Spoken Language in Dictionaries: Does It Really Matter?

Lars Trap-Jensen
Det Danske Sprog- og Litteraturselskab (DSL)
Society for Danish Language and Literature
Christians Brygge 1
1219 COPENHAGEN K
DENMARK
ltj@dsl.dk

Abstract
The paper asks in retrospect whether the end product has justified the time-consuming task of incorporating spoken data in the corpus underlying a new six volume dictionary of contemporary Danish. In what way has the inclusion of spoken language affected the overall appearance, which parts of the vocabulary are particularly affected, and was it altogether worth the effort?

1. Introduction: Spoken language in corpora and dictionaries

We probably all know it: speech makes up the bulk of language being produced, and yes, we should pay more attention to spoken language when making our dictionaries. But it is such a nuisance with all its hesitation signals, self-corrections, false starts, repetitions and slips of the tongue. And to top it all, it is even expensive. A dictionary project starting from scratch with building its own corpus faces a time-consuming and expensive task, in particular if access to ready-made transcripts in an appropriate electronic format is unavailable. Representing a dictionary project, The Danish Dictionary, that did in fact make the cumbersome effort of incorporating a substantial proportion of spoken language in the underlying corpus, I shall in this paper try to evaluate the outcome of the effort. The dictionary has now been completed and is presently in the stage of publication, so now is the right time to ask: how widely has the spoken material been used, has spoken language left a mark on the general flavour of the dictionary, and was it really worthwhile taking all the trouble if, perhaps, the outcome is modest?

2. Spoken language in the corpus of The Danish Dictionary

In the corpus of The Danish Dictionary spoken language amounts to c 7-8 million spoken words out of a total of 40 million words, equivalent to 17-20 per cent of the corpus. The amount is deliberately given in rough figures as the boundary between spoken and written language is not always clear-cut. The higher figure also includes what we call 'speech paper', which covers texts written down in support of a speech or another oral presentation, as well as 'paper speech', which includes television subtitles, transcripts of verbatim reports from parliamentary negotiations, and the like. Both media clearly display features of spoken language, especially in vocabulary and syntactic patterns, but cannot be categorized entirely as such as the texts in other respects resemble the written medium: utterances have to a

certain extent been normalized to well-formed sentences, hesitation signals and other kinds of 'noise' have been omitted, and so on. In other words, they represent intermediate categories.

It follows from the 80-20 distribution between written and spoken language that the language being described is chiefly written texts. This is in fact in accordance with the official mandate given when the project was initiated. Here it is stated that the dictionary should "*cover* the written language and *consider* the spoken language". On the other hand, it does not mean that facts about spoken language do not occur or are not described. If a linguistic item or phenomenon is found exclusively or predominantly in the spoken material it is certainly interesting and merits description. Rather, it is the other way round that may be problematic. If a linguistic item or phenomenon occurs rarely or not at all in the spoken material, does it then follow that it is characteristic of written language? Not necessarily. We would expect the finding to be represented in the spoken material by a frequency of only one fifth anyway, so the number of absolute occurrences must be substantial before absence in the spoken material is significant from a purely statistical point of view. In addition, the absence might be explained by another, conditional variable such as formality or age of the language user as the composition of the spoken material deviates somewhat from that of the written.

The practical consequence of this is that the lexicographer tends to be much more cautious in using a label such as "particularly in written language" even if no example is in fact attested in the spoken material. As a minimum, the lexicographer would have to check against his or her own intuition as a native speaker and to consider other possible lables. Examples in the dictionary of words that have been described as particularly characteristic of written language include the following: **førstnævnte** ('first mentioned', 'former'), **sidstnævnte** ('last mentioned', 'latter'), **andetsteds** ('elsewhere'), and synonym pairs with a marked preference for either of the expression types have been labelled accordingly: **samt** ('and') and **ofte** ('often') are more common in written language than the near-synonyms **og** and **tit**.

