Swedish and Danish, spoken and written language

A statistical comparison

Peter Juel Henrichsen
Copenhagen Business School

Jens Allwood
University of Gothenburg

The aim of much linguistic research is to determine the grammar and the lexicon of a certain language L. The spoken variant of L – in so far as it is considered at all – is generally taken to be just another projection of the same grammar and lexicon. We suspect that this assumption may be wrong. Our suspicion derives from our contrastive analyses of four corpora, two Swedish and two Danish (covering spoken as well as written language), suggesting that – in the dimensions of frequency distribution, word type selection, and distribution over parts of speech – the mode of communication (spoken versus written) is much more significant as a determining factor than even the choice of language (Swedish versus Danish).

Keywords: language comparison, spoken language corpora, speech versus writing, Danish versus Swedish, Scandinavian languages

1. Introduction

Politically, Danish and Swedish are considered to be two different languages; indeed (and perhaps for this reason) many Swedes declare themselves unable to understand Danish, and vice versa. At the same time, spoken Danish and written Danish are taken to be one language as a matter of course (similarly for Swedish). Our corpus data tell quite another – and perhaps surprising – story.

International Journal of Corpus Linguistics 10:3 (2005), **367–399.**
ISSN 1384–6655 / E-ISSN 1569–9811 © John Benjamins Publishing Company

Previous investigations of the differences between the two languages have been based on intuitions and comparison of traditional grammars. Quantitative corpus-based approaches have, so far as we are aware, not been tried. In this paper, we will make use of this kind of approach by comparing:

(i) a spoken language corpus for Danish
(ii) a spoken language corpus for Swedish
(iii) a written language corpus for Danish
(iv) a written language corpus for Swedish

The reason for the employment of both spoken and written corpora is that previous work has shown considerable differences between spoken and written Swedish (Allwood 1998) and between spoken and written Danish (Henrichsen 2002b). The differences are, in fact, so significant that they raise the question of whether intralinguistic variation between the spoken and written variants of a language might actually be greater than the difference between what has traditionally been regarded as different languages. As we will see below, this might indeed be the case. The spoken languages of Swedish and Danish, in many ways, seem to be more similar than spoken Danish or Swedish are to written Danish or Swedish.

In our comparison we will take variation between social activities into account by matching similar activities in spoken Danish and spoken Swedish. To be more explicit, we will match Danish informal interviews with Swedish informal conversations and interviews. Likewise, in order to keep invariant the style of writing, we will match Danish newspaper texts with Swedish newspaper texts.

The implications of our observations are at the moment not entirely clear, but they may have practical as well as theoretical consequences important in such areas as conversational training, studies in intercultural communication, and localization of language technology.

As we shall see, certain parts of speech are typical of the spoken variants of the Scandinavian languages (and for English and other languages as well), notably interjections, pronouns, attitudinal adverbs, and conjunctions. Those parts of speech are not always studied as intensively as e.g. nouns, verbs, and prepositions in language courses, and so one might speculate that fluency in conversation could be improved by rehearsing the related communicative functions, such as feedback (attention, accept, reluctance, etc.), own communication management (hesitation, resumption, repair), attitude (empathy, scepticism, irony), and pronominal reference (keeping track of the discourse referents in a spoken narration) – as a supplement to more traditional language exercises focussed on grammatical constructions and categories typical of the written language.

In Scandinavia today, speech technological projects are almost exclusively intra-national.[1] Adaptation of written language dictionaries and grammars for use in speech synthesis and recognition is commonplace in our countries and, we suspect, in other small language areas around the world too. Given the many common features that we find in the spoken variants of the Scandinavian languages – and the relatively large discrepancies between spoken and written styles within each language – we speculate that Denmark, Sweden, and other small countries might benefit from founding a speech technology on advanced components developed for neighbouring languages rather than on existing national dictionaries and grammars based on written text sources.

The organization of the paper is as follows. After a short discussion of the linguistic background of our investigations, we present our four reference corpora in Section 3, while in Section 4 we comment on our methodology. Section 5 presents the details of our contrastive analysis of the Danish corpora, one written and one spoken, and in Section 6 we present – in a condensed form – a Swedish investigation along similar lines. In Section 7 our findings are summed up and a general picture is drawn of spoken and written language, and finally in Section 8 we suggest some guidelines for further research.

English translations of Scandinavian word types are given in a reduced form in the main text; for fuller translations, see the Appendix.

2. Background

Structuralist linguistics for a long time has favored (sometimes explicitly but perhaps mostly implicitly) the view that the difference between spoken and written language is of minor importance to linguistic theory. With the work reported in this paper, we wish to examine this belief more critically. Is the difference between spoken and written language really without theoretical importance? Let us consider some reasons why this view might be challenged. A basic reason is that spoken language has evolutionary primacy and probably is genetically facilitated. Unless we assume very rapid genetic change, this is not the case for written language.

Another reason is that the structure of spoken and written language, although similar in some respects, is also very different in many respects. Face-to-face spoken language is interactive (in its most basic form), multimodal (at the very least containing gestures and talk) and highly context dependent. Further, it is basically organized into utterances which are often no longer than a word. Written language, on the other hand, in its most typical form is non-interactive, monological and monomodal with a lesser degree of contextualization, organized in sentences which are governed by normative rules of the type that a proper sentence should contain a subject and a predicate. The norms of spoken language are usually not of this sort, rather they concern intelligibility and adequacy in different social activities.

As part of the mechanisms that make spoken language into the efficient and finely tuned means of communication that it is, we find ways of changing your mind or ‘own communication management’ (for example, what from a normative written language perspective might be called ‘disfluencies’, ‘false starts’, ‘self repair’ etc.). We also find short and unobtrusive ways of giving feedback (for example, by words like *yeah* and *uhuh*) while overlapping with another speaker’s utterance. None of these phenomena that are typical and central to the functionality of spoken language have any place in written language.

The differences between spoken and written language have previously been discussed by, for example, Allwood 1998, Biber et al. 1999 and Leech et al. LRW2001. Generally, we may say that estimates about the importance of the distinction vary, from holding that it is merely a genre difference, similar to the difference between texts from novels and texts from newspapers (Biber 1988), to claiming that the difference is of a more radical nature, such as McKelvie (1998), Debaisieux and Deulofeu (2001). Our belief is that the differences are fairly significant and that their true nature, to a considerable extent, has remained hidden because most research that has been done, has been focussed only on those aspects of spoken and written language which are comparable. What this means is that since it is unsurprising that spoken language does not contain punctuation marks, this feature is often not even noted as a significant difference. In fact, the most common meaningful signs in written language are indeed ‘,’ (comma) and ‘.’ (period). The term ‘word’ is avoided here, in order not to block the comparison. Similarly, very significant features of spoken language, such as overlap between speakers, own communication management and feedback are absent in written language and therefore left out of the comparison. The missing types of comparison become even more evident, if we bring in body movements and various types of gestures, which are also a *sine qua non* of face-to-face spoken language communication, but absent in written language.

In our investigation, to be presented below, we will, however, restrict ourselves to features that exist in both spoken and written language and we will argue that even with this restriction, the differences that can be found are considerable.

3. Data

In this section we present the linguistic data of our investigation together with our analytical methods.

3.1 Four reference corpora

Our comparison is mainly based on four corpora referred to as DanSPO, SweSPO, DanWRI, and SweWRI. To enhance comparability, in line with what was noted above, these corpora were all adjusted by removing all non-lexical markup (such as punctuation in written corpora and details of pronunciation, pauses etc. in speech). Each reference corpus consists of orthographically controlled words only, organized with one sentence per line in the written corpora and one utterance per line in the spoken corpora.

DanSPO is identical to the Danish speech corpus BySoc, established in the late eighties in the socio-linguistic project "Bysociolingvistik" (The Copenhagen Study in Urban Sociolinguistics). It consists of so-called Labovian interviews (Labov 1984), i.e. informal conversations without preset topic – about

Table 1. Composition of reference corpora

Reference corpus	DanSPO	SweSPO	DanWRI	SweWRI
Size (words)	1,335,247	380,338	1,334,944	785,986
Style	Labovian interview (informal conversation)	Informal interview and informal conversation	Mixed newspaper genres	Mixed newspaper genres
Source corpus	BySoc	GSLC	Berlingske 99 (Danish daily newspaper)	Göteborgsposten 2001 (Swedish daily newspaper)
Selection	All of BySoc	Gbg-fragment of GSLC	Fragment of text body (articles only)	Fragment of text body (articles only)

80 in total, mostly recorded in the informants' own homes. BySoc is described in Gregersen et al. (1991), Henrichsen (1998), and is available at www.id.cbs.dk/~pjuel/BySoc

SweSPO is identical to the gbg-fragment of the Göteborg Spoken Language Corpus (GSLC). GSLC was mainly recorded in the period 1978–2000 as part of many different projects. GSLC contains 1.3 million word tokens organized in around 25 sections containing different social activity types such as auction, patient-doctor consultation, and shopping (Allwood 1999; Allwood et al. 2002a; Allwood et al. 2002b). The gbg-fragment – alias SweSPO – consists of informal interviews and informal conversations.

