Design and Evaluation of Shared Prosodic Annotation for Spontaneous French Speech: From Expert Knowledge to Non-Expert Annotation

Anne Lacheret[1] Nicolas Obin[1, 2] Mathieu Avanzi[1, 3]

[1] Modyco Lab, Paris Ouest University, Nanterre, France
[2] Analysis-Synthesis team, Ircam, Paris, France
[3] Neuchâtel University, Neuchâtel, Switzerland
anne@lacheret.com, nobin@iracm.fr; Mathieu.avanzi@unine.ch

Abstract

In the area of large French speech corpora, there is a demonstrated need for a common prosodic notation system allowing for easy data exchange, comparison, and automatic annotation. The major questions are: (1) how to develop a single simple scheme of prosodic transcription which could form the basis of guidelines for non-expert manual annotation (NEMA), used for linguistic teaching and research; (2) based on this NEMA, how to establish reference prosodic corpora (RPC) for different discourse genres (Cresti and Moneglia, 2005); (3) how to use the RPC to develop corpus-based learning methods for automatic prosodic labelling in spontaneous speech (Buhman *et al.*, 2002; Tamburini and Caini 2005, Avanzi, *et al.* 2010). This paper presents two pilot experiments conducted with a consortium of 15 French experts in prosody in order to provide a prosodic transcription framework (transcription methodology and transcription reliability measures) and to establish reference prosodic corpora in French.

1 Introduction

In this paper the case of the prosodic annotation of spontaneous French speech is discussed. Ever since the ToBI system was introduced in the international speech community (Silverman *et al.*, 1992), it has been considered by some – irrespective of the language to be annotated[1] - as a standard for prosodic annotation, while others contend that ToBI cannot be regarded as a universal annotation tool, *i.e.* it is not appropriate to capture the prosodic properties of certain languages. This is especially true when dealing with spontaneous speech, for which new methods of annotation must be found. In other words, a better prosodic labelling is essential to improve linguistic analyses of prosody (Martin 2003, as well as research in speech technology (Wightman 2002). Linguistics and speech technology have dealt with prosodic transcription from various points of view, which makes a precise definition of the task difficult. An initial distinction can be drawn between (i) phonological approaches (Silverman *et al.*, 1992; Hirst and Di Cristo, 1998; Delais-Roussarie, 2005; etc.), and (ii) acoustic-phonetic prosodic analysis (Beaugendre *et al.*, 1992; Mertens, 2004). Nowadays, these two approaches still remain problematic. The coding schemes of the former reflect not only a specific, and rather narrow, phonological point of view, but also the phonetic poverty of the transcription (most of the time, only information about the fundamental frequency is delivered, and no information regarding intensity, vocal quality, variations in syllabic length and speech disfluencies is provided). In the second approach, very fine-grained descriptions and modelling have been conducted (House, 1990; Mertens, 2004), but they are too rich to be easily exportable. The question therefore remains: what is the best compromise between an overly detailed phonetic description and a phonological annotation which is too narrow from a theoretical point of view? In an attempt to answer this question, the following prerequisites underpin our approach to prosodic annotation. First, it should be based on a **theory-independent phonological labelling**. To achieve this, we have designed an inductive prosodic processing which does not impose a phonological (generative) mould, but in which various existing notation systems (such as ToBI, Intsint, IVTS, see references below) could be integrated. Second, the annotation proposed by the expert should be easily reproducible by non-expert annotators and finally carried out by computers (in order to reduce the cost of human processing and

[1] For French, see the work of Post (2000) and Jun & Fougeron (2002).

to avoid the subjectivity and variability of manual treatment).

