In: P. Juel Henrichsen & P. Rossen Skadhauge (eds.) (2006) Treebanking for Discourse and Speech. *Copenhagen Studies in Language*, *32*, Copenhagen: Samfundslitteratur Press, pp. 29-42

Treebanks for spoken language – some reflections

Jens Allwood, Department of Linguistics, Göteborg University

1. Why treebanks for spoken language?

Multimodal face-to-face communication involving spoken language and gestures is the basic form of communication for human beings. It has evolutionary primacy over written language and is probably genetically encoded. To have access to a well structured database containing spoken language would be of great value both for the development of empirical resources for linguistics and for the development of linguistic theory, in order that it might more adequately describe and explain the nature of language and of linguistic interaction. It is likely that such descriptions and explanations then, in turn, will lead to the development of new applications involving language and communication technology (Allwood, 2001a).

2. What is a corpus of spoken language?

Let us define a "computerized multimodal corpus of spoken language" as a collection of instances of multimodal spoken language communication, organized in a database.

Although this definition might seem straight forward, it involves a problem in deciding on what is to count as instances of spoken language data. Is the corpus going to be a corpus of digitized audio and video files or is it going to be a corpus of transcriptions of the audio and video files, i.e. a corpus of texts representing the audio and video files in a written format?

If the former option is chosen, one of the consequences will be that we will have continuous chunks of speech without distinct segmentation of sentences, phrases, words or phonemes. It will also be the case that we will have data on the communicative body movements (hand gestures, facial gestures, head movements etc.) that normally are part of spoken language interaction.

If the second option is chosen, what we will have will be a text segmented into words by space, possibly segmented into morphemes (if this is added in the transcription) and possibly with additions that enable us to graphically represent such phenomena as intonation, pause, stress, overlap, etc.

3. What is a Treebank for a corpus of spoken language?

Depending on what type of primary data the corpus of spoken language consists of, the question of what constitutes a tree bank can be answered in fairly different ways.

If the corpus primarily consists of digitized sound and video files, the first step will have to be to find or to impose some sort of structure on the data, i.e. segmentation in terms of, for example, acoustic features, phonic n-grams, prosodic features, syllables, morphemes, words, phrases, utterances, gestures or some other sort of unit. The properties and/or units that are chosen will then decide on what further structure can be imposed on the data. Traditional grammatical analysis in general, (e.g. parts of speech) will presuppose words, if not morphemes, so that if, for example, utterances rather than words are chosen as primary units, it is not clear that any traditional grammatical categories can be used. This will also be the case if acoustic features, prosodic features, phonic n-grams, syllables or gestures are chosen. We do not, at present, have any way to make meaningful sense of such categories without units like morphemes or words. In spite of this difficulty, in the long run it would be desirable to create spoken language corpora based on digitized audio and video data. A crucial question in this connection will be if we can find new ways of automatically finding/imposing meaningful structure based on properties of the speech signal. (Can we learn anything from speech recognition?) Analytical work of this type is really the only way to guarantee that we are doing justice to the meaning bearing structure that is really present in authentic interactive face-to-face spoken language communication.

This means that we must gain a more complete understanding of what is presupposed by segmentation into words, morphemes, utterances (e.g. speaker change) and gestures. Are acoustic/optic features of some sort both necessary and sufficient for finding segmental and non-segmental meaning or are they merely necessary features having to be complemented by some analog of human perceptual, cognitive processing in order to yield a meaningful output. How can we best make use of properties and units like acoustic features (for n-grams and prosodic features), syllables, utterances and gestures.

However desirable a more direct analysis of spoken language would be at the present time, most spoken language corpora exist in transcribed form as a form of written language. This means that a lot of structure normally present in written language has already been assigned to spoken language. There is therefore, of course, a risk that such a corpus, in certain ways, is as an artifact with properties which are not necessarily a true part of spoken language.