Nowadays, many dictionaries claim to be corpus based and descriptive, even if it can in many cases be questioned how this should be understood. Probably, most people would take it to mean that the dictionary truthfully reflects the language as it is being used by its speakers without rejecting attested material for being wrong, ugly, offensive, politically incorrect, or including unattested material for the opposite reason. But obviously, it does not suffice to use a corpus and describe it truthfully if the corpus is in fact not representative of the language which it is supposed to cover. The question of representativeness is an old one in corpus linguistics and one which I shall not engage in here, but when it comes to spoken language it is probably where most corpora are most remarkably biased. If a truly representative corpus should reflect proportionately the total number of language tokens being produced within a given period of time, spoken language would probably make up well over half of the material. I am not aware of the exact figure of the proportion, but even if it should be known, it is by no way obvious how corpus representation should mirror the proportion. For instance, shouldn't public language carry a greater weight than private as more people are exposed to it? You could argue that representativeness is not necessarily the same thing when viewed from a language consumption as opposed to a production point of

view (cf. Rundell & Stock p. 49). However interesting from a theoretical point of view this discussion is, lexicographers are also practical workers with a budget and a deadline and have to take these more mundane matters into account.

The overall guiding principle for corpus composition has been the broadest possible coverage (for a more detailed exposition, see e.g. Asmussen & Norling-Christensen 1998). This also holds for its spoken part, but an important additional factor has been, for obvious economic reasons, availability. The editorial team have carried out some of the transcription work themselves, but the majority of the material was generously donated in electronic format from various institutions and individuals. We have received radio and television programmes from the Danish Broadcasting Corporation, transcribed sociolinguistic and sociological interviews from university colleagues, unedited reports from political debates from the Danish parliament and from the city council of Copenhagen, speeches, lectures, church sermons, various telephone answer services, and even loudspeaker announcements from train stations. So, although the sub-corpus of spoken language is not optimally balanced in every respect, it does include a variety of both uses and users, covering private as well as public language, and the language of experts as well as that of laymen.

With nearly 8 million words of semiscripted and unscripted spoken words the corpus was at the time of its compilation the largest spoken corpus in Northern Europe, and even today it is still comparable to, for instance, the 10 million words of the British National Corpus.

3. Lexical items

At the lexical level, a substantial number of words can be found that are certainly characteristic of spoken language. However, it does not follow that words of this kind will occur in a spoken language corpus only. On the contrary, many of them are familiar words that are frequent in written texts as well, and the reason is, of course, that it is common to reproduce spoken language in the written medium, in fiction as well as in journalism. An interesting question is therefore: are there any words (or linguistic sound units) that are genuinely oral in the sense that they occur only in the spoken medium? In our experience, the answer is tentatively affirmative. We have indeed included a number of words and phrases that we have found almost exclusively in the spoken corpus. These have been marked in the dictionary with the comment "particularly in spoken language", and in the following, the most conspicuous groups of lexical items are examined in more detail.

Most obvious are *interjections* and *onomatopoeic words* which are almost by nature confined to spoken language. With the above-mentioned reservation in mind, we have recorded some interjections not found in any other dictionary, e.g. **ad** signalling disgust ('yuck'), **arh** or **ahr** signalling hesitation or doubt, and the positive response particle **jaha** ('yeah').

One highly significant characteristic of spoken language is its volatility. Once a word or a chain of words has been uttered, it is normally gone and cannot be retrieved – it was never meant for storage. Consequently, an important keyword in this connection is conventionalization. As far as sound-words are concerned, it is required for a sound or an exclamation to become part of the lexicon that there is agreement in the language community on the expression side (the *signifiant* in Saussurean terms) of the utterance, i.e. how it is