This paper deals with an initial set of fundamental questions: (i) What does it mean to develop a theory-independent method of annotation? What does it imply in terms of methodological choices? (ii) Can we consider a type of annotation which is based on a categorical processing of prosody as well as continuous judgment, or is the latter too difficult to implement and process in a shared prosodic annotation? (iii) What kind of preliminary analysis is required in order to write a well-documented guideline sharable in the community for French prosody annotation? These three questions led us to conduct two pilot experiments in 2009, which are presented here. Each section is structured as follows: description of the corpus, the task, and the results, and a brief discussion of the experiment in question to explain the final choices made for the reference prosodic labelling summarized in the conclusion.

2 Pilot experiment one

This first experiment was conducted on a 63 sec. (335 syllables) recording, consisting in a monologue of spontaneous speech (interview with a shopkeeper in southern France). The recording was processed by 15 expert annotators (native French researchers in phonology and/or phonetics). The goal of this section is to present (§2.1) the task and its different steps, (§2.2) the results of the coding regarding inter-annotator agreement and (§2.3) the major problems revealed by the results concerning the coding method.

2.1 The task

The prosodic annotation is based first on the marking of two boundary levels, second on the identification of perceptual prominences, and finally on the labelling of disfluencies and hesitations.

Given our bias neutrality theory, no constraint was set *a priori* regarding prosodic domain and constituents separated by a prosodic break (rhythmic, syntactic or pragmatic units; this point concerns the functional interpretation to be conducted later). Concerning prominences, we considered that prominence was syllabic and had not to be merged with the notion of stress. This means that a prominent syllable is considered as a perceptual figure emerging from its background. Finally, we defined disfluency as an element which breaks the linear flow of speech, whatever the element is: it can be a syllable, a word, a morpheme unit, part of a sentence, etc.

The starting point of the procedure is a semi-automatic alignment processing (Goldman, 2008) conducted under Praat (Boersma and Weenink, 2010 which provides a 3-layer segmentation structure: segmentation within a phones string, syllabic string, and words string. They are all displayed on 3 temporally aligned tiers. Three empty tiers aligned on the syllabic tier have to be annotated (FRONT for marking the prosodic boundaries, PROM for annotating prominences and DYSF for coding disfluencies). Finally, a COMMENTS tier can be used to point out some mistakes in the annotation task and/or errors in the pre-processing (wrong segmentation or transcription, etc). An example of an annotated output file is given in figure 1.

Since the annotators do not have access to the acoustic parameters (melodic and intensity line, spectral information), the identification of prosodic boundaries, prominences and disfluencies is based only on perceptual processing. The coding methodology (categorical scale for the annotation) is structured in the following way: each annotator browses the file from left to right and organises the work in 3 steps.

- **First step: FRONT Tier processing, two degrees of prosodic boundary**

First, each annotator has to identify **breath groups** (henceforth BG, marker '2' at the end of the BG). A BG is defined as follows: it corresponds to a string of syllables bounded left and right by a silent pause, regardless of the function or duration of the pause.

Example:

#C'est clair$_2$#
(#it is obvious#)

Second, in each BG, the expert indicates where he perceives the end of an **internal prosodic group** (IPG, marker '1').

Example:
#mais$_1$ je vais aussi$_1$ leur donner de moi-même$_2$#
(#and I will also give them of myself#)

If the annotator is not sure about the presence of a prosodic boundary, he uses the indeterminacy marker '?'. In this way, two degrees of prosodic boundary are identified (major: BG and minor: IPG). Then, IPG are used to determine internal prosodic segments, which form the new anchor

points (coding span) for the following processing steps (prominences and disfluencies annotation).

- **Second step: PROM tier processing**

The marker '1' is associated to syllables perceived as prominent (± terminal: *la re$_1$lation$_1$: the relationship*), and the indeterminacy marker '?' indicates the locations where the annotator hesitates between the presence and the absence of a prominence.

Example:

La personne$_?$ va vous ra$_1$conter sa vie$_1$
(the man will tell you his life).

The accentual clash rule (Dell, 1984; Pasdeloup 1990) is not taken into account. In other words, two or more contiguous syllables can be annotated as prominent.