As seen, the spoken corpora were collected with different purposes. We have sought to keep constant the activity influence on language style in our investigation by selecting for SweSPO the segment of GSLC containing informal interviews and conversations, since (only) this style matches the style of BySoc with respect to number and kind of participants (linguist+informant) and purpose (free-style conversation).

3.2 DanSPO

DanSPO consists of all files in BySoc sliced one utterance per line. An 'utterance' is defined as a sequence of lexical words delimited by any of these events: pause (notated `£`, `££`, `£££` for normal/long/very long pause), hesitation with phonation (~), audible breathing (#), non-verbal communication (e.g. laughter, notated "`(ler)`"), turn shift, incomplete words (interruption point marked with "`-`"), partly unintelligible passage (transcription enclosed in square brackets). Other kinds of sound-related information (e.g. rising/falling intonation, hesitation with phonation, prolonged syllable) are ignored, as are passages marked by the transcriber as being atypical (e.g. read-

```
A> aha ££ jamen hvor- hvor-~ hvord- hvordan hvordan~
1>
------------------------------------------------
A> skete det havde jeg nær sagt (ler) hun blev
1>                           ja (ler) sådan
------------------------------------------------
A> hun blev gravid som syttenårig
1> [ mente du det nok ]         (ler) #
```

Figure 1. Sample from corpus BySoc in so-called 'score-format' showing the onsets of the interviewer (A>) and the informant (1>) relative to each other.

aha	*aha*
jamen	*but*
hvordan hvordan	|:*how*:|
skete det havde jeg nær sagt	*did it happen I almost said*
ja	*yes*
hun blev hun blev gravid som syttenårig	|:*she became*:| *pregnant aged 17*
sådan mente du det nok	*that's probably what you meant*

Figure 2. The corpus sample from Fig. 1 is here shown in DanSPO format (left column). English glosses are given in the right column. |: xyz :| means xyz repeated.

aloud text, foreign language, etc.). Figure 1 shows how the original transcription text was transformed to the present corpus DanSPO.

The sample in Figure 1 is rendered in DanSPO as shown in Figure 2.

3.3 SweSPO

SweSPO consists of all words in the gbg-fragment of corpus, one utterance per line.

The transcriptions of GSLC can be rendered in several formats depending on the closeness desired to standard orthography. The present study employs GSLC with transcriptions in standard Swedish orthography (excluding punctuation), being the style most equivalent to BySoc.

The GSLC transcription format thus differs somewhat from the BySoc format. Among the differences is the representation of overlapping utterances (demarcated in GSLC with square brackets and in BySoc by using a relative time axis) and extra-lexical information (rendered in separate lines in GSLC, while interspersed in the transcription in BySoc). In the corpora used in this paper we have abstracted away from such differences retaining the lexical word forms only; compare Figure 2 and 4.

```
$B: hon är min maka
$A: < ja >
@ < giggling >
$A: ja ja ja det är bara att fylla i det här
$B: ja
$A: [19 ska väl inte va ]19
$B: [19 hur ofta ]19 träffas ni / ganska sällan
```

Figure 3. Sample from GSLC. The transcription format includes information on relative timing of utterances (pauses, points-of-interruption, overlaps etc.).

hon är min maka	*she is my partner*
ja	*yes*
ja ja ja det är bara att fylla i det här	|: *yes*:| *it's just to fill in this*
ja	*yes*
ska väl inte va	*should not be*
hur ofta träffes ni	*how often do you meet*
ganska sällan	*fairly seldom*

Figure 4. Sample from Figure 3 shown in SweSPO format (left column). English glosses are in the right column. |: xyz :| means xyz repeated.

See Allwood et al. (2002b) for a contrastive analysis of the transcription formats of BySoc and GSLC.

3.4 DanWRI and SweWRI

DanWRI and SweWRI are copied from large newspaper corpora, as mentioned above. Punctuation is removed, and sentence-initial capital letters are lowered except for lexically governed capitalizations (proper names, certain pronouns and abbreviations).

4. Methodology

As we intend to compare words in different languages and different modes of communication, our project involves translation as well as transcription. Neither activity can be claimed to be semantically neutral. Translating – or transcribing – a word potentially changes its meaning. How, then, can we expect our comparisons to be meaningful? Aren't we comparing apples to oranges?

4.1 Comparing äpplen and æbler

Cognate languages often have 'false friends' – words that are etymologically related and phonetically similar, and yet do not mean the same. One example is the Swedish-English pair *även* – *even*, "även" meaning *also, too, likewise* rather than *even*.

False friendship is also widespread between Danish and Swedish. The lexeme spelled "rolig" means *quiet* in Danish, while *amusing* in Swedish; "spring" is *run* in Swedish, but *jump* in Danish etc., and such superficial similarities often mislead inexperienced translators. In other cases, a certain semantic do-main is structured differently in the two languages. For example, Swedish "kusin" (*cousin*) is ambiguous in Danish between "fætter" (*male cousin*) and "kusine" (*female cousin*). Like English, Swedish has no single term translating Danish "fætter" while Danish lacks a collective term for "fætter"∪"kusine".

Of course, as long as the semantic conflicts are as clear-cut as in these examples, they can be controlled using bilingual dictionaries. However, many semantic displacements are so tiny or subtle that not even the most advanced of dictionaries are aware of them. In Sweden, for instance, the term "mjölk" (*milk*) in general has the default reading *whole milk* which is what you'll get in the dairy shop if you simply ask for mjölk. Danish "mælk" (*milk*) has no similar default. Ask for mælk, and you'll probably be met by the question: whole milk or low-fat?

Semantic displacements,[2] large and small, are likely to pervade any translation list, and the lack of semantic control is hence intrinsic to all projects involving translation. On the other hand, this does not mean that dictionaries are meaningless things. Language users that are firmly rooted in two languages, often have precise and inter-subjective judgments in questions of adequate and inadequate translations. What is important to remember is simply that preferred translations should be understood as being *best possible* rather than *exact*.

In our project, the problems of semantic displacement has to be considered and, if possible, quantified. As we are comparing varieties of language organized along two orthogonal axes – viz. national language and mode of communication – we must consider in which dimension the problems can be expected to be greatest.

Even very extensive monolingual dictionaries do not usually distinguish between the meanings of written and spoken realizations of a word, or do so for a small number of lexemes only. Likewise, linguistic literature on word semantics usually does not state explicitly whether the claims and observations made count for spoken or for written language. It hence seems to be a common understanding among linguists that, concerning word semantics, the mode of communication is not of great importance. Also the fact that children can learn to write within a few years is easier to explain assuming that they, at least by default, reuse the semantics of the words they know. So even if neither of these arguments are completely conclusive, it is fairly uncontroversial to claim that changing the mode of communication leaves the semantic content of lexemes largely intact.[3]

In contrast, we know for a fact that the transition from language to language does imply semantic displacements, as exemplified above and amply documented in bilingual dictionaries.

Given these facts and assumptions, we still believe our experimental setup to be meaningful. Our main hypothesis is that – in the statistical dimensions we are studying – the choice of mode (written versus spoken) is more significant than the choice of tongue (Danish versus Swedish) as a determining factor. Assuming that the meaning of words is less well preserved under translation than under transcription, alignments of Swedish and Danish words can be expected to be less equivalent, more 'noisy', than alignments of written and transcribed words within the same language. It could be argued, then, that this will render any result in support of our main hypothesis even more significant; if two hunters compete and the one with the bent gun wins, this certainly adds to his achievement. However, we do not want to press this point too hard, and for now we just observe that our assumptions concerning comparability of languages and modes of communication are generally shared among linguists.

4.2 Three dimensions of comparison

We have chosen to compare our corpora using three different statistical approaches,

- frequency distribution
- word type ranking
- distribution over parts of speech

Since these three levels of description are largely independent of each other, tendencies observable at more than one level are particularly significant. We will expand on this in the sections to follow.

