Quantitative experiments on prosodic and discourse units in the Corpus of Interactional Data

Klim Peshkov, Laurent Prévot, Roxane Bertrand, Stéphane Rauzy, Philippe Blache
Laboratoire Parole et Langage
Aix Marseille Université & CNRS
Aix-En-Provence, France
firstname.lastname@lpl-aix.fr

1 Introdution

The recent years have seen growing the number initiatives related to the interface between syntax, prosody and discourse. While in English the computational counterpart of this perspective has been largely advanced both from more formal modeling and machine learning perspectives, in French the situation is much less clear. Some automatic tools for analyzing prosody (Avanzi et al., 2010), (Goldman et al., 2007) have been developed but tested so far mostly on monologue data. The determination of the relevant units of the different linguistic domains is a crucial issue for this kind of work. In this poster, we will present a series of quantitative evaluations of the output of various automatic tools dealing with prosody, syntax and discourse.

The data we are using is the Corpus of Interactional Data. This is a corpus made of 8 conversations of one hour involving two speakers.

2 Automatic tools

We ran Analor and Momel-Intsint on our corpus. Moreover, we implemented a version of the Simon and Degand (2011) characterization of the prosodic units by using (as they do) the output of Prosogram (Goldman et al., 2007). In the discussions below, the units of Analor are called *Periods* while the ones of Simon and Degand (2011) are called UIM (Major Intonative Units). We also have Interpausal Units (IPU) as a baseline.[1]

At the syntactic level, at the current stage, we simply used a projection of the punctuation learned on a large balanced corpora.[2] More precisely, from a tagged corpus we have learned where strong (periods, exclamation marks etc.) and weak (commas) punctuations occur.

3 Manual Annotations

Among other linguistic elements, prosodic and discourse units have been annotated in the framework of the OTIM project (Blache et al., 2010).

Concerning prosody, several kinds of segmentation have been produced. Originally experts have segmented about 2 hours of corpus into Accentual Phrases (AP) and Intonative Phrases (IP). More recently, an annotation campaign involving naive annotators has been realized. The whole CID corpus has been double-annotated. The task for naive annotators consisted in marking prosodic boundaries of different levels (*1, 2, 3* ; 0 being the default non-annotated case of no boundary).[3]

Concerning discourse, the annotation campaign also involved naive annotators that have segmented the whole corpus (half of it being cross annotated). This was realized thanks to a discourse segmentation manual, inspired by (Afantenos et al., 2010) but largely adapted to our interactional spoken data and simplified to be used by naive annotators. The manual combined semantic (eventualities identification) and discourse (discourse markers)

[1] We do not pretend that these different units are supposed to capture the same prosodic level. We simply want to experiment with the units produced by these tools to decide how to use them later.

[2] This corpus was mostly a written corpus which could be an issue. However, we consider the information captured to be relevant.

[3] This was realized according to a coding manual developed by Roxane Bertrand and Cristel Portes.

Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue, September 19–21, 2012, Paris, France.

		Recall	Precision
IPU	Start	0.415	0.838
	End	0.376	0.736
	Units	0.177	0.245
Period	Start	0.353	0.843
	End	0.339	0.783
	Units	0.153	0.364
UIM	Start	0.478	0.794
	End	0.428	0.710
	Units	0.218	0.360

Table 1: Precision and recall. Reference segmentation: manual IP

	Reference segmentation		
	IP	Discourse	Pseudosyntax
IP	-	0.322	0.603
Discourse	0.238	-	0.435
Pseudosyntax	0.369	0.364	-

Table 2: Divergence between linguistic domains

and pragmatic (recognition of specific speech acts) instructions to create the segmentation. Such a mixture of levels has been made necessary by the nature of the data featuring both rather monologic narrative sequences and highly interactional ones. Manual discourse segmentation with our guidelines has proven to be reliable with κ-scores ranging between 0.8 and 0.85.

4 Experiments

Prosodic units Concerning prosodic units, we first compare the automatic tools to the manual annotation into Intonative Units (See Fig. 1). Overall, IP manual annotation is much more fine grained than the segmentation performed by the automatic tools. It is therefore no surprise to find that precision is rather good (at least for boundary detection) while recall is extremely low. It is noticeable to remark how low are the scores when we shift our attention to unit determination rather than simple boundary detection. By the time of the conference, we will have also compared all the naive annotations (in terms of strength of frontiers) with the automatic tools.

Interfaces In order to try to shade a new light on the interfaces we attempted rather rough quantitative comparison (using the WindowDiff measure (Pevzner and Hearst, 2002)) of the units from the different linguistic domains (See Fig. 2). This was done by using (i) the expert segmentation into Intonative Units, (ii) the manual discourse segmentation and (iii) the projection of the punctuation for the syntactic level.

c_1 (strong punctuation+period)	0.285
c_3 (Pseudophrase+UIM)	0.264
c_5 (strong punctuation+IPU)	0.241
IPU	0.198
UIM	0.217
Period	0.265
strong punctuation	0.419

Table 3: WindowDiff comparison of segmentations combining prosody and pseudosyntax, reference: manual discourse units

Finally, we evaluated automatic tools against the manual discourse segmentation (See Fig. 3). The results is that the IPU baseline provides the closest segmentation to the one of the naive annotators. Quite depressingly, both more sophisticated tools are less related to the manual annotation, adding syntactic information create significant divergences with the manual annotations. Lack of space prevent us from a deeper analysis of this results but the poster will be focused on explaining them and finding solutions.

References

Afantenos, S., Asher, N., et al. (2010). Discourse segmentation manual. Technical report, Toulouse University.

Avanzi, M., Lacheret-Dujour, A., and Victorri, B. (2010). A corpus-based learning method for prominence detection in spontaneous speech. In *Speech Prosody 2010-Fifth International Conference*.

Blache, P., Bertrand, R., et al. (2010). Multimodal annotation of conversational data. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 186–191, Uppsala, Sweden.

Goldman, J., Avanzi, M., et al. (2007). A methodology for the automatic detection of perceived prominent syllables in spoken french. In *Eighth Annual Conference of the International Speech Communication Association*.

Pevzner, L. and Hearst, M. A. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Simon, A. C. and Degand, L. (2011). L'analyse en unités discursives de base: pourquoi et comment? *Langue francaise*, (2):45–59.