- **Third step: DISF tier processing**

As for the coding of prominences, the experts use the symbol '1' to indicate the disfluencies clearly identified and '?' to point out a hesitation. The latter context is often linked to lengthening and final post-tonic schwa.

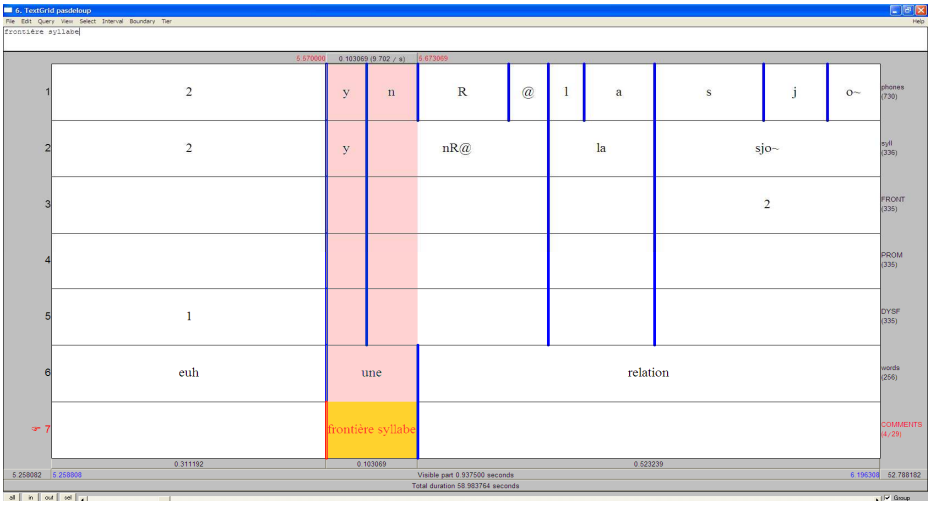
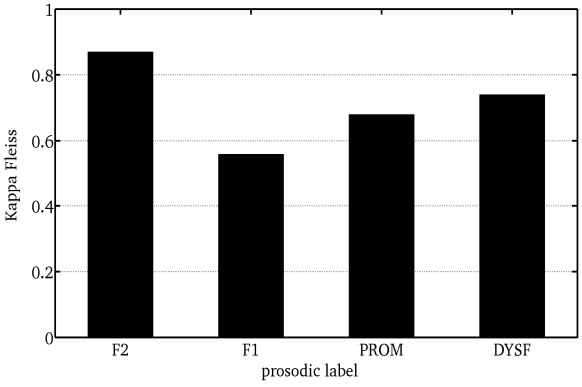


Figure 1. Example of prosodic annotation in pilot experiment one. Tiers indicate, from top to bottom: phones, syllables, boundaries (FRONT), prominences (PROM), disfluencies (DISF), graphemic words and comments. The empty segments correspond to any prosodic events detected in which the comment points out an incorrect syllabic labelling.

2.2 Results of the coding: inter-annotator agreement in pilot experiment one

- **Agreement measure**

The kappa statistic has been widely used in the past decade to assess inter-annotator agreement in prosodic labelling tasks (Syrdal and McGory, 2000), and in particular the reliability of inter-annotator agreement in the case of a categorical rating, (Carletta, 1996). Among the many versions proposed in the literature, we selected the *Fleiss' kappa* (Fleiss, 1971), which provides an overall agreement measure over a fixed number of annotators in the case of categorical rating (unlike Cohen's Kappa which only provides a measure of pairwise agreement).



- **Results**

Figure 2 presents the Fleiss' kappa agreement for each prosodic label. Indeterminacy markers were simply processed as missing values and removed from the annotation data.