In including descriptions of the **frequency distribution** we adhere to the Zipfian tradition. George Kingsley Zipf claimed that certain distributional patterns are universal, i.e. are found in any (large) sample of any language (Zipf 1936). An important aspect of Zipf's programme is the demonstration that languages can be compared based solely on the frequency distributions in text collections, one advantage being that arbitrarily different languages become directly comparable since frequency distribution functions make no reference to the actual inventory of word types.

Word type ranking refers to the ordered frequency lists of word types. Questions to be addressed at this level include: Which word type is the overall most frequent in each corpus? To what degree are the lists of frequent types shared among the corpora (modulo word-to-word translation)? Does the corpus suite subdivide naturally according to the individual word type preferences? If yes, which corpora prefer the same types? As the answering of these questions does require translation as well as transcription, we must not forget the risk of semantic displacement when interpreting the results.

Parts of speech (POS) distribution concerns parts of speech rather than types. For the bilingual comparisons, we settled on a smallish tagset of 7 tags corresponding to the traditional major parts of speech. The tagset will be presented and commented on in later sections.

In comparing the POS-distributions we apply methods similar to those of the word type selections, except that the analytical objects are POS tags rather than word types. The central questions of this session include: which tag is the most frequent in each corpus? How are the tags distributed in general? Does the corpus suite subdivide naturally based on the POS preferences? If yes, which corpora agree?

Again a caveat is in place. Automatic tagging is of course fast and convenient, but not without its problems. Our taggers were trained mainly on written text sources, so the quality of the DanSPO and SweSPO tagging cannot be expected to match that of the written corpora. More importantly, however, the very idea of analyzing speech in categories developed for written text is questionable. Our tagset (and most other POS-tagsets) lacks labels for sound and vision based features such as intonation, stress pattern, pause distribution, turn shift, facial expression, and gestures – features playing an essential role in spoken communication. So in general, comparative studies of spoken and written corpora based on standard POS-tagging are blind to certain aspects of spoken language expressivity and must be careful not to jump to wrong conclusions concerning the diversity of expression. One of our reasons for including this type of analysis in the present study is that we wish to be able to relate to similar investigations in other languages – English in particular. Word-to-word translation between English and the Scandinavian languages is often not possible on a word-to-word basis (cf. Appendix). On the other hand, all Germanic languages largely share the same grammatical taxonomy allowing POS-based comparisons. For more details on the automatic POS tagging of speech corpora, see Nivre et al. (1996) and Nivre and Grönqvist (2001).

We are now ready to enter the main part of this report, save a concluding remark. At each of the three levels of description we wish to determine whether the corpus suite divide naturally into pairs. If this is the case, the essential question is which corpora go together – those sharing national language, or those sharing mode of communication?

5. The Danish case

This section reports on our frequency-based contrastive analysis of the two Danish corpora DanSPO and DanWRI.

5.1 Frequency distribution

As explained in Section 4, we first study the frequency distributions of DanSPO and DanWRI irrespective of the actual word types ('Zipf style').

Figure 5 shows the number of occurrences for the 30 most frequent words in DanSPO and DanWRI. From here on, we adopt the notation #*n* for the word type with rank number *n* (in a specified corpus). Notice the large difference between the leftmost columns in the graph: #1 of DanSPO has 74,159 occurrences while #1 of DanWRI has only 44,981. In general, the top ranked types are seen to cover larger parts of DanSPO than of DanWRI. #1-#10 cover 28.5% of DanSPO, while only 20.8% of DanWRI. Put in another way, it takes only 30

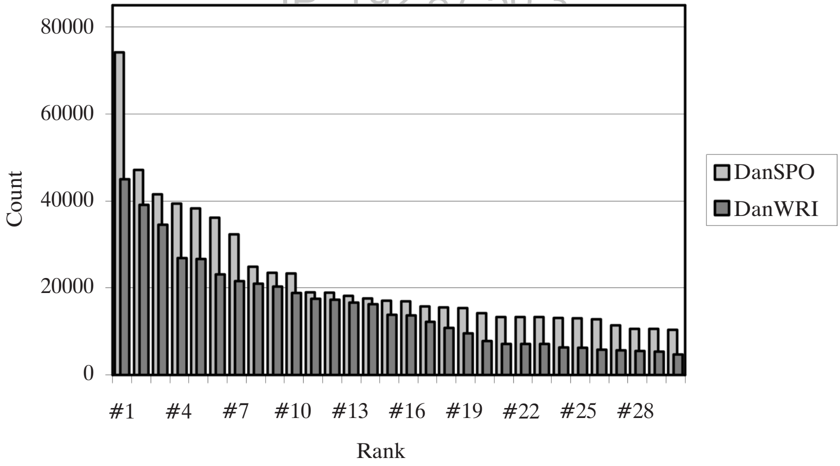

Figure 5. Danish frequency distribution. The graph shows the number of occurrences of the 30 most frequent word types in DanSPO and DanWRI respectively.

Table 2. Type-token ratio for DanSPO and DanWRI in various frequency ranges, shown in absolute and relative measures (Accumulated count / Coverage in %)

Range	DanSPO		DanWRI	
#1–10	380,599	**28.5%**	277,161	**20.8%**
#1–20	549,283	**41.1%**	412,762	**30.9%**
#1–30	671,223	**50.3%**	473,882	**35.5%**
#1–50	806,525	**60.4%**	536,450	**40.1%**
#1–100	940,834	**70.5%**	618,383	**46.3%**
#1–200	1,046,036	**78.4%**	696,591	**52.2%**
#1–500	1,144,585	**85.7%**	797,258	**59.7%**
#1–1000	1,197,630	**89.7%**	876,435	**65.6%**
#1–10000	1,302,670	**97.6%**	1,144,510	**85.7%**

types to cover 50% of the DanSPO text mass while 154 types in DanWRI. In total, DanWRI has 104,968 different types while DanSPO has only 35,112, or one third. The same words are thus reused to a much larger extent in speech than in writing. Table 2 shows the type-token ratio for selected frequency ranges.

5.2 Word type ranking

Even if the distributional patterns of DanSPO and DanWRI do differ substantially, of course their preferred types could still be the same. Is this the case? In Table 3 below we present the 10 most frequent word types in DanSPO and DanWRI, respectively.

Only four types appear in both top-10 lists, namely "det", "og", "er", "i". Of these four, only "og" (*and*) cover similar parts of DanSPO and DanWRI while "det" (*it*/*this*/*that*/*the*) is almost three times as frequent in DanSPO as in DanWRI. Also the frequencies of "er" (*BE*$_{PRES}$) and 'i' (*in*) differ markedly.

Let us pick out the ten most frequent types in DanSPO for a closer study. Each of them does occur in DanWRI, but most are not as frequent. *How* frequent is illustrated in Figure 6b below showing the relative coverage of DanSPO's #1–10 in DanWRI. On the logarithmic axis, value "1" means equal coverage in DanSPO and DanWRI, "0.1" means 10% coverage in DanWRI relative to DanSPO, etc. Figure 6b, analogously, shows the coverage of DanWRI's #1–#10 in DanSPO.

Examples: From Figure 6a we learn than "der" and "ikke" are only half as frequent in DanWRI as in DanSPO (the exact values are 0.53 and 0.49, respectively). In Figure 6b we see that "til" and "for" are similarly weak in DanSPO

Table 3. Frequency lists: 10 most frequent word types in DanSPO and DanWRI

Rank	DanSPO			DanWRI		
	Type		Count	Type		Count
#1	**det**	*it, this, …*	74,159	**i**	*in*	44,981
#2	ja	*yes*	47,127	**og**	*and*	39,145
#3	**og**	*and*	41,538	at	*to*$_{INF}$, …	34,560
#4	jeg	*I*	39,371	**er**	*BE*$_{PRES}$	26,913
#5	**er**	*BE*$_{PRES}$	38,317	**det**	*it, this, …*	26,644
#6	så	*then, so, …*	36,193	en	*a, one*	23,145
#7	der	*there, it, …*	32,305	til	*to*$_{PREP}$, …	21,615
#8	ikke	*not*	24,869	på	*on, at*	20,972
#9	var	*BE*$_{PAST}$	23,467	af	*of, off, …*	20,364
#10	**i**	*in, …*	23,341	for	*for, …*	18,822

Types occurring in both subtables (DanSPO and DanWRI) are in **bold** typeface. Only minimal English translations are given in the table, cf. Appendix for fuller translations.