But even if all transcriptions of spoken language, to a certain extent are artifactual (e.g. they usually include space between words), they can be more or less close to written language. They can be written with normal standard written language orthography, or possibly with something closer to phonetic representation. They can include punctuation marks, capital letters etc. or they can attempt to impose less written language structure by excluding punctuation and capital letters. They can also try to include some features of spoken language, not normally included in written language, such as intonation, voice quality, stress, pauses, overlaps, gestures, etc. (e.g. Allwood et al 2000). Some of these extra features like intonation, voice quality and gestures, have turned out to be hard to represent graphically and hard to transcribe reliably. Other features, like pauses and overlaps, are somewhat easier to handle. Whatever is chosen, the question of whether what has been chosen is artifactual remains relevant. It is also clear, that what further structure may be assigned to the corpus will depend on the properties which, in this way, are present as primary given data.

4. Tree bank annotations for a corpus of transcriptions of spoken language

Some of the ways a corpus of transcriptions of spoken language can be annotated are the following:

1) Parts of speech (presupposes word segmentation)
2) Morphological categories (presupposes morpheme segmentation)
3) Phrase categories (presupposes words and possibly phrases)
4) Grammatical functions, e.g. subject, predicate, object (presupposes words and sentences)
5) Other dependencies within utterances, e.g. semantic roles (presupposes some type of unit, e.g. morphemes, words or phrases, but could also involve gestures or utterances)
6) Communicative acts/functions (presupposes utterances and gestures and possibly words)
7) Dependencies between utterances (presupposes communicative acts/functions)
8) Exchange types (presupposes communicative acts/functions)

Examples of many of these annotations of spoken language corpora can be found in Allwood (2001b), Allwood et al. (2002), (2003), Allwood, Grönqvist et al.(2003), Allwood, Juel Henrichsen et al. (2003,) Grönqvist and Gunnarsson (2003) and Nivre and Grönqvist (2001).

5. Examples of annotations

Below we will now present some examples of annotations of spoken language that could possibly be used in a treebank. The transcriptions are taken from the GSLC, a corpus based on about 30 different types of social activities, cf
http://www.ling.gu.se/projekt/tal/index.cgi?PAGE=3&SUBPAGE=3
The annotations exemplified are:

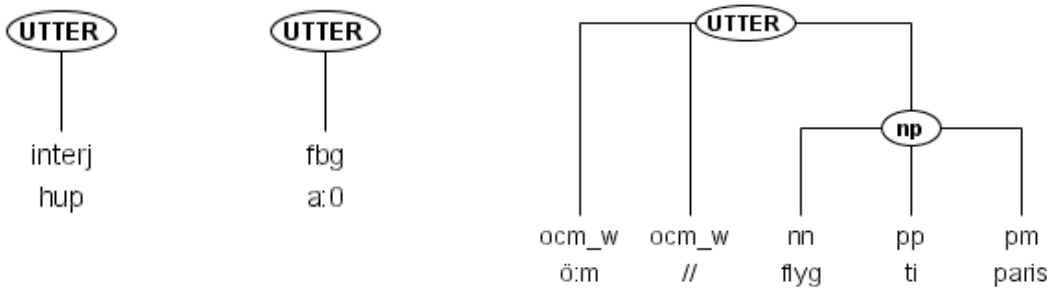
(i) Parts of speech
(ii) Phrase categories
(iii) Communicative acts/functions
(iv) Dependencies between utterances
(v) Exchange types

Example 1. Automatic parts of speech coding of a travel bureau dialog

Dialog with translations		Dialog with parts of speech		
A:	hup (hup)	A:	hup	(interj)
B:	a: (yeah)	B:	a:	(fb)
A:	ö:m // (ehrm //)	A:	ö:m	(ocm)
	flyg ti Paris (flights to Paris)	A:	//	(ocm)
			flyg	(nn)
			ti	(pp)
			paris	(pm)

A customer A enters a travel bureau and requests information about flights to Paris. The parts of speech coding introduces two new paths of speech, fb (feedback) and ocm (own communication management). Ocm includes words, but also pauses, //. The dialog can also be annotated with phrases structure categories as in example 2 below.

