Rhapsodie: a Modular approach to the Syntactic and Prosodic Annotation of a Spoken French Corpus

Sylvain KAHANE (Université Paris Ouest, Modyco)
Anne LACHERET (Université Paris Ouest, Modyco)
Paola PIETRANDREA (Università Roma Tre, Lattice)
Frédéric SABIO (Université de Provence, LPL)

The Rhapsodie project aims at annotating and exploiting a corpus of 33000 words of Spoken French with the aim of modeling the interface between syntax and prosody and of identifying the existing correlations between prosodic and syntactic boundaries.

One of the basic tenets of the Rhapsodie approach concerns the modularity of the annotation. The syntactic and the prosodic annotations are conducted separately by two different teams. This approach aims at providing an independent characterization of each of the two levels of analysis to be studied at their interface.

For the prosodic annotation, we have chosen to provide a complete annotation of acoustic prominences. This has led to the identification of three degrees of prosodic boundaries (rhythmic, intonational and periodic boundaries), as well as to tonal annotation and disfluencies annotation. We have developed a specific methodology presented in Obin et al. (2011) for the manual annotation of syllabic prominences and disfluencies. The semi-automatic detection of prosodic boundaries of different ranks is based on the work of Lacheret and Victorri (2002), Avanzi et al. (2008). The automatic tonal annotation has been conducted by P. Mertens with the Prosogram software.

For the syntactic annotation, we have chosen to provide a complete morphosyntactic tagging of the corpus and to annotate the dependency relations between words. The syntactic models adopted for treating our corpus have often needed to be readapted and extended. In particular, we have found that the annotation of the dependencies between words is not enough to account for the whole of connections between speech units characterizing spoken texts and permitting to identify (presumably prosodically marked) syntactic units.

We have therefore extended our syntactic annotation in two directions.

On the one hand we have provided a systematic annotation of all listing phenomena in the corpus. Building on Blanche-Benveniste (1990) we claim that all cases of multiple realization of the same structural position constitute a list: coordinations, repetitions, disfluencies, etc. (Gerdes and Kahane 2009, Kahane 2011, Bonvino et al 2009). We have annotated all lists and tagged them as for their semantic function (e.g., addition, approximation, hesitation, etc.)

On the other hand we have included a level of macrosyntactic annotation in our corpus. For the annotation of this level we have borrowed and refined the apparatus of units defined in the macrosyntactic theories elaborated at Aix-en-Provence (Blanche-Benveniste et al. 1990), at Fribourg (Berrendonner 1990) and at Florence (Cresti 2000) (Benzitoun et al. 2010, Duffort et al. 2010). We have segmented the corpus in macrosyntactic major and minor units: the illocutionary units and their components (noyau, pre-noyau, postnoyau, that we suggest to translate as nucleus, pre-nucleus, post-nucleus). Our annotation schema has allowed us to provide at the same time a segmentation and an indirect tagging of the linear position and pragmatic function of macrosyntactic units.

The first studies conducted at the interface between the syntactic and the prosodic annotation

have made clear that both lists and macrosyntactic units are prosodically marked by prosodic boundaries (Lacheret et al. 2011).

In a sense, the annotation of our corpus is not so distant from the annotation of the Spoken Italian corpus Lablita (Cresti and Moneglia 2005) and the Spoken Brazilian Portuguese Corpus C-Oral Brazil (Mello and Raso 2009). These corpora, treated within the frame of the Language in Act Theory (Cresti 2000) include a thorough annotation and tagging of the prosodic and the macrosyntactic structure.

Two facts, though, distinguish the Rhapsodie experience from the Lablita and the C-Oral Rom Brazil experiences.

First of all, the annotation of the Rhapsodie corpus is based on a genuine syntactic rather than a pragmatic interest. This different perspective has led us to highlight the role played by lists (defined as syntactic objects) in spoken languages, which is quite neglected in the Lablita and C-Oral Rom Brazil experiences. On the other hand, our corpus is not endowed with the rich pragmatic tag set developed for the Italian and Brazilian corpus.

Secondly, the modularity of the annotation led us to develop an independent annotation for the two levels of analysis and to study the prosodic/syntactic interface only at the exploitation stage. This approach is quite different from the Lablita and the C-Oral Rom Brazil approaches which identify macrosyntactic units through a preliminary analysis of prosodic units.

The present work aims at precisely identifying (and accounting for) the similarities and the differences between the pragmatic/prosodic approach and the modular syntactic prosodic approach to spoken corpus annotation.

We have applied our annotation schema to two small samples of the Lablita and the Rhapsodie Corpus which had been previously annotated with the Lablita schema (Cresti et al. 2011a, Cresti et al. 2011b).