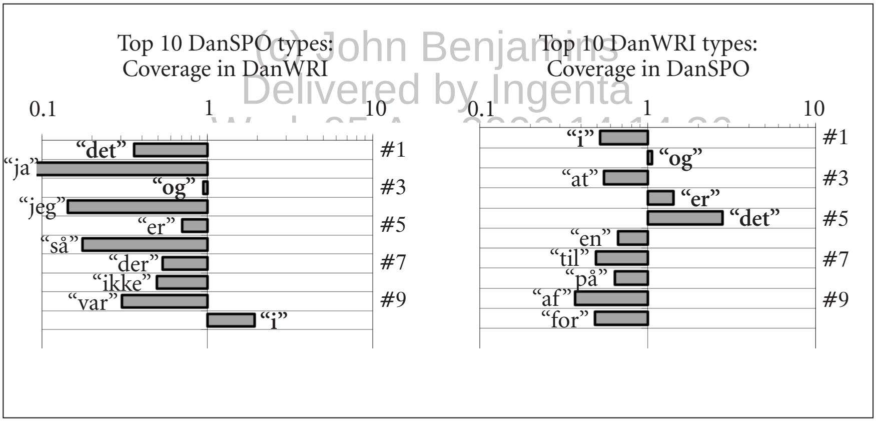

Figure 6. Relative coverage of types #1–10 in DanSPO (Figure 6a, left) and DanWRI (6b, right). Types appearing in both tables are in **bold** typeface. Eng. glosses: see Table 3.

(0.49 and 0.48). In both tables the only short bar is that of "og", being the only top ranked type of equal coverage. In Figure 6a, "ja" is extreme – actually extending over the left edge of the graph by more than one order of magnitude ("ja", *yes*, being 128 times less frequent in DanWRI than in DanSPO).

In general we find a substantial difference in coverage for almost all top ranked types, in about half of the cases by a factor 2 or more (corresponding to types that are more than twice as frequent in one corpus than in the other). This pattern is repeated in the range #11–#20 where we also find but a single type of similar coverage (#14 "de", *they*, covering just 8% more of DanSPO than DanWRI) while all others differ widely (half of them by a factor 2 or more).

The two corpora thus clearly diverge concerning word type ranking (the divergence is even more pronounced for low frequency word types as we will see shortly), but we still need to show that the discrepancy is related to the mode of communication rather than e.g. genre or topic. We therefore introduce three new written corpora in various genres: daily newspapers (referred to as W-1), magazines (W-2), and journals (W-3) (from Maegaard 1975). Compare now Figure 7 with Figure 8 below. Figure 7 repeats Figure 6b expanding the data series up to #20 while leaving out some details. Figure 8 presents the same 20 word types showing their coverage in W-1, W-2, and W-3.

Unsurprisingly, the choice of written genre (Figure 8) does have some impact on the word type selection, but the disagreements among the written corpora are generally much smaller than in Figure 7. Also the general picture is far less chaotic. Among the new corpora, W-2 ("magazines") diverges most from DanWRI, yet no type in this corpus is over-selected by a factor greater than 1.5 (or under-selected by less than 0.66). This is clearly in contrast to the DanSPO versus DanWRI case where we found half of the types being over-selected by a factor 2+ (or under-selected by 0.5).

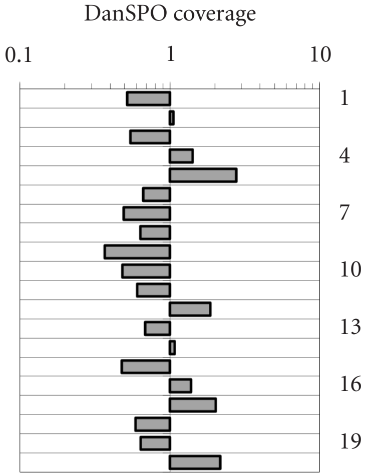

Figure 7. Coverage of DanWRI's word types #1–20 in DanSPO

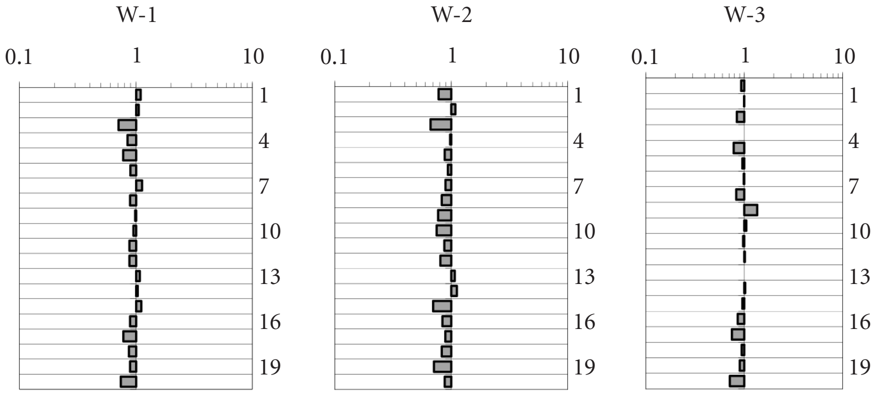

Figure 8. Coverage of DanWRI’s word types #1–20 in newspapers (W-1), magazines (W-2) and journals (W-3)

The next question is whether this picture is repeated for the infrequent types. To answer this question we adopt a formula for quantifying the deviation in word type preference in two corpora *X* and *Y*.

Deviation in Word Type Preference (DWTP)[4]

$$\mathrm{DWTP}(X, Y, \#a, \#b) = \frac{\displaystyle\sum_{r=a}^{b} \left| \log \frac{Freq(X, Type_r^X)}{Freq(Y, Type_r^X)} \right|}{b - a + 1}$$

Function DWTP$(X, Y, \#a, \#b)$ measures the mean deviation of types *#a–#b* with regard to coverage in X and Y. DWTP is thus a function of the types in range *#a–#b* of *X*. $Type_{r'}^{X'}$ is the type ranked *r′* in (the frequency list of) corpus *X′*. *Freq*(*X*’, *T*’) is the frequency of type *T*’ in *X*’. Value 0 corresponds to total agreement (each type in *#a–#b* covers equal parts in *X* and *Y*). Other values represent disagreement – more so, the larger the value. Examples:

DWTP(DanSPO, DanWRI, 1, 10) = 1.315
DWTP(DanWRI, DanSPO, 1, 10) = 0.599

These values correspond to Figures 6a and 6b above, as DWTP values are equivalent to the mean length of the bars. Notice that, due to the asymmetry of the formula (range $\#a - \#b$ referring to X rather than Y) DWTP(X', Y', a', b') is not in general equal to DWTP(Y', X', a', b'). The tables in Figure 8 compute as:

Table 4. Deviation in word type preference for assorted type ranges (shown as DWTP values)

Range	(DanSPO, DanWRI)	(DanWRI, DanSPO)
#1–10	1.135	0.599
#11–20	1.453	0.509
#21–30	1.211	0.920
#31–50	1.014	0.553
#51–100	0.825	0.837
#101–200	0.932	1.039
#201–500	1.108	1.437
#501–1000	1.036	1.711
#1001–10000	1.109	1.642

DWTP(DanWRI, DanSPO, 1, 20) = 0.554

DWTP(DanWRI, W-1, 1, 20) = 0.147
DWTP(DanWRI, W-2, 1, 20) = 0.183
DWTP(DanWRI, W-3, 1, 20) = 0.108

For DWTP values of <0.2, the deviation is less than 22% (insignificant). Value 0.69 (= log 2) means a deviation of a factor 2; value 1.1 is factor 3, and value 1.61 is factor 5.

In Table 4, DanSPO and DanWRI are compared by lexical selection for various type ranges. As seen, disagreement is generally stronger in the low end of the frequency scale. In particular, the DanWRI types ranked #201+ (mainly content words) have markedly different frequencies than have the same types in DanSPO (differing by more than a factor 4 on average).

We conclude that spoken and written Danish (as represented in our reference corpora) prefer different types to a large extent. While this holds true for all type ranges, the disagreement becomes extreme in the lower part of the DanWRI list (#201+).