rendered as a linguistic item. It is important to realize that the items contained in a corpus of spoken language are not spoken words proper, but rather transcriptions of spoken words. Therefore, an affirmative answer to the question above is in the strict sense a contradiction in terms: once the utterance is put on paper and occurs in, say, a language corpus or in a dictionary (i.e. a written medium), it is obviously no longer 'genuinely oral'. So we need to recognize a conventional way of writing an utterance before we can consider it a candidate for inclusion in the dictionary. Such conventionalizations are moreover often language specific, which is probably the reason why words for the same sound may vary from one language to the other. For instance, the English particle of assent *uh-huh* is written *aha* in German, Spanish and other languages, even though the physical sound is probably very much the same. However, once the convention has been established, there is nothing to prevent an impact from the convention on the spoken form such that, for example, the English language community agrees that the sound uttered by the male hen is spelled, by convention, *cock-a-doodle-doo* and that – by a subsequent convention – the lexeme is pronounced correspondingly, whereas the Germans agree on *kikeriki*, the French on *cocorico*, the Russians on *kookarekoo,* and so on. Similarly, languages have different words for yawning, sneezing, gasping, farting etc. even though they are motivated by the same bodily sounds. Because of the requirement of a certain degree of conventionalization, the number of words found exclusively in spoken language is rather limited. Good writers have a subtle feeling for language and are often keen observers of linguistic behaviour. And they make use of these observations when they let their characters engage in dialogues in fictive texts. Indeed, it is likely that authors play an active and important role in the conventionalization process itself. A very good place to look for evidence is in comic strips. This genre often depicts informal conversations resembling everyday real-life situations, and the skilled cartoonist may succeed in creating new conventions lending a highly personal flavour to the strip, or as the ultimate success even in creating new words. I personally associate words like *aughh* and *bleagh* with Charles M. Schultz' *Peanuts*, and words like *hrmpf* and *wak* (or *uak*) are Barksisms, I believe. The Danish word *bvadr* (an interjection indicating disgust), which has now become an established word found in several Danish dictionaries, was in fact created in 1960 by the translator of *Peanuts* as a representation of *bleagh.*

If we accept a less strict criterion and define spoken language as remarks being produced by one or more speakers as monologue or dialogue (including written versions of them), the number of lexical items of lexicograhic relevance increases considerably.

Another group of items are *discourse markers*, i.e. words and expressions that bracket units of talk occurring outside the propositional content of a sentence (cf. Schiffrin 1987). Their primary function is to indicate the relationship between speaker and hearer, or between speaker and text. In English they include items such as *well, y'know, oh, now, then, so* and *I mean*, and arguably also text coherence markers like *as I just said, on one hand .. on the other* (these and many of the following examples are described at the semantic level. As they are often multi-functional or polysemous, they may well occur in more than one category, and they may have additional general senses not accounted for here). Related to this group are *fillers*, words that speakers use to keep the floor while planning the utterance ahead (in English words like *er, mmh, well*, in Danish words like *æh, hmm, tjah*), and *tag-*

questions (in Danish *ikk'* or *ikk' også* (as in *hun kommer, ikk'?* 'she will come, won't she?') and *vel* (as in *hun kommer ikke, vel*? 'she won't come, will she?')). It should be stressed that the categories as such are by no means confined to spoken language: textual coherence markers, for instance, are clearly common in written texts. However, the point is rather that they are not the same; the inventory seems to have a clearly delimited subset characteristic of spoken language. In The Danish Dictionary we have included from this group items like **altså** ('then'), **du ved** ('you know'), **for øvrigt** ('by the way'), **hva'** ('huh?'), **hvordan det** ('how's that'), **jeg mener** ('I mean'), **lissom** ('like'), **om jeg så må sige** ('so to speak'), **se nu ..** ('take ..'), **se så** ('there now'), **så** ('then, so'), and **så'n** ('like, kind of').

Pragmatic phrases are much more frequent in conversation than in written texts. Conversations are social acts where one must give feedback constantly and assure the interlocutor of one's interest, sympathy and agreement. Some of this takes linguistic form, often as formulaic phrases which are exchanged according to conventional rules of appropriateness. Even if they are often semantically transparent, their pragmatic linkage to particular conversational situations has made us include a number, for example: **det kan du bande på** ('you bet'), **det er jo det** ('that's it'), **en gang til** (lit. 'once more', that is 'beg your pardon, sorry'), **klart** (lit. 'clearly', that is 'sure', cf. German **klar**), **det skal jeg love for** ('I'll say so, you bet'), **jeg kan godt sige dig** ('I tell you'), **det må du nok sige**('you can say that again'), **det må jeg sige** ('what d'you know, well how about that'), **det siger du ikke** ('you don't say'), **nej, ved du nu hvad** ('come on'), **helt ærligt** ('honestly').

