest share in both corpora. The fact that the type share of verbs is larger for speech than for writing probably depends on the fact that a larger part of the spoken corpus has been coded.

Adverbs are the third largest p o s in speech and the seventh largest in writing and thus seem to play a more important role in speech than in writing. Probably, this is related to the fact that adverbs like pronouns often are dependent on and allow for an integrated use of contextually given information.

Prepositions are the third largest p o s in writing (speech fifth) and conjunctions are the fourth largest p o s in speech (writing fifth). What this might show is that conjunctions are a more important means of linking in speech than in writing. In writing, prepositions are more important. This claim is also supported by the analysis of the 10 most frequent words which was presented in section 3.

The next big difference concerns the occurrence of FB (feedback words) and OCM (own communication management words). These words almost exclusively occur in speech. This is hardly surprising since both word types are functionally connected with direct on-line interaction which is carried out more often in speech than in writing.

Turning to nouns and adjectives, we see that nouns occupy fourth position in writing but only eighth position in speech. We also see that adjectives are somewhat more common in writing than in speech (speech position 7, writing position 6). Probably this is related to the fact that(Swedish) written language has a strongly nominalizing character.

The high frequency for nouns, adjectives and prepositions, in written language, all point in this direction. Spoken language (Swedish), on the other hand, seems to make less use of nominalizations and instead to be more oriented toward using and highlighting relations to other (mostly previous) utterances, and to context in general, which can be done through use of pronouns, adverbs, conjunctions, feedback words and words for own communication management.

5. Conclusions

The following similarities and differences between speech and writing have been found.

Similarities

1. Word types are mostly the same in speech and writing. When they are not, they are usually easy to relate to each other.

2. The most frequent words (95%) in both speech and writing are syncategorematic or function words, ie words which are needed to structure the content or to connect to the situation.
3. Pronouns and verbs are the most frequent p o s in both speech and writing which shows, that in spite of differences in degree, the need for contextual anchoring and dynamics is considerable in both forms of language.

Differences

1. The number of word types is smaller in speech than in writing. In speech we use roughly 50% of the word types we use in writing.
2. Common words and collocations are more common in speech than in writing, ie expressions which have a high frequency have a relatively speaking higher frequency in speech than in writing. Highly frequent words also occur in highly frequent collocations to a greater extent in speech than in writing.
3. As a consequence of points 1) and 2) the space for variation (for example to meet the needs of different individuals or activities) ought to be smaller in speech than in writing. We all seem to speak more similarly (at least on a word or phrase level) than we write. It will be a future task to investigate whether this hypothesis is correct.
4. The most common words in speech show that more contextual information is used than in writing. They also show that talk is more anchored in the immediate participants of talk – the speaker ((**jag** (I), **vi** (we) and **man** (one)) and - the listener (**du** (you) and **ni** (you)). In writing, more commonly a 3rd person perspective (**han** (he), **hon** (she), **den** (it), **det** (it), **de** (they) etc) is used instead.
5. In speech, there are a number of features such as feedback, own communication management and use of interjections which are needed since speech is interactive and on-line. Writing instead exhibits features which can be associated with a monological use of language.
6. As a consequence of points 3, 4 and 5, pronouns, conjunctions , adverbs, feedback words, OCM words and interjections are

used more in speech, while prepositions, nouns and adjectives are used more in writing.

References

Allwood, J. (1988) Om det svenska systemet för språklig återkoppling. In P. Linell, V. Adelswärd, T. Nilsson, & P.A. Pettersson (Eds.) Svenskans Beskrivning 16, Vol.1. (SIC 21a), University of Linköping, Tema Kommunikation.

Allwood, J. (1996) Talspråksfrekvenser. Gothenburg Papers in Theoretical Linguistics, University of Göteborg, Dept of Linguistics.

Allwood, J., Nivre, J., & Ahlsén, E. (1989) Speech Management - On the Non-Written Life of speech. Gothenburg Papers in Theoretical Linguistics 58, University of Göteborg, Dept of Linguistics.

Allwood, J., Nivre, J., & Ahlsén, E. (1991). On the Semantics and Pragmatics of Linguistic Feedback: Gothenburg Papers in Theoretical Linguistics 64, University of Göteborg, Dept of Linguistics.

Biber, D. (1988). Variation across speech and writing. Cambridge: Cambridge University Press.

Dahl, Ö. Unpublished MS. Dept of Linguistics, University of Stockholm.

Jörgensen, N. (1976). Meningsbyggnader i talad svenska. Lund