Clause Boundary Detection in Transcribed Spoken Language

Fredrik Jørgensen
Department of Linguistics and Scandinavian Studies
University of Oslo
fredrik.jorgensen@iln.uio.no

Abstract

We argue that finite clauses should be regarded as the basic unit in syntactic analysis of spoken language, and describe a method that automatically detects clause boundaries by classifying coordinating conjunctions in spoken language discourse as belonging to either the syntactic level or the discourse level of analysis. The method exploits the special role that coordinating conjunctions play in organizing spoken language discourse, and that coordinating conjunctions at discourse level mark clause boundaries.

1 Introduction

Syntactic analysis of written language rests on the *sentence* as the object of analysis, and aims at describing the inner structure of a sentence. The Cambridge Encyclopedia of Language (Crystal, 1987) states this explicitly:

> "A sentence is the largest unit to which syntactic rules apply."
> (Crystal, 1987, p. 94)

But do sentences exists *per se* in spoken language? Capitalized first words and full stops, which are used to identify sentences in written language, are parts of the writing system. For spoken language, we need to investigate properties of the spoken language systems, e.g. in terms of intonation units (Chafe and Danielewicz, 1987). But the notion of intonation units is problematic:

> "Speakers are sloppy in this respect, often producing a sentence-final intonation before they mean to, or neglecting to produce one when they should." (Chafe and Danielewicz, 1987, p. 103)

Looking at spoken language data, it is hard to find evidence that sentences, as we find them in written language, are the basic units of syntactic analysis. Miller and Weinert (1998) argue extensively that sentences are found in written language only, and that they should be treated as a low-level discourse unit rather than syntactic units. Instead, the *clause* should be regarded as the basic syntactic unit of spoken language (Miller and Weinert, 1998, p. 71). Levelt (1989) also treats clauses as the basic grammatical units, where speaker utterances are partitioned into finite clauses, which in turn may be partitioned into one or more basic clauses (finite or nonfinite).

We adopt the view that clauses are the natural building blocks of a spoken language discourse, and that clauses also should be the starting point for syntactic analysis of spoken language. This view can be summarized by rephrasing The Cambridge Encyclopedia of Language: *For spoken language, the finite clause is the largest unit to which syntactic rules apply.*

Note that although we want to treat the finite clause as the *largest unit* to which syntactic rules apply, this does not mean that the finite clause is the *only unit* to which syntactic rules apply, neither is the finite clause the only unit corresponding to a complete utterance. Utterances may consist of single words or phrases, which may or may not be elliptical from a syntactic or broader communicational or informational perspective. When we speak of clause boundaries in this paper, what we mean is boundaries between segment units or topic units which are maximally a finite clause, but which may consist of smaller syntactic units.

Turning now to coordinating conjunctions, we find that in written language, a coordinating conjunction is a word that links words, phrases, or clauses into compound elements. This is also true for spoken language. But conjunctions have another important role in spoken language, namely organizing the clauses of the discourse, or having other pragmatic functions, e.g. to avoid turn shifts, to mark questions, to express modality etc. Schiffrin (1987) explicitly discusses the English conjunctions *and, but* and *or*, and their role as what she calls "Discourse connectives" These "discourse connectives" or "discourse conjunctions" are better understood at the discourse level, rather than at the syntactic level. Conjunctions also relate larger parts of the discourse, which are not necessarily adjacent to the conjunction (Webber et al., 2006). For instance, the Norwegian and Swedish conjunction *men* ('but') can be used to signal a shift from a digression and back to the main topic of the conversation (Svennevig, 1999; Horne et al., 2001).

As coordinating conjunctions organize the clauses of the discourse, they also turn out to be potential indicators of clause boundaries within an utterance. Detecting clause boundaries may be important for several linguistic disciplines, such as morpho-syntax (e.g. Part of Speech tagging), syntax (e.g. parsing and (semi-)automatic treebank construction) and semantics (in dialogue systems, where e.g. propositions, events or speech acts are necessary for semantic representations).