Example 2. Utterance based phrase structure categories



Another option is to annotate the dialog with communicative acts: see example 3 below.

Example 3. Communicative acts/functions in a travel bureau dialog

Flight to Paris

Dialog utterances	**Communicative acts/functions**
# 00:00:00	
$P1: hup	Summons/Request for contact
$J1: [1 {j}a:]1	Acceptance (P1)
$P2.1: [1 ö:m]1 //	Hesitation/Keep turn +
$P2.2: flyg ti{ll} <1 paris >1	Request for information/Statement of main task/ statement of main information need
@ <1 name >1	
$J2.1: mm <2 >2 <3 /	Hesitation/Acceptance of task(P2.2) +
$J2.2: ska [2 du ha:]2 en returbiljett >3	Question/Request for specification of type of ticket
@ <2 event: P opens her bag >2	
@ <3 event: people are talking in the background >3	
$P3: [2 ö:{h}]2	Hesitation
$P4: va{d} sa du	Request for clarification(J2)/Question
$J3: ska du ha en tur å0 retur	Answer(P4)/Clarification(J2)/Repetition(J2)/ Question
$P5.1: ja <4 / >4	Answer(J3)/affirmation(J3)+ specification(J2,J3)

Example 4. Dependencies between utterances in a dialog between a customer and a cashier in a supermarket

Yet another option is to try to represent the dependencies between the communicative acts in the dialog. Consider the examples below, where we first present a dialog with English translation and then in example 5, a coding of such dependencies.

Dialog between Customer and Cashier in Supermarket

$A: hej	(hi)
$B: va{r} de{t} bra så	(will that be all)
$A: ja	(yes)
$B: hundrasjutton kronor tack	(hundred and seventeen crowns please)
$A: *Action: Payment*	
$B: trehundraåtti{o}tre kronor ti{ll}baka	(three hundred and eighty three crowns back)
$A: tack så mycke{t}	(thanks a lot)
$B: va{r}sågod	(you´re welcome)

Example 5 Dependencies between Utterances in a super market closing dialog

Dialog utterances	**Dependencies between communicative functions**	**Activity link**
	Opening greeting/Request for contact	Activity sequence
2. <-$B: va{r} de{t} bra så	Accepting contact	
2. $B->: va{r} de{t} bra så	Inquiry if service needs are met	Question-Answer
3. <-$A: ja	Affirmative answer	
3. $A->: ja	Readiness for continuation	Activity sequence
4. <-$B: hundrasjutton kronor tack	Continuation	
4. $B->: hundrasjutton kronor tack	Request for payment	Activity sequence
5. <-$A: *Action: Payment*	Non verbal, Payment	
5. $A->: *Action: Payment*	Non verbal, Payment	Activity sequence
6.<- $B: trehundraåtti{o}tre kronor ti{ll}baka	Indirectly acknowledging payment	
6. $B->: trehundraåtti{o}tre kronor ti{ll}baka	Signaling return of change	Transfer-Acknowledgement/ Gratitude
7. <-$A: tack så mycke{t}	Signaling that change has been received/Thanking	
7. $A->: tack så	Thanking	

mycke{t}		Gratitude-Acknowledgement
8. <-$B: va{r}sågod	Saying "you´re welcome"/Ending interaction	

The dependencies are coded so that every utterance is repeated twice, first as linked backwards and then as linked forward in time. Many of the links and dependencies result from the fact that the utterances are successive actions in the subactivity of paying and leaving a supermarket. Some, in addition, strengthen the link or dependency given by the activity by inserting questions, greetings or transfers of goods that require answers, greetings or expressions of gratitude in return. When a particular type of communicative act regularly evokes a particular type of responsive communicative act this might be seen as an interactive unit type (exchange type)

One or more of the exemplified types of coding might be a good choice in creating a larger Treebank of transcriptions of spoken language. In order to be of general use, whatever is chosen should be as "theory-neutral" as possible, without becoming trivial, so that some of the interesting properties of spoken language can be revealed.