5.3 Grammatical observations

We now add a grammatical dimension to our investigations. Using Eric Brill's algorithm (Brill 1994), we annotate DanSPO and DanWRI with part of speech tags, employing the Danish PAROLE tagset (Bilgram & Keson 1998; Henrichsen 2002a). This tagset contains about 150 tags distributed over 10 major parts

Table 5. Personal pronouns

Type		DanSPO		DanWRI		Weighted-count
		Rank	Count	Rank	Count	W-1 / W-2 / W-3
jeg	*I*	#4	39,371	#27	5,638	4,800 / 11668 / 2,787
du	*you*$_{SG}$	#18	15,553	#197	557	1,166 / 3,610 / 342
vi	*we*	#22	13,326	#28	5,498	5,415 / 5,810 / 4,790
han	*he*	#23	13,314	#23	7,122	5,722 / 7,348 / 2,115
hun	*she*	#40	6,609	#59	2,073	1,653 / 5,036 / 422

W-1,2,3 figures are weighted for direct comparison with DanSPO/DanWRI figures

Table 6. Particles with special discourse functions

Type		DanSPO		DanWRI		Weighted-count
		Rank	Count	Rank	Count	W-1 / W-2 / W-3
så	*so, then, ...*	#6	36,193	#24	6,351	4,909 / 8,448 / 4,614
der	*there, ...*	#7	32,305	#12	17,277	15,213 / 13,953 / 17,606
ikke / ik’	*not / y’know*	#8 / #15	41,968	#17	12,257	9,580 / 10,856 / 9,217
altså	*so, well, ...*	#20	14,239	#234	453	340 / 486 / 561
sådan	*thus, like, ...*	#26	12,837	#152	754	590 / 988 / 758

W-1,2,3 figures are weighted for direct comparison with DanSPO/DanWRI figures

of speech: noun, verb, adjective, pronoun, conjunction, preposition, adverb, interjection, ‘unique’ (for grammatical particles etc.), and ‘residual’ (a rarely used category for www-addresses, smileys, unidentified tokens etc).

The POS-tagged versions of DanSPO and DanWRI allow us to pose new questions: Which parts of speech dominate in written Danish, and which in speech? Is there any agreement?

Consider first some examples, the cases of personal pronouns, prepositions, determiners, and ‘discourse particles’ (the latter referring to a loose collection of conjunctions and adverbials etc. acting as modifiers, “altså”, “så”, “sådan”, “der”, or as feedback triggers, “ik’ ”. They are often found in utterance final position, and often have a prolonged vowel/sonorant, e.g. “ altså∼ “, ” sån∼”).

As seen, pronouns and discourse particles are in general far more frequent in DanSPO than in DanWRI. There is, however, considerable variation within the categories: while “vi”, “han”, and “hun” are 2–3 times more frequent in DanSPO, the corresponding figures for “jeg” and “du” are 7 and 28, respectively. So, 1st person is extremely common in informal conversation, while quite rare in the daily paper.

Table 7. Determiners

	Type	DanSPO		DanWRI		Weighted-count
		Rank	Count	Rank	Count	W-1 / W-2 / W-3
en	a_{UTT}, ...	#19	15,406	#6	23,145	20,736 / 21,691 / 22,214
den	the_{SG+UTR}, ...	#29	10,590	#11	17,509	15,392 / 15,326 / 17,067
et	a_{NEU}, ...	#41	6,456	#18	10,799	9,394 / 8,961 / 10,173

W-1,2,3 count figures are weighted for direct comparison with DanSPO/DanWRI figures

Table 8. Prepositions

	Type	DanSPO		DanWRI		Weighted-count
		Rank	Count	Rank	Count	W-1 / W-2 / W-3
i	*in*, ...	#10	23,341	#1	44,981	49,508 / 35,116 / 41,716
til	to_{PREP}, ...	#28	10,628	#7	21,615	24,453 / 19,374 / 21,397
på	*on*, ...	#21	13,343	#8	20,972	18,699 / 17,387 / 17,430
af	*of*, *off*, ...	#35	7,511	#9	20,364	20,025 / 15,689 / 27,779
for	*for*, ...	#34	9,099	#10	18,822	17,974 / 14,130 / 19,731
med	*with*, *by*, ...	#27	11,384	#13	16,615	17,910 / 17,814 / 16,688

W-1,2,3 figures are weighted for direct comparison with DanSPO/DanWRI figures

Determiners and prepositions, in contrast, are favored in DanWRI (see Tables 7 and 8).

"I" (*in*) is the favorite preposition of both DanSPO and DanWRI (by a safe margin). Otherwise there is little or no agreement among the most frequent prepositions:

- DanWRI: til > på > af > for > med
- DanSPO: på > med > til > for > af

DanSPO and DanWRI tokens are distributed over parts of speech as follows:[5]

DanWRI: **Noun**-Verb-**Prep**-Pro-**Adj**-Adv-Conj-Unique-Residual-Interjec
DanSPO: **Pro**-Verb-**Adv**-Noun-Conj-Prep-**Interjec**-Adj-Unique-Residual

In each row the order of the elements reflects the *absolute* numbers of tokens in the corresponding categories. DanWRI, hence, has more nouns than verbs, more verbs than prepositions, etc. Interjections is the smallest category in DanWRI.

Categories in **bold** are larger in relative measures, i.e. much larger in one corpus than in the other. By way of example, interjections are not the largest

Table 9. DanSPO's favored categories

PAROLE tag	POS	Examples	Count	
			DanSPO	DanWRI
RGU	adverb	så, sådan, altså	200,151	87,813
I=	interjection	ja, nej, nå, mm	88,115	498
PP;	personal pronoun	jeg, han, jeres	192,415	57,899
PT;	interrogative pron.	hvem, hvad	7,404	2,458

Table 10. DanWRI's favored categories

PAROLE tag	POS	Examples	Count	
			DanSPO	DanWRI
SP	preposition	under, i	89,870	178,696
NC;	common noun	dag, pigernes	123,404	296,005
AN;	adjective	stort, bedre	77,278	128,737
NP;	proper name	Bo, Norge	18,875	92,995

category of DanSPO, yet much larger in DanSPO than in DanWRI (7th in the DanSPO sequence, 10th in DanWRI).

In conclusion, written Danish prefers nouns, prepositions, and adjectives, while spoken Danish prefers pronouns, adverbs, and interjections (including all sorts of particles used as attention signals, feedback, response elicitors etc).

5.4 Concluding remarks on the Danish data

We have used statistical measures to compare the verbal material of corpora DanSPO and DanWRI of spoken and written Danish. We have found substantial differences between the two, not only in frequency distribution, but in word type selection and in categorical preference as well. Each of these three dimensions is independent of the two others, in the sense that large disagreement in any one dimension may well co-occur with near-agreement in the other two. It is therefore interesting to observe that DanSPO and DanWRI diverge substantially in each dimension.

In Section 7 – after having presented the Swedish data – we follow up on these preliminary observations.

6. The Swedish case

We now turn to the next question: How does Swedish relate to Danish? Where are the most significant similarities – and the most pronounced differences? Are the conclusions drawn in the previous section specific for Danish, or do they hold for Swedish too?

For ease of comparison we present the Swedish and Danish figures side by side in most of the tables below, repeating some Danish data from the previous sections.

6.1 Frequency distribution

We first compare the frequency distribution of spoken and written Swedish in Figure 9, and compare it to the corresponding Danish graph, repeated below as Figure 10.

As the graphs 9 and 10 show, the frequency distributions for spoken and written Swedish closely match those of spoken and written Danish. This actually is valid for all frequency ranges, as seen in Table 11.

The remarkably close match in columns SweSPO and DanSPO is not quite matched by the written corpora. At rank 100, the accumulated frequency for DanWRI is about 3% higher than for SweWRI. This difference may be due to

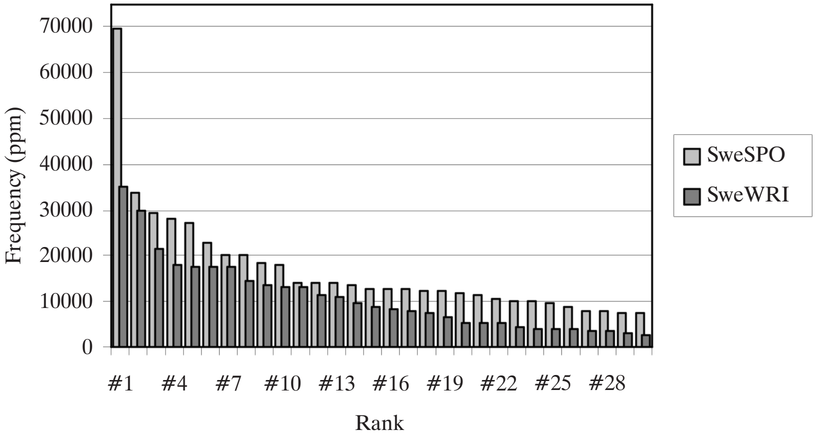

Figure 9. Swedish frequency distribution, 30 top-ranked types (frequencies shown in parts-per-million).