In this paper, we propose a method to automatically identify clause boundaries in a spoken language corpus of Norwegian. The method described is a supervised machine learning method, where we propose a partial solution to clause boundary detection by reducing the task to a classification problem, classifying conjunctions as either belonging to the discourse level (clause boundaries) or the syntactic level (linking sub-clausal elements). The clauses may in turn be combined into larger discourse structures, or used to define discourse relations, based on the discourse conjunctions and any other feature of the discourse or insight from pragmatic theory. However, these tasks are beyond the scope of this paper.

2 Clause Boundary Detection

The clause boundary detection experiment is structured as follows:

1. Assign a category type (discourse or syntactic) to a set of coordinating conjunctions in a spoken language corpus
2. Extract a set of features from the context of the conjunctions
3. Apply a machine learning method (memory based learning/TiMBL)
4. Evaluate the results (using ten-fold cross validation)

2.1 Data: The NoTa Corpus

The NoTa corpus[1] is an on-line corpus of approximately 1 million words of transcribed spoken Norwegian. Approximately 80.000 words have been manually Part of Speech-tagged, and the current experiment has been run on this part of the corpus.

The discourse in NoTa is divided into *turns* and *segments*, where segments are the lowest discourse units, corresponding to single utterances. Segments in NoTa correspond quite closely to the intonation units of Chafe and Danielewicz (1987), but are based on a combination of intonation, pauses and length of the utterance. Segment may include several clauses, and in some cases clauses are spread throughout more than one segment.

2.2 Annotation of Conjunctions

We have already described two of the conjunction types in this experiment; *Discourse* and *Syntactic* conjunctions. In addition to these two main categories of conjunctions, I've included a third category, *Indeterminable* conjunctions. This is due to the discourse particle *sånn* ('like', 'you know', 'stuff', 'that' etc.), which is very frequent in spoken Norwegian (ranked 12 in spoken Norwegian and 575 in written Norwegian). *Sånn* may function as a pro-word for a number of phrase types, and phrases

[1] http://www.tekstlab.uio.no/NoTa/Oslo/index.html

starting with *sånn* belong to the syntactic level. But the attachment site if *sånn* phrases is often difficult or even impossible to determine. In (1), *sånn* may be conjoined to the VP (*lekte sisten*/'played tag') or the NP (*sisten*/'tag').

(1) fløy etter hverandre og lekte sisten og
ran after each other and played 'tag' and
sånn
stuff

Due to these complications, any conjunction followed by a phrase where the first word is *sånn* is classified as an *Indeterminable* conjunction. Thus, all conjunctions in the training data are annotated belonging to one of the three categories:

Syntactic Conjunctions: Conjunctions that combine two syntactic constituents below the finite clause. This category is assigned to all conjunctions where two or more conjuncts are identified in the context. Whenever the conjuncts are not identifiable in the context, the conjunction will *not* be annotated as a Syntactic Conjunction, even though the missing conjunct may be due to e.g. self-interruption. This in order to avoid speculations and arbitrary decisions about the speakers' intentions.

Indetetminable Conjunctions: Conjunctions followed by a phrase where *sånn* is the first word.

Discourse Conjunctions: All other conjunctions. These conjunctions may combine clauses, may be discourse fillers etc. These conjunctions all share the property that they do not disrupt a syntactic constituent below the clause level.

In total, 853 conjunctions in the NoTa were assigned one of these three categories.

2.3 Feature Set

Decision on and extraction of the feature set is the core of any application using machine learning methods. The feature set used in this experiment is grouped into one basic feature set, which is incremented with new features, as described below:

Basic: Word form of the conjunction and word form, lemma and part of speech for ± 4 tokens (tag set size = 17). (Features 1-25).
+Match: Binary value stating if the PoS tags and/or word forms of the preceding and succeeding tokens are identical. (Features 26-27).
+Filter: Filter out tokens which does not fill any syntactic role (pragmatic and spoken language words: *interjections*, *conjunctions*, *unknown words*, *pauses*, *punctuation*. (Features 28-35).
+RelFreq: Relative frequencies for the previous word-form and/or PoS ending a segment, and the next word form and/or PoS starting a segment (after applying filtering). Inclusion of relative frequencies has been proved useful for sentence boundary detection in written language (See e.g Stevenson and Gaizauskas (2000)). (Features 36-41).