6. Some problems

None of the proposed types of coding are unproblematic. For example, as we have already seen, parts of speech tagging presupposes segmentation into words, which is actually quite problematic because of the continuous nature of spoken language.

Below I would now like to briefly discuss some of the problems we have to face in doing analysis of spoken language.

1. Do we have different words in spoken and written language?

As an example of this problem, consider table 1.

Table 1 Words in spoken and written language

Spoken	English	*Written*	
de (94%)	(it)	det	(3.4%)
ja (94%)	(I)	jag	(5.1%)

The table shows that the most common word in spoken Swedish is pronounced *de* (94%) and *det* (3.4%). *Det* is the written form. The spoken word *ja* and the written word *jag* have a similar relation. The following question may now be raised as to whether we have one or two more basic word forms. If we assume that we have one basic word form, this could either be the spoken language forms (*de* or *ja*), with written language additions, for illiterates and preschool children, or the written language forms (*det* or *jag*), with spoken language reductions, for literates. If we assume that there are two related basic words, we have to have an account of how they are related. Are all three options possible varying with speaker, or should we choose one of these options?

2. What should we do when one or the other mode (speech or writing) does not uphold distinctions made in the other mode? Consider the example in table 2.

Table 2 Different marking of verbal grammatical categories

	Spoken	*Written*	English
Pret	ja stanna	jag stannade	(I stopped)
Pres.	ja stanna	jag stannar	(I stop)
Int.	ja ville stanna	jag ville stanna	(I wanted to stop)
Imp.	stanna	stanna	(stop)
Infinitive marker	å	att	(to)
Subordinating conjunction	att	att	(that)

In the first case, distinctions which are made in written language, are not made in spoken language and in the second case distinctions which are made in spoken language are not maintained in written language. The problem already noted in the case of spoken and written language words reappears. Should one system be derived from the other or should we somehow make room for parallel variants.

3. Do we have the same parts of speech in spoken and written language?

Certain communicative functions are very much more common or almost exclusive to speech. For this reason, their role has been downplayed in traditional grammatical analysis. Two such functions are feedback (fb) and Own communication management (OCM). Spoken language contains many feedback words like the following; *ja, jaha, ha, a, jo, joho, ho, o, nä, nähä, hä, ä*, and OCM words like the following; *eh, e, äh, ä*. The translation of feedback words is difficult, since there are no exact equivalences in English. The words exemplified above provide functional variants based on the meaning of "yes" and "no", while OCM words express different types of hesitation.

Traditionally, such words have either been denied word status and been labeled extralinguistic or paralinguistic or, if admitted, usually been classified as interjections. Is this sufficient given the very prominent functional role such words play in spoken language?

4. What should be the basic unit of analysis of spoken language, e.g. "sentence" or "utterance" (contribution)?

Traditional grammar is usually based on the word as a unit, while more modern grammars have given the sentence as a unit more importance. A sentence (in its classical form) is usually viewed as consisting of a subject and a predicate verb. The verb is, for this reason, given a central role in many types of modern grammatical analysis. The following two findings from the GSLC Spoken Language Corpus (1.45 million words) are a little problematic to harmonize with the assumption of the central roles of the sentence and the verb in most types of linguistic analysis.

A. Around 25% of all utterances are one-word utterances.
B. Around 40% of all utterances have no verbs.

The first finding shows that many utterances are not sentences, i.e. they have no subject-predicate structure. The second finding seems to indicate that much linguistic structure can be achieved without the help of verbs. There is therefore reason to consider whether the "utterance" (or better "contribution", to make room for gestures) might be a better basic category for spoken language than the "sentence".

