Treebanks for spoken language – some reflections

Jens Allwood, Department of Linguistics, Göteborg University

Margus Treumuth

Ülevaade artiklist

- räägitakse suulise keele korpuse koostamisest
- tuuakse viis näidet võimaliku märgenduse kohta
- sõnastatakse üheksa probleemi, millega suulise keele korpuse koostamisel tuleb tegeleda

Mida märgendada saab?

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)

Mille kohta näiteid tuuakse?

- Tuuakse näited:

1. Parts of speech
2. Phrase categories
3. Communicative acts/functions
4. Dependencies between utterances
5. Exchange types

- Ei tuua näiteid:

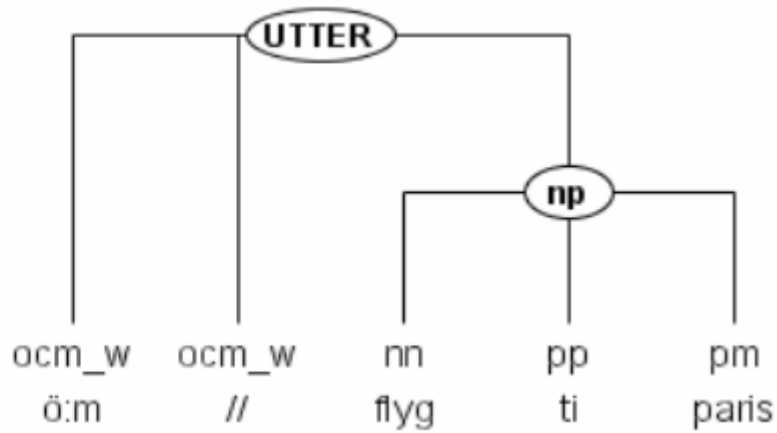
1. Morphological categories
2. Grammatical functions
3. Semantic roles

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)

Kaks uut märki: fb (feedback) and ocm (own communication management).
Ocm includes words, but also pauses, //.

Utterance based phrase structure categories



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

Dependencies between utterances in a dialog between a customer and a cashier in a supermarket

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	

Arutlus

1. Do we have different words in spoken and written language?
2. What should we do when one or the other mode (speech or writing) does not uphold distinctions made in the other mode?
3. Do we have the same parts of speech in spoken and written language?
4. What should be the basic unit of analysis of spoken language, e.g. “sentence” or “utterance” (contribution)?
5. How are features of “own communication management” going to be classified?
6. What tags to choose for communicative acts/communicative functions?
7. Which dependencies should primarily be analyzed?
8. How do we provide formal representation for prosody, for multimodal (gestures) features, for relations between utterances?
9. How should we graphically present prosodic or multimodal features and relations between utterances?

Do we have different words in spoken and written language?

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

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

	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)