Number	Feature Set	Description
1	*Basic*	word form of the conjunction
2-9	*Basic*	window of ± 4 word forms
10-17	*Basic*	window of ± 4 lemmas
18-25	*Basic*	window of ± 4 PoS
26	+*Match*	prev word and next word identical?
27	+*Match*	prev PoS and next PoS identical?
28-33	+*Filter*	filtered, ±1 word, PoS and word/PoS
34	+*Filter*	filtered, prev word and next word identical?
35	+*Filter*	filtered, prev PoS and next PoS identical?
36	+*RelFreq*	filtered, rel. freq. for prev PoS ending a segment
37	+*RelFreq*	filtered, rel. freq. for next PoS starting a segment
38	+*RelFreq*	filtered, rel. freq. for prev word ending a segment
39	+*RelFreq*	filtered, rel. freq. for next word starting a segment
40	+*RelFreq*	filtered, rel. freq. for prev word/PoS ending a segment
41	+*RelFreq*	filtered, rel. freq. for next word/PoS starting a segment

Table 1: Feature Set for Conjunction Classification

3 Results

All experiments were run with the memory based learning application TiMBL (Daelemans et al., 2004), using 'modified value difference' as similarity metric and k-value = 3. The results were evaluated using ten-fold cross validation.

For each feature set, *forward search* was applied, an algorithm for automatic feature selection. *For-*

ward search, implemented as follows:

1. Run the classifier and rank the features by a information gain metric.

2. Run the classifier with only the highest ranked feature. For every feature, starting from the top of the ranked list, add this to the feature set and run the classifier again. If the accuracy increases, keep the feature. Otherwise, lose it.

The results with the different feature sets are shown in Table 2. The columns show the feature set, number of features before (*n(all)*) and after (*n(fwd)*) forward search, and accuracy before (*acc(all)*) and after (*acc(fwd)*) forward search.

Baseline is the most frequent class (Discourse conjunctions), and gives an accuracy of 71.16%. The basic feature set gives 89.09%, while the maximum feature set gives 90.85% before forward search and 94.37% after forward search.

The accuracy without forward search increase for each feature set added, in total an accuracy gain of 1.76%. It is interesting to note that after forward search, the basic feature set is the second best, and the total accuracy gain with the maximum feature set is only 0.47%.

In the maximum feature set, the following 13 features were kept after the automatic feature selection (ranked by information gain):

1. *PoS match (filtered)*
2. *PoS match (unfiltered)*
3. *conjunction word form*
4. *relative frequency for next PoS beginning a segment*
5. *relative frequency for previous PoS ending a segment*
6. *+1 PoS (filtered)*
7. *+2 PoS*
8. *-1 PoS*
9. *-3 PoS*
10. *+3 PoS*
11. *+1 word form*
12. *+2 word form*
13. *+3 word form*

Feature set	**n(all)**	**acc(all)**	**n(fwd)**	**acc(fwd)**
Baseline	-	71.16%	-	-
Basic	25	89.10%	4	93.90%
+Match	27	89.68%	6	93.08%
+Filter	35	90.27%	11	93.79%
+RelFreq	41	90.86%	13	94.37%

Table 2: Results for Conjunction Classification

The maximum feature set after forward search has several advantages. Not only does this feature set give better accuracy, it is half the size as the basic feature set, which means less processing time, and it requires fewer training instances. Figure 1 shows the learning curve for the basic feature set before forward search, and the maximum feature set after forward search. Figure 1 implies that for the basic feature set, approximately 600 training instances are needed, while for the maximum feature set after (forward search), 400 training instances are sufficient.