5. Another problem is presented by the phenomenon of "own communication management", e.g. cases of hesitation or self correction. How are they going to be classified?

6. Tagging for communicative acts/communicative function runs into the problem concerning what tags to choose. There is no generally agreed on finite list. In fact, the best solution might be to allow slightly different lists depending on what activity the spoken language occurs in. The problems connected with choice of types of communicative functions are to some extent also inherited in deciding what types of exchange (e.g. question-answer, statement-confirmation etc.) to code. As with types of communicative function, types of exchange are related to what kind of activity the recorded spoken language occurs in.

7. There are intriguing problems concerning coding dependencies across utterances. There are many such dependencies, e.g. co-reference, dependence of communicative functions, providing information in cases of ellipsis etc. Which dependencies should primarily be analyzed?

8. Providing formal representations for spoken language raises many hard questions, such as: How do we provide formal representation for prosody, for multimodal (gestures) features, for relations between utterances? Are our present notions of syntax, semantics and pragmatics adequate for the task?

9. The last problem, I want to point to is the issue of graphical representation. How should we graphically present prosodic or multimodal features and relations between utterances?

Most existing schemas and graphical tools are based on the notion of sentence and are not easily extended to cover relations between sentences of between utterances. There is a need for more flexible graphical tools, in order to capture prosodic or multimodal features and relations between utterances.

7. Concluding remarks

The creation of spoken language treebanks is an important objective both for theoretical and more practically oriented linguistics. In establishing such treebanks, it is important not to forget digital corpora and their properties. It is also important not to forget that spoken language is multi-modal, making high use of gestures and prosody.

Even if most spoken language corpora today consist only of transcriptions of speech, we are at an early stage of development, where many interesting findings can be made also on this basis. However, there are many open questions, so there is a need for experimentation,

perhaps constructing several treebanks of spoken language, built on different theoretical assumptions.

References

Allwood, J. 2001a. Corpus Based Spoken Language and Computational Research. In Holmboe, H. (Ed) *Nordisk Sprogteknologi: Nordic Language Technology*. Museum Tusculanums forlag, pp. 59-68.

Allwood, J. 2001b. Dialog Coding - Function and Grammar: Göteborg Coding Schemas. *Gothenburg Papers in Theoretical Linguistics 85*, Dept of Linguistics, University of Göteborg, pp. 1-67

Allwood, J., Björnberg, M., Grönqvist, L., Ahlsén, E., & Ottesjö, C. 2000. The Spoken Language Corpus at the Dept of Linguistics, Göteborg University. FQS - *Forum Qualitative Social Research, Vol. 1, No. 3*. - Dec. 2000, pp 22.

Allwod, J., Grönqvist, L., Ahlsén, E.,& Gunnarsson, M. 2002. Annotations and Tools for an Activity Based Spoken Language Corpus. In van Kuppevelt, J. (ed.) *Current and New Directions in Discourse and Dialogue*. Kluwer Academic Publishers.

Allwood J., Grönqvist L. & Hendrikse A.P. 2003. *Developing a tag set and tagger for the African languages of South Africa with special reference to Xhosa*, Journal of Southern African Linguistics and Applied Language Studies 2003, 21(4): 223-237

Allwood, J. Juel Henrichsen, P., Grönqvist, L., Ahlsén, E., & Gunnarsson, M 2003. Transliteration between Spoken Language Corpora: Moving between Danish BySoc and Swedish GSLC. *Gothenburg Papers in Theoretical Linguistics 86*, Dept of Linguistics, University of Göteborg.

Grönqvist L. & Gunnarsson M. 2003. *A method for finding word clusters in spoken language*, in proceedings from Corpus Linguistics 2003: 265-273, Lancaster, March 2003

Nivre J. & Grönqvist L. 2001. Tagging a Corpus of Spoken Swedish, *International Journal of Corpus Linguistics*, Nov 2001, vol 6, issue 1, pp 47-78, John Benjamins Publishing Company.