As conversation is an activity that takes place in time and space and involves two or more speakers, it is no surprise that *deictic pronouns* and *adverbs* occur frequently in this text type when speakers refer and orientate themselves in such a setting. Again, we find a subset which is confined to the spoken medium: ***hersens*** and **dersens**, **den her** and **den der** (both pairs meaning 'this' and 'that', with additional connotations of informality and reservation, cf. the English use in narrative: *I looked up and saw this huge bloke coming towards me*), and the adverbs **henne** (indicating location away from the speaker) and **her** ('here' used in a special time sense as in *her til foråret* 'this (not so distant) spring'). For the groups mentioned so far, and for the following groups in particular one could argue that they might as well have been labelled 'informal'. And it is true that informality is a recurrent feature of most, if not all, of the lexical items under discussion. One might therefore speculate if informality is an intrinsic feature of spoken language, but clearly this is not the case: the spoken medium is represented in the corpus by a variety of genres, some of which certainly are rather formal, like sermons, parliamentary debates, speeches and railway announcements. So, instead one must content oneself with noting that the relevant lexical items seem to originate in informal conversation like the personal interviews and radio and television programmes.

Another recognizable group are *swearwords*. Again, these words are not confined to the spoken medium, but they are, like interjections, much more frequent here and display a wider and more varied range of elements than is commonly found in written texts. We include among others **allerhelvedes**, **dæleme**, **eddermame**, **hammer-**, **knageme**, **kraftedeme**, **kraftstejleme**, **pokkerme**, **saftsuseme**, **sateme**, **sgisme**, **sørenjenseme** (not translatable one by one, but they serve of course the same function as *bleeding, God damn it*, etc. in English) The same is true of *slang* and *colloquialisms*, and a possible reason is the

informality and volatility associated with spoken conversation. Many, if not all, neologisms of this nature originate in informal conversation where speakers test their linguistic creativity. Some of them may remain nonce creations, whereas others gradually become integrated as an established part of the lexicon, but we find them all in the spoken medium first. Still others are not neologisms, but have not managed to diffuse from the spoken to the written medium. This is to some extent true of a number of colloquialisms, and it is probably the informal nature of these expressions that makes them remain in the spoken medium. Examples include expressions like **og alt sådan noget** ('or something, and that kind of thing'), **i den dur**, ('along those lines'), **ikke en fis** ('not .. a shit'), **gider du lige** ('give me a break'), **jeg skal give dig .. skal jeg** ('I'll give you ..'), **noget i den stil** ('something like that'), **går den, så går den** ('if it comes off'), **hvad hulen** ('what the heck'), **du skulle snakke** ('you're a fine one to talk'), **skulle jeg hilse og sige** ('let me tell you'), **skulle du spørge fra nogen** ('who wants to know, what business is that of yours'), **langs ad vejen** ('as you go along, bit by bit'), **det er for vildt** ('wicked!').

One last group of words has been labelled as spoken language, but it can be questioned if this is really justified. The group consists of dialectal words and regionalisms. These are strictly speaking beyond the scope of the dictionary, but naturally they show up in some of the interviews conducted in different parts of the country. A few of them have entered into the dictionary because their use has extended beyond the local origin and into the general language. Even if it is true, as seen from the editor's point of view, that they are found exclusively in spoken texts, it would perhaps be more appropriate to label them according to their geographical origin.

4. Grammatical characteristics

The Danish Dictionary brings information on a number of grammatical relations, including valency, syntactic patterns and other constructional information. Some of the notes on grammar also specify that a certain construction is particularly frequent in spoken language. Examples are word order (e.g. *en til kage* rather than *en kage til* 'another cake'), construction with or without the infinite particle *at* (as with the semi-modal auxiliaries *turde* 'dare' and *gide* 'feel like' which occur more frequently with the particle in spoken Danish), repetition used as intensifier (*slet, slet ikke* 'not at all') and older forms surviving in the spoken medium (*i steden for* rather than *i stedet for* 'instead of'). Needless to say, the use of the label is quite restricted as focus in a dictionary is on the lexical level, and we respect the division of labour between grammar and dictionary. So, for instance, the widespread use of main clause word order in subordinate clauses receives no comment as it belongs to another level of language description than the lexical.