On the one hand, we have observed an interesting overlapping between the segmentations provided by the two approaches, which prove the robustness of both of them.

On the other hand, we have observed some differences and some inconsistencies in one and the other annotation that can be overcome by using the criteria adopted within the other framework.

As a result of our analysis, we expect to provide a precise evaluation of the two annotation systems and to indicate whether and to what extent the two approaches can be integrated for a richer and finer annotation of spoken languages.

References

Avanzi M., Lacheret A., Victorri B. (2008) : « Analor, a Tool for Semi-automatic Annotation of French Prosodic Structure » ; Speech Prosody 2008, P. Barbosa (ed.), Campinas, Brésil.

Benzitoun, Ch. Dister, A., Gerdes, K., Kahane, S., Pietrandrea, P., Sabio F. (2010). Tu veux couper là faut dire pourquoi. Propositions pour une segmentation syntaxique du français parlé. Les Actes du Congrès Mondial de Linguistique Française (CMLF 2010), Nouvelle-Orléans, Juillet 2010.

Berrendonner, A. (1990). Pour une macro-syntaxe. Travaux de linguistique, 21, 25-31.

Blanche-Benveniste, C. (1987), Syntaxe, choix du lexique et lieux de bafouillage, DRLAV, 36-37, 123–157.

Blanche-Benveniste Claire, 1990, « Un modèle d'analyse syntaxique "en grilles" pour les productions orales », Anuario de Psicologia, 47, p. 11-28.

Blanche-Benveniste, C. (1997), Approches de la langue parlée en français, Paris, Ophrys.

Blanche-Benveniste, C., M. Bilger, Ch. Rouget & K. Van den Eyende (1990), Le français parlé. Etudesgrammaticales, Paris, Editions du Centre National de la Recherche Scientifique.

Bonvino, E., Masini, F., Pietrandrea P. (2009). List Constructions: a semantic network. Troisième Conférence Internationale de l'AFLiCo, Nanterre. Accessible at http://francescamasini.caissa.it/Presentations_files/parigi_draft.pdf.

Cresti, E. (2000). Corpus di italiano parlato. Florence: Accademia della Crusca.

Cresti, E., M. Moneglia, I. Tucci (2011). Annotation de corpus selon la théorie de la langue en acte F. Lefeuvre, E. Moline (eds.) Unités Syntaxiques et Unités Prosodiques, Langue Française 170, 95-110.

Cresti E. and I. Tucci 2011b The Analysis of the Italian Texts paper delivered at the VI LABLITA International Workshop in Corpus Linguistics Firenze 16-17 juin 2011

Cresti, E.; Moneglia, M. (eds) C-Oral-Rom: Integrated Reference Corpora For Spoken Romance Languages. Amsterdam: John Benjamins. 2005.

Deulofeu J., Duffort L., Gerdes K, Kahane S., Pietrandrea P, (2010). Depends oWhat the French Say Spoken Corpus Annotation With and Beyond SyntacticFunctions. Proceedings of the Fourth Linguistic Annotation Workshop (LAW IV) Uppsala, Sweden 15-16 July 2010Gerdes and Kahane 2009,.

Kahane S., to appear, "De l'analyse en grille à la modélisation des entassements", in Caddéo S., Roubaud M.-N., Rouquier M. & Sabio F. Penser les langues avec Claire Blanche-Benveniste, Aix-en-Provence, PUP.

Lacheret-Dujour A., S. Kahane, P. Pietrandrea, M. Avanzi, B. Victorri, (2011). Oui mais elle est où la coupure là ? Quand syntaxe et prosodie s'entraident ou se complètent. F. Lefeuvre, E. Moline (eds.) Unités Syntaxiques et Unités Prosodiques, Langue Française 170, 61-79.

Lacheret-Dujour A., Victorri B. (2002). La période intonative comme unité d'analyse pour l'étude du français parlé : modélisation prosodique et enjeux linguistiques » . Verbum, M. Charolles (éd), Nancy, 55-72.

Mello, H.; Raso, T. (2009). Para a transcrição da fala espontânea: o caso do C-ORAL-BRASIL. Revista Portuguesa de Humanidadespp. 301=325.

Mertens P. (2004) : "The Prosogram : Semi-Automatic Transcription of Prosody based on a Tonal Perception Model " B. Bel & I. Marlien (eds.) Proceedings of Speech Prosody 2004, Nara (Japan), 23-26 March.

Obin N., Avanzi M., Lacheret A. (2011) : “Transcription of French prosody in Discourse: the Rhapsodie protocole” Colloque IDP, Interface Prosodie-Discours, Manchester, September 2001