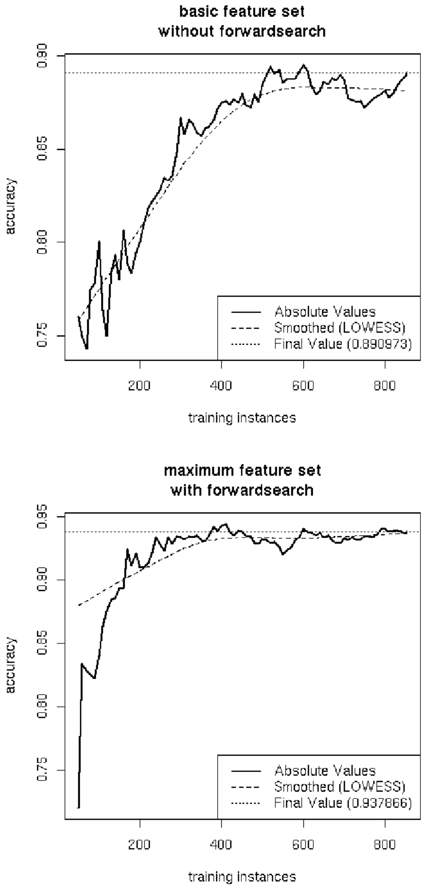

Figure 1: Learning curves for Basic Feature Set and Maximum Feature Set

An example of output of the system is given in (2), where conjuncts marked D(iscourse) are interpreted as belonging to the discourse level, and at the same time functioning as clause boundaries. In this example, the three units are (i) a clause containing a repair, (ii) an interrupted clause or a fragment, and (iii) a finite clause.

(2) **og/D** der bodde jeg til jeg var like
and/**D** *there lived I until I was just*
før jeg fylte seks år **for/D** jeg
before I turned six years ***because/*****D** *I*
hus- **og/D** det husker jeg veldig godt
re- ***and/*****D** *that remember I very well*

'And I lived there until I was just before I turned six. Because I re-. And I remember that very well'

The partitioning into clauses, as shown in (2), may prove useful both for parsing and for detecting so-called "disfluencies" (see e.g. Shriberg (1994)), as the three segments to be analyzed are shorter and simpler than the original segment.

Note that the results reported in this paper only states how many of the conjunctions are classified correctly. It does not state anything about the proportion of clause boundaries correctly identified in the data, as the NoTa corpus is not currently annotated with this information. The proportion of clause boundaries detected in the corpus is crucial to evaluating the usefulness of the method described. Thus, the method described above is only claimed to be a partial solution to the problem of spoken language clause boundary detection.

4 Conclusion

This experiment describes a method for partially solving the clause boundary detection by exploiting the special role coordinating conjunctions play in structuring a discourse, and reducing clause boundary detection to a classification problem. The method gives promising results with relatively few training instances.

References

Wallace Chafe and Jane Danielewicz. Properties of written and spoken language. In S. Jay Samuels and Rosalind Horowitz, editors, *Comprehending Oral and Written Language*, pages 83–113. Academic Press, 1987.

David Crystal. *The Cambridge Encyclopedia of Language*. Cambridge University Press, New York, 1987.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot K, and Antal van den Bosch A. *TiMBL: Tilburg Memory Based Learner, version 5.1, Reference Guide. ILK Technical Report 04-02*, 2004.

Merle Horne, Petra Hansson, Gösta Bruce, Johan Frid, and Marcus Filipsson. Cue words and the topic structure of spoken discourse: The case of swedish men 'but'. *Journal of Pragmatics*, 33(7): 1061–1081, July 2001.

Willem J. Levelt. *Speaking*. The MIT Press, 1989.

Jim Miller and Regina Weinert. *Spontaneous Spoken Language*. Oxford University Press, 1998.

Deborah Schiffrin. *Discourse Markers*. Cambridge University Press, 1987.

E.E. Shriberg. *Preliminaries to a Theory of Speech Disfluencies*. PhD thesis, University of California at Berkeley, 1994.

M. Stevenson and R. Gaizauskas. Experiments on sentence boundary detection. In *Proceedings of the sixth conference on Applied natural language processing table of contents*, pages 84 – 89, Seattle, Washington, 2000.

Jan Svennevig. Talespråket - mellom pragmatikk og grammatikk. In M. Engebretsen and J. Svennevig, editors, *Mediet teller! Tverrfaglige perspektiver på skrift og tale*, pages 101–116. Høgskolen i Agder, 1999.

Bonnie Webber, Aravind Joshi, Eleni Miltsakaki, Rashmi Prasad, Nikhil Dinesh, Alan Lee, and Katherine Forbes. A short introduction to the penn discourse treebank. In Peter Juel Henrichsen and Peter Rossen Skadhauge, editors, *Treebanking for Discourse and Speech*, volume 32 of *Copenhagen Studies in Language*, pages 9–28. Samfundslitteratur Press, Copenhagen, 2006.