5. Quotations

Apart from the question of availability, the main reason why the relationship between spoken language and dictionaries is not always a heartfelt one, is probably the difficulty involved in getting suitable examples out of a spoken language corpus. Many of the features of spoken language are intrinsically bound to the medium, and when transformed into the written medium, a spoken utterance inevitably looks hopelessly halting. For that reason, it is often impossible to excerpt readable quotations in unaltered form from the spoken language

corpus, not least because we have adopted a policy of bringing authentic quotations supplied with the source of origin, i.e. the editors should only change a quotation if they explicitly indicate that an alteration has taken place: two dots indicate that something has been omitted, whereas text in square brackets signals that the content has been changed, added or reformulated in relation to the original text. In addition to these general principles, the editors have been allowed to change a quotation from a spoken source if the change has to do with transcription irregularities (unauthorized spelling, punctuation, hyphenation etc.) that cannot be attributed to the original utterance, and also where the transcriber has added metacomments to the text. An example may illustrate the case: in the article **afslutning**, sense 3 (here 'end-of-term') we find the following quotation:

> *den 10. juni har vi afslutning .. hvor vi får vores uddannelsesbevis overrakt*
> talespKbh87
> (June 10 we celebrate end-of-term .. where we receive our certificate of education)

The passage where the quotation is taken from reads in full:

> <replik id=ZZ1> ja den tiende juni har vi afslutning simpelthen hvor vi får {tøven} vores {pause} uddannelsesbevis overrakt med {pause} karakterer og alt det fine der som man skal have {pause} og så kan man så g- {pause} gå ud og håbe på man kan få et job {pause} </replik>

In the quotation a word has been omitted (the filler *simpelthen*) and replaced by two dots, and two metacomments (on hesitation and pause) have simply been left out.

Obviously, too many breaks with dots or brackets disturb the reading and should preferably be avoided. As a general editorial principle, not more than a single, or in rare cases two breaks, has been allowed. Given this, it is hardly surprising that the proportion of spoken quotations does not reach the 17-20 per cent which the statistics would suggest. Nevertheless, there are altogether 7,148 quotations taken from spoken language out of a total of c 100,000 in the dictionary, or about 7 per cent. That is not at all a negligible figure and certainly adds to the flavour of the dictionary: on average, almost two quotations on each page come from the spoken language corpus.

6. Conclusion

No doubt, it is tedious and time-consuming to put together a corpus of spoken language. It is also true that spoken passages are not always very suited as authentic examples in transcribed, written form. Having said that, however, there can be no doubt that inclusion of spoken Danish has contributed significantly to The Danish Dictionary. 315 explicit notes referring to spoken language tell us that a dictionary is not complete without accounting for it. And with more than 7,000 authentic quotations from spoken sources the user will certainly notice and, we hope, also appreciate this often-neglected, but by no means rare variety of language.

References

Asmussen, J. and Norling-Christensen, O. 1998. 'The Corpus of the Danish Dictionary', in *Lexikos 8 (AFRILEX-reeks/series 8).* (pp. 223-242) Stellenbosch.

Landau, S. I. 2001. *Dictionaries: the Art and Craft og Lexicography.* (second edition) Cambridge: Cambridge University Press.

Moon, R. 1998. 'On Using Spoken Data in Corpus Lexicography', in T. Fontenelle et al. (eds.), *EURALEX' 98 Proceedings.* (Volume II, pp. 347-357) Liège.

Rundell, M. & Stock, P. 1992 (October). 'The Corpus Revolution', in *English Today 32.* (pp.45-51)

Schiffrin, D. 1987. *Discourse Markers.* Cambridge: Cambridge University Press.