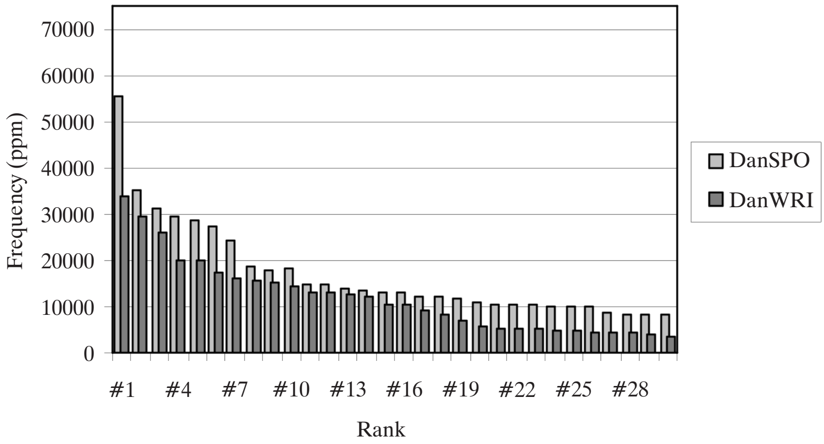

Figure 10. Danish frequency distributions, 30 top-ranked types. Same data as in Figure 5, rendered here in ppm.

Table 11. Accumulated frequencies

Rank	SweSPO	DanSPO	SweWRI	DanWRI
... #10	28.8%	28.5%	19.8%	20.8%
... #20	41.7%	41.1%	28.8%	30.9%
... #30	50.9%	50.3%	32.8%	35.5%
... #50	60.4%	60.4%	37.1%	40.1%
... #100	71.0%	70.5%	43.5%	46.3%
... #200	78.4%	78.4%	49.6%	52.2%
... #500	86.2%	85.7%	58.1%	59.7%
... #1000	90.7%	89.7%	64.6%	65.6%

factors beyond our control. There may for instance be minor differences in the word token definitions applied in corpus Berlingske-99 (DanWRI) and corpus Göteborgsposten (SweWRI), or the two newspaper corpora may not consist of exactly the same text genres. In any case, the difference is fairly small and certainly insignificant in comparison with the spoken-written discrepancy.

6.2 Word type ranking

We now take a comparative look at the word type ranking in all four corpora. Table 12 provides a comparison of the 10 most highly ranked words in the two languages.

Table 12. Aligned Swedish and Danish frequency lists

Rank	SweSPO (type / count)		DanSPO (type / count)		SweWRI (type / count)		DanWRI (type / count)	
#1	det [∼det]	69,622	det [+det]	55,552	i [∼i]	34,887	i [+i]	33,687
#2	är [∼er]	33,641	ja [+ja]	35,303	och [∼og]	30,030	og [+och]	29,317
#3	och [∼og]	29,534	og [+och]	31,116	att [∼at]	21,633	at [+att]	25,883
#4	ja [∼ja]	28,093	jeg [+jag]	29,493	en [∼en]	17,783	er [+är]	20,158
#5	att [∼at]	27,270	er [+är]	28,703	det [∼det]	17,710	det [+det]	19,954
#6	jag [∼jeg]	22,811	så [+så]	27,112	på [∼på]	17,542	en [+en]	17,334
#7	man [∼man]	20,190	der [–]	24,200	som [∼som]	17,519	til [+till]	16,188
#8	så [∼så]	19,964	ikke [+inte]	18,629	är [∼er]	14,494	på [+på]	15,706
#9	som [∼som]	18,318	var [+var]	17,579	av [∼af]	13,398	af [+av]	15,251
#10	inte [∼ikke]	18,058	i [+i]	17,479	med [∼med]	13,182	for [+för]	14,096

Count figures are in parts-per-million. Lexemes with +/∼ are nearest Swe./Dan. equivalents according to Palmgren et al. (2001) and Molde (2000). Cf. Appendix for Eng. translations.

Table 13. Word type ranking (DWTP values): Swedish speech vs. writing

Range	(SweSPO, SweWRI)	(SweWRI, SweSPO)
#1–#10	1.313	0.571
#11–#20	1.251	0.799
#21–#30	1.289	1.064
#31–#50	1.150	0.988
#51–#100	1.187	0.819
#101–#200	1.038	1.152
#201–#500	1.005	1.468
#501–#1000	0.994	1.462

Types ranked #1001+ are not considered in this table due to the small size of corpus gbg.

The top-10 lists of spoken Danish and spoken Swedish share seven types (i.e. nearest-equivalent translations according to two leading dictionaries). The three remaining Swedish types are *att, man* and *som*, the Danish residual being *der* (which lacks a Swedish equivalent), *var* and *i*. For the written language, the overlap is also considerable with eight out of 10 equivalents, the Swedish residual being *som* and *med*, the Danish residual *til* and *for.* In contrast, the SweSPO and SweWRI top-10 lists share five types only, and the DanSPO and DanWRI lists share only four. We may also note that the rank order is more similar between the two spoken language variants than it is between the spoken and written variant of the same language. The same holds for the two written variants.

Table 14. Word type ranking (DWTP values): Danish versus Swedish

Range	(DanSPO, SweSPO)	(DanWRI, SweWRI)
#1–#10	0.300	0.143
#11–#20	0.316	0.169
#21–#30	0.401	0.158
#31–#50	0.616	0.291
#51–#100	0.646	0.332
#101–#200	0.694	0.373

Table 13 shows that written Swedish and spoken Swedish are quite distinct in their word type preferences – in parallel with the Danish case as seen in Table 4 in Section 5.2.

Table 14 is based on a word-to-word Dan-Swe translation list compiled from a standard dictionary (Svensk-Dansk Ordbog 2001) and carefully examined by three linguists (not including the authors). In preparing the data for Table 14, all content words were excluded from the calculation being presumably typical of the activity type rather than a more general structural feature of the language. The common noun “naturen” (*nature*) is e.g. very frequent in GSLC while absent in BySoc, whereas the proper noun ‘Nyboder’ (a suburb of Copenhagen) is frequent in BySoc, but absent in GSLC.

The table shows that the distinct preferences noted in Table 4 and Table 13 are indeed upheld, so that (i) spoken Danish and Swedish are similar (especially concerning the top-ranked types), and (ii) written Danish and Swedish are also similar.

6.3 Grammatical observations

Spoken and written tokens are distributed over the major parts of speech as follows:

DanWRI:	Noun-Verb-Prep-Pro-Adj-Adv-Conj
DanSPO:	Pro-Verb-Adv-Noun-Conj-Prep-Adj
SweWRI:	Noun-Verb-Prep-Pro-Adv-Adj -Conj
SweSPO:	Pro-Verb-Adv-Noun-Conj-Prep-Adj

The tagsets employed in the tagged versions of the four reference corpora are not identical, but at least compatible. Certain specific categories had to be omitted from the investigation being absent in at least one of the corpora. The ignored categories are: Feedback, Own Communication Management, Unique, Residual, Interjections, Numerals. The SweSPO category list is compiled from Allwood (1999). The exclusion of the categories feedback, own communication management and interjection has the consequence that the apparent differences between spoken and written language are actually diminished, since these three categories are all very much more common in spoken language than in written language. If we disregard the caveat of Section 4.2 above, it is striking that the distributions of the parts of speech are near-identical in spoken Danish and Swedish, and again in written Danish and Swedish. In contrast, the distributions differ widely when holding constant the language rather than mode of communication, i.e. the similarity is much greater between spoken Danish and spoken Swedish than it is between spoken and written Danish or spoken and written Swedish.

7. Language versus mode of communication

Having aligned the Swedish and Danish figures, we are now in a position to draw some overall conclusions concerning the relative importance of mode of communication in comparison with national language. Table 15 and 16 below illustrate the relations between the four reference corpora. In each cell is represented the DWTP value for a certain combination of corpora.

Several conclusions can be read off these tables.

Table 15. Word type ranking (DWTP values, 10 top-ranked types)

Rank #1-#10	**SPOKEN versus ...**	**WRITTEN versus ...**
... SPOKEN	(DanSPO,SweSPO) = 0.30 (SweSPO,DanSPO) = 0.29	(DanWRI,DanSPO) = 0.60 (SweWRI,SweSPO) = 0.57
... WRITTEN	(DanSPO,DanWRI) = 1.14 (SweSPO,SweWRI) = 1.31	(DanWRI,SweWRI) = 0.14 (SweWRI,DanWRI) = 0.11

Table 16. Word type ranking (DWTP values, 100 top-ranked types)

Rank #1-#100	**SPOKEN versus ...**	**WRITTEN versus ...**
... SPOKEN	(DanSPO,SweSPO) = 0.45 (SweSPO,DanSPO) = 0.46	(DanWRI,DanSPO) = 0.74 (SweWRI,SweSPO) = 0.85
... WRITTEN	(DanSPO,DanWRI) = 0.98 (SweSPO,SweWRI) = 1.14	(DanWRI,SweWRI) = 0.27 (SweWRI,DanWRI) = 0.25

First, the choice of *mode of communication* is clearly more significant than the choice of *national language* with respect to the distributional patterns discussed in this paper. Written Swedish is far more similar to written Danish (DWTP = 0.11 in Table 15) than it is to spoken Swedish (0.57). Spoken Danish and Swedish are much more similar (0.30) than spoken and written Danish (1.14), and so forth.

Secondly, concerning lexical preferences, spoken language seems more idiosyncratic than written language meaning that the top ranked types of written language are all moderately frequent in speech as well while a number of the top ranked words of speech are extremely rare in writing. Compare values

DWTP(SweSPO,SweWRI,#1,#10)=1.31
DWTP(SweWRI,SweSPO,#1,#10)=0.57,

a highly significant difference recalling that DWTP-values are logarithmic.

Thirdly, the two main conclusions above persist when extending the range under consideration from #1-#10 to #1-#100, but the distinctions become less pronounced: large DWTP-values (corresponding to large discrepancies) tend to decrease, while small values (close similarities) increase. In other words, it is to a large degree the top-frequent types that account for the differences between the spoken and written mode of communication. When including more types, the main conclusions still hold, but less clearly so.

Finally, observe that the figures within each cell are equal within a small margin, indicating that the transfer between languages (keeping the mode of communication constant) is largely symmetrical: the difference between written and spoken Danish (DWTP = 0.60) closely matches the difference between written and spoken Swedish (DWTP = 0.57), and so forth. This symmetry is perhaps not surprising, yet encouraging since the opposite situation would blur the otherwise quite clear conclusions that we have been able to draw.

Figure 11 below illustrates the data of Table 15. Each data point represents a pair of corpora (C1, C2), the X-value corresponding to DWTP(#1,#10,C1,C2), and the Y-value to DWTP(#1,#10,C2,C1). The geometrical distance to (0,0) hence measures the disagreement in word type selection (larger distance meaning larger disagreement). In other words, points near to (0,0) represent corpora with similar word type preferences, while data points far from (0,0) represent corpora that disagree on which word types to prefer.

The perhaps surprising conclusion is that Danish speech and Swedish speech are much more similar to each other in the dimensions investigated here – frequency distribution, word type selection, and distribution of parts

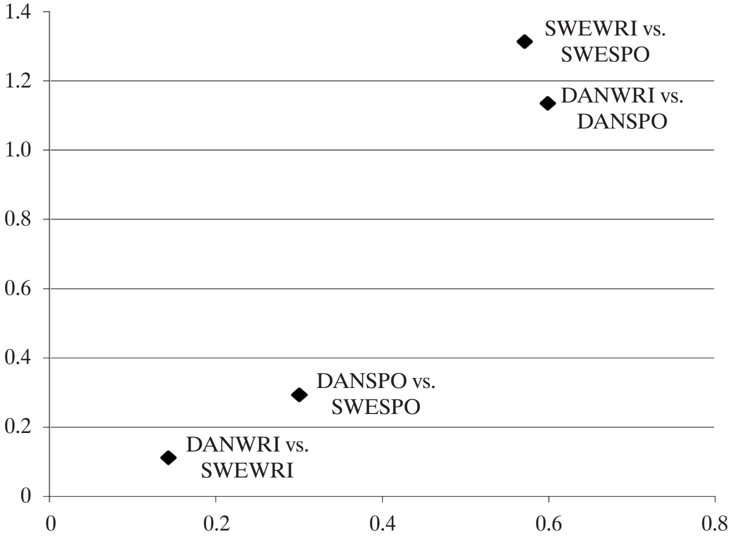

Figure 11. Pairwise similarities mapped as DWTP-values

of speech – than Danish speech is to Danish writing and Swedish speech is to Swedish writing. There seem to be grounds for claiming that *if* spoken Danish and written Danish are to be upheld as one language, certainly spoken Danish and spoken Swedish should be considered as one language as well – and similarly for Swedish. Whether this conclusion holds in other dimensions of language like syntax, semantics and pragmatics remains to be seen; for now we leave this interesting question to further investigation. Our conjecture is that the perceived difference between Danish and Swedish speech is mostly a matter of minor lexical and phonological transformations.

Open word classes (proper nouns, common nouns, adjectives, content verbs) are strong in written language while function words including pronouns dominate in speech.[6] Typical written language includes large numbers of different words, while spoken interaction to a much greater extent recycle the same words. Each type in DanWRI thus has less than 13 occurrences on average, while each DanSPO type has 38.

However important these general observations may be, a single type seems to provide the biggest surprise: the strikingly high frequency of multi-purpose pronoun *det*. Ask a Dane or Swede which word is the most frequent in his everyday vernacular, and chances are that (s)he will suggest “og”/“och”, “i”, "at"/"att", "jeg"/"jag" or "ja" – hardly ever "det" (we tried this many times). Even in the process of answering the question, he/she will inevitably use "det" a dozen times – without noticing. Why is "det" so frequent, and yet so invisible?

7.1 Biber et al. 1999

The Longman Grammar of Spoken and Written English (Biber et al. 1999) provides grounds for interesting comparisons with our findings. The book reports on a large-scale corpus-based comparison of various styles of English speech and writing (termed 'registers'), including newspaper text (5.4 million words) and transcriptions of conversations (3.9 million words). Fitting together pieces of information scattered in various sections (2.3.5, 2.4.14., 4.1.2, 4.10.5, 14.3.3 et pass.), we arrive at this POS distribution:

English speech (conversations):	**Pro**-Noun-Verb-**Adv**-Prep-Adj
English writing (news):	**Noun-Prep**-Verb-**Adj**-Pro-Adv

Categories larger in one mode ("register") than in the other are in **bold** typeface (our category of conjunctions (Conj) is incompatible with the Longman POS inventory). Comparing with our findings (see 6.3 above) we observe that – in English as in Scandinavian – nouns, prepositions, and adjectives are typical of writing, while pronouns and adverbs (and obviously interjections) are typical of speech.

Why are pronouns, adverbs, and interjections typical of speech? What do these three categories have in common? Certainly not their morphology. In so far as these items have morphological features at all, pronouns are more like nouns, adverbs like adjectives, and interjections like grammatical particles – all categories typical of the written language.

Perhaps some of the reasons are the following: Spoken language draws on contextually available information more than written language, where less context is available and the text has to be more explicit. One consequence of this is that written language has to introduce and maintain reference by explicit use of descriptive nouns and noun phrases, while spoken language, relying on context, can make do with pronouns. Interactive spoken language is also generally more impulsive and reactive. This means that there is a greater need for and use of interjections (including words for feedback and own communication management) and attitudinal adverbs. In written language, as already has been mentioned, on the other hand, there is a greater need of contextual explicitness, which is often met by using longer descriptive noun phrases containing prepositions, binding the phrases together, and adjectives to provide more explicit descriptive information. In spoken language instead there are often more conjunctions helping to flesh out content, that may be more compressed in written language, into several short statements.

8. Conclusions and implications for further research

In this paper, we have studied three dimensions of language based on word frequencies: frequency distribution, word type ranking and the distribution of parts of speech in spoken and written Danish and Swedish. We have found that in all of the three dimensions spoken Danish and spoken Swedish are more similar to each other than are spoken Danish to written Danish or spoken Swedish to written Swedish.

At the moment, however, it is a little unclear what conclusions can be drawn from these observations. A first conclusion might be that the differences between spoken and written language are the same in at least three important respects in Danish and Swedish. Given the compatibility with English data discussed above, it is not unlikely that the same differences might be found between the spoken and written modes of other languages, i.e.,

(i) Common words are reused more often in spoken language than in written language and written language has a richer vocabulary in frequent use than spoken language.

(ii) The discourse functions expressed by certain very frequent words could represent a constant functional need in the spoken languages of a certain language type. The picture for written language is less clear.

(iii) The discourse functions expressed by pronouns, adverbs, interjections and conjunctions are more typical of spoken language than written language, while the discourse functions expressed by nouns, adjectives and prepositions are more typical of written language.

A second conclusion might be that spoken Danish and spoken Swedish are more closely related than spoken and written variants of Danish and Swedish are to each other. The plausibility of this conclusion depends on how the properties we have studied are related to other properties that give a language its identity. It also depends on whether the properties we have observed are general and universal features of the difference between spoken and written language variants rather than a feature of the particular relationships between spoken and written Danish and Swedish.

We therefore believe our study should be extended in two ways: (i) We should investigate the difference between spoken and written variants of other languages than Swedish and Danish, in order to see if our results reoccur. (ii) We should attempt to correlate our present findings concerning spoken and written Swedish and Danish with other features of these languages, to see if what we have found is part of a more general picture of the relationship between the languages.[7]

9. Acknowledgments

We wish to thank our two anonymous reviewers for their precise and highly value-adding comments.

Notes

[1] – at least since the decline of NST (Nordic Speech Technology) in Voss, Norway.

[2] 'Displacement' is an intuitive term coined to denote the semantic distance between Danish-Swedish synonyms (i.e. translations preferred by language users with a solid understanding of both languages). The term 'displacement' is thus used for informal presentation only, not for data analysis.

[3] This is also the stance taken in the extensive corpus-based Longman Grammar of Spoken and Written English (Biber et al. 1999) where spoken language transcripts (conversations) are compared directly with written sources like newspaper texts and academic prose.

[4] To avoid illegal 0s, tokens appearing in *X* but not in *Y* are ignored in computing DWTP. For simplicity we don't use *smoothing* to level out the granularity effects of such zeroes; consequently, DWTP values may be too small for low frequency ranges (greatest discrepancies being ignored).

[5] Due to the methodological uncertainties concerning speech tagging (cf. 4.2 above) we will not present the actual sizes of the categories.

[6] Certain function categories are actually more common in writing: determiners, prepositions.

[7] Some recent, affirmative results are reported in Henrichsen (2004).

References

Allwood, J. (1998). *Some Frequency based Differences between Spoken and Written Swedish.* Proceedings of XVIth Scandinavian Conference of Linguistics (pp. 18–29). University of Turku.

Allwood, J. (1999). Talspråksfrekvenser. *Gothenburg Papers in Theoretical Liguistics* S21, Gothenburg University.

Allwood, J., Grönqvist, L., Ahlsén, E., & Gunnarsson, M. (2002a). Göteborgskorpusen för Talspråk. In P. J. Henrichsen (Ed.), *Korpuslingvistik* (pp. 39–58). Copenhagen: Akademisk Forlag.

Allwood, J., Henrichsen, P. J., Ahlsén, E., Grönqvist, L., & Gunnarsson, M. (2002b). Transliteration Between Spoken Language Corpora – Moving between Danish BySoc and Swedish GSLC. *Gothenburg Papers in Theoretical Linguistics* 86, Gothenburg University.

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

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English.* London: Longman.

Bilgram, T. & Keson, B. (1998). The Construction of a Tagged Danish Corpus. *Proceedings of NODALIDA-1998* (pp. 129–139). University of Copenhagen.

Brill, E. (1994). A Report of Recent Progress in Transformation-based Error-driven Learning. *Proceedings of the ARPA Workshop on Human Language technology 1994*, Princeton, N. J. [Brill's tagger is available at http://www.cs.jhu.edu/~brill]

Dansk-Svenska Ordbok. (2000). 3. edition. Molde, B. (Ed.), Stockholm: Norstedts Förlag.

Debaisieux, J.-M., & Deulofeu, J. (2001). Grammatically Unacceptable Utterances are Communicatively Accepted by Native Speakers; Why are They? *Proceedings of DISS-01* (pp. 69–72), University of Edinburgh.

Gregersen, F., & Pedersen, I. L. (Eds.). (1991). *The Copenhagen Study in Urban Sociolinguistics.* Copenhagen: Reitzel, Vols. 1+2.

Henrichsen, P. J. (1998). Peeking Into The Danish Living Room – Internet access to a large Danish speech corpus. *Proceedings of NODALIDA-1998* (pp. 109–119). University of Copenhagen.

Henrichsen, P. J. (2002a). Fyrre Kilometer Kryds og Bolle – metoder til grammatisk opmærkning i største skala. In P. J. Henrichsen (Ed), *Korpuslingvistik* (pp. 68–88). Copenhagen: Akademisk Forlag.

Henrichsen, P. J. (2002b). Some Statistically based Differences Between Spoken and Written Danish. *Gothenburg Papers in Theoretical Linguistics* 88, Gothenburg University.

Henrichsen, P. J. (2004). Siblings and Cousins – Statistical Methods for Spoken Language Analysis. *Acta Linguistica Hafniensia, 36*, 7–33. Copenhagen: Reitzel.

Labov, W. (1984). Field methods of the Project on Linguistic Change and Variation. In J. Baugh et al. (Eds.), *Language in Use: Readings in Sociolinguistics* (pp. 28–53). Englewood Cliffs, NJ: Prentice Hall.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word Frequencies in Written and Spoken English.* London: Longman.

Maegaard, B. (1975). *Hyppige ord i Danske Aviser, Ugeblade og Fagblade.* Copenhagen: Gyldendal.

McKelvie, D. (1998). *The Syntax of Disfluency in Spontaneous Spoken Language.* Human Communications Research Centre, University of Edinburgh, Research Paper HCRC/RP-95.

Nivre, J., Grönqvist, L., Gustafson, M., Lager, T., & Sofkova, S. (1996). Tagging Spoken Language Using Written Language Statistics. *16th ICCL* (at COLING-96) (pp. 1078–1081). University of Copenhagen.

Nivre, J., & Grönqvist, L. (2001). Tagging a Corpus of Spoken Swedish. *International Journal of Corpus Linguistics, 6* (1), 47–78.

Svensk-Dansk Ordbog. (2001). Palmgren, V., Munch-Petersen, V. P. & Hartmann, E. (Eds.). 3 edition. Copenhagen: Gyldendal.

Zipf, G. K. (1936). *The Psycho-biology of Language – Introduction to Dynamic Philology.* London: Routledge.

Authors' addresses

Peter Juel Henrichsen
Center for Computational Modelling of Language
Dept. of Computational Linguistics
Copenhagen Business School
Bernhard Bangs Allé 17B
DK-2000 Copenhagen F
Denmark
E-mail: pjuel@id.cbs.dk

Jens Allwood
Dept. of Linguistics,
University of Gothenburg
Box 200
SE-40530 Göteborg
Sweden
E-mail: jens@ling.gu.se

Appendix – English translations

Approximate English translations of top-frequent Swedish and Danish types

Rank	SweSPO	DanSPO	SweWRI	DanWRI
#1	**det** *it*, *this*$_{PRO}$, *that*$_{PRO}$, *the*$_{DEF+NEU}$	**det** = Swe. "det"	**i** *in*$_{PREP}$, *in*$_{ADV}$	**i** *in*$_{PREP}$, *in*$_{ADV}$
#2	**är** *BE*$_{PRES}$	**ja** *yes, yeah*	**och** *and*	**og** *and*
#3	**och** *and*	**og** *and*	**att** cf. SweSPO	at = Swe. "att"
#4	**ja** *yes, yeah*	**jeg** *I*	**en** *a*$_{UTRUM}$, *one*$_{NUM}$, *one*$_{PRO}$	**er** *BE*$_{PRES}$
#5	**att** *to*$_{INF}$, *that*$_{SUBORD}$	**er** *BE*$_{PRES}$	**det** cf. SweSPO	**det** = Swe. "det"
#6	**jag** *I*	**så** = Swe. "så"	**på** *on, at*	**en** = Swe. "en"
#7	**man** *one*$_{GENERIC+NOM+SG}$, *you*$_{GENERIC+NOM+SG}$	**der** *there*$_{PRO}$, *there*$_{ADV}$	**som** *that*$_{SUBORD}$, *who*$_{SUBORD}$, *whom*$_{SUBORD}$, *which*$_{SUBORD}$, *like*$_{CONJ}$	**til** *to, till*
#8	**så** *so*, *that*$_{SUBORD}$, *then*$_{COORD}$	**ikke** *not*	**är** cf. SweSPO	**på** *on, at*
#9	**var** *BE*$_{PAST}$	**var** *BE*$_{PAST}$	**av** *of, off, by*	**af** *of, off , by*

#10	**inte**
	not
		i
		in$_{PREP}$, *in*$_{ADV}$
			med
			with, by
				for
				for, too

Capitalized symbols refer to paradigms (e.g. BE$_{PRES}$ *= am/is/are)*