Figure 2. Inter-annotator agreement for each prosodic label

These results show moderate agreement on prosodic boundaries for FRONT1 (0.56) and FRONT2 (0.86). While agreement on major prosodic boundaries seems to be strong, it should be remembered that this marker was formally imposed on the annotators in the instructions. Consequently, the score questions the relevancy of the task: if a few annotators did not follow it, it is probably because in specific distributions, the end of a BG does not correspond to a major prosodic boundary. Furthermore, experts noticed that a prosodic break could be stronger at the end of an IPG than at the end of a BG where the silent pause is not necessarily due to a prosodic break, especially in spontaneous speech. Prominence labeling provides moderate agreement (0.68), better than FRONT1, and better than the agreement scores found in the literature for other prominence labelling tasks for French speech (Morel *et al.*, 2006)[2]. Finally, disfluency labelling shows substantial agreement, disagreements being mostly due to confusion between the prominent or disfluent status of a syllable.

2.3 Conclusion on pilot experiment one

The results of this first experiment call for the following comments. While identification of hesitations and disfluencies seems to be an easy task, the annotation of prosodic boundaries and prominences raises a set of methodological and linguistic questions: (i) Are the concepts sufficiently well-defined to represent the same prosodic reality for each annotator? (ii) How far are the experts influenced by their theoretical background or phonological knowledge? (iii) To what extent does the fixed coding methodology introduce noise in the labelling (for instance, does the end of a BG systematically correspond to a major prosodic boundary)? (iv) Is a 3-step annotation coding too heavy a cognitive task, incompatible with the principle of economy required by a sharable prosodic annotation scheme?

3 Pilot experiment two

For this second experiment, we chose the same recording (speaker from southern France, 63 sec. of speech) and a second one that was more difficult because of its interactive dimension and because it contains many speech overlaps and disfluencies (3 speakers of Normandy, 60 seconds of speech, 284 syllables to label). The data were processed by 11 experts. This section follows the same organization as section 2.

3.1 The task: focus on prosodic packaging

For this second experiment, we selected to focus the annotation on the most problematic point in the first experiment, namely the coding of prosodic breaks. We conjectured that the lack of agreement derived first from the terminology that the experts were asked to use: the concept of *prosodic boundary*, which is phonologically marked and also theory-dependent, might explain the lack of consensus between experts belonging to different schools. Consequently, each annotator was asked to carry out only one task, called *prosodic packaging*. In this task, the expert had to segment the flow of speech into a string of prosodic packages (Mertens, 1993; Chafe 1998) as far as possible according to his perceptual processing, *i.e.* independently of any underlying functional and formal constraints.

Given the nature of the task, the method of annotation was not imposed, unlike the first experiment. In other words, each annotator fixed his own coding span. Finally the experts were required to carry out a meta-analysis, justifying their coding span and trying to understand and explain the cues they had used for the packaging task (acoustic, rhythmic, syntactic, pragmatic criteria).

Each Praat textgrid is composed of five tiers (see figure 3 below): three tiers are used as anchor points for the annotation (syllables, words and "Loc.", which indicates the speaker changes), and only one tier has to be annotated (prosodic packages); the Comments tier is also displayed with the same function as in experiment one. Four symbols are used for the annotation (continuous scale rating): "?": hesitancy regarding the end of a package; "1": end of a package, weak break with the following package; "2?": indeterminacy regarding the degree of the transition between two packages (weak or strong); "2": strong breaks between two packages.

[2] These better results are probably due to the more stringent method of annotation imposed.

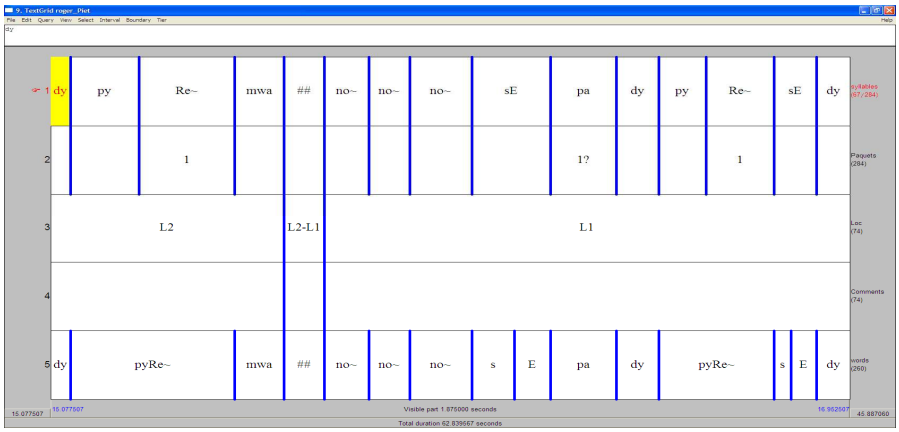


Figure 3. Example of transcription in prosodic packages in pilot experiment 2.Tiers indicate, from top to bottom: syllables, boundaries (FRONT), speakers (LOC, where L1 and L2 mean speaker one and speaker 2, L1-L2 = overlap between the 2 speakers), comments and phonetic words.

3.2 Results of the coding: inter-annotator agreement in pilot experiment two

- **Agreement measures**

In addition to the Fleiss'kappa test used in the first experiment, we introduced here the *Weighted Cohen's Kappa* (Fleiss and Cohen, 1973) which provides a pairwise agreement measure in the case of ordinal categorical rating (categorical labels are ordered along a continuous scale). In particular, weighted Cohen's Kappa weights disagreement according to the nature of the disagreed labels. Linear Cohen's Kappa was used in this experiment.

In this second experiment, we addressed three kind of inter-annotator agreement: (i) **Presence of the end of a prosodic package** (PPP), *i.e.* to what extent did annotators agree about the end of a prosodic package? (ii) **Location of the end of a prosodic package***:* annotators may agree on a PPP, but disagree on the exact location of this boundary. This was measured by adding a tolerance on the location of the PPP (1-order syllable context). (iii) **Strength of the end of PPP**, *i.e.* how much annotators agree about the degree of a prosodic boundary.

Fleiss' kappa was estimated for the first two problems, and Linear Cohen's Kappa for the last (indeterminacy markers being considered as intermediate degrees).

- **Results**

Figure 4 presents the agreement scores for the three cases mentioned above and for the two corpora used.

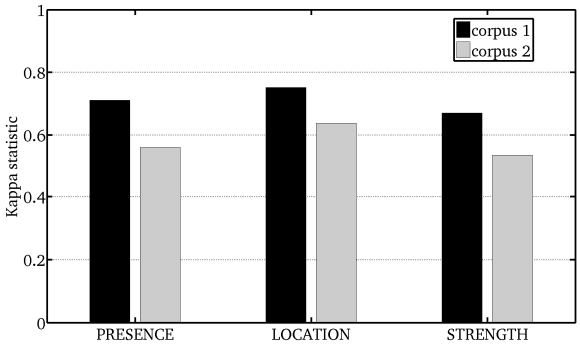


Figure 4. Inter-annotator agreement according to presence, location, and strength of the end of prosodic package.

Overall agreement scores indicate a significantly lower agreement for the second corpus. This is probably related to its higher complexity (low audio quality, high level of interaction, many disfluencies, regional accent) which made the task harder to process. The comparison of **presence** (corpus 1 = 0.71; corpus 2 = 0.56) versus **strength** (corpus 1 = 0.67; corpus 2 = 0.53) of the end of a prosodic package agreements suggests that categorical rating is more reliable than ordinal rating. In other words, annotators appear to perform better at rating the categorical status of a syllable rather than its precise degree. On the **location** problem, it is first interesting to note that the occurrence of such a location shift is significant in the prosodic labelling. In the present study, the location shift represents respectively 12% and 18% of syllables that were rated as PPP by at least one of the annotators (**balance effect**, see figure 5). Thus, merging these shifts leads to a higher agreement score (corpus 1 = 0.75 and corpus 2 = 0.63 after merging).

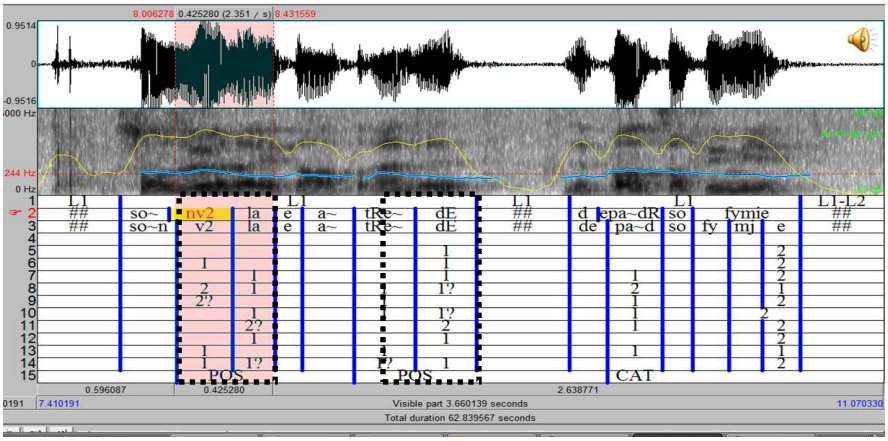


Figure 5. Examples of balance effect in the segment *"son neveu là est en train d'é-"* (*his nephew is there now*)

- **Annotator clustering**

Finally, we investigated whether the experts' phonological models affected the way in which they perceive prosodic objects.

First, annotators were labelled by the authors according to their assumed underlying phonological model. This resulted in 4 groups (3 different phonological models + a residual group: two speech engineers involved in signal processing with no phonological model).

The annotators were then hierarchically clustered according to their agreement score (see figure 6). This hierarchical clustering was achieved through complete linkage on semi-euclidean distance between annotator agreement (see Hastie *et al.*, 2009 for details)

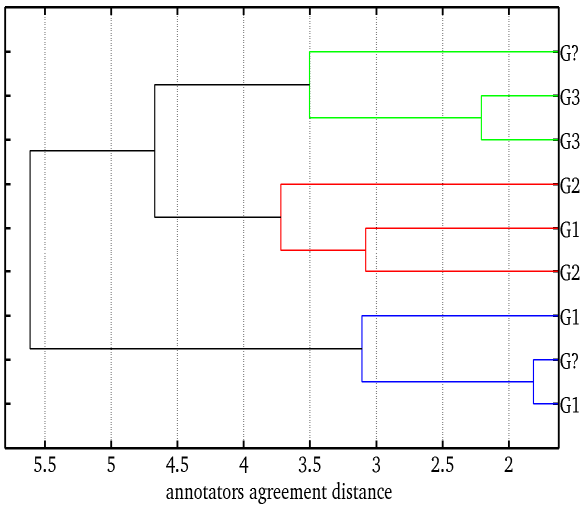


Figure 6. Agglomerative hierarchical clustering of the annotators according to their agreement on both corpora.

Interestingly, this results in three main clusters that significantly match the three previously defined groups for process annotation: (i) A tonal perception (G1) and syntactic functional approach (Mertens, 1993); (ii) Cognitive processing (G2), trying to segment the flow of speech independently of syntactic constraints (Lacheret, 2007; see the notion of flow of thought in Chafe, 1998); (iii) a formal approach (G3) based on prosodic phonology (Nespor and Vogel, 1986) and the problem of mapping between prosodic structure and generative syntax (Selkirk, 1984).

3.3 Conclusion on pilot experiment two

Two main conclusions emerge from this second experiment. (i) Even if prosodic constructions are in many respects continuous mechanisms, it seems more realistic for the time being to consider a method based on a categorical annotation. (ii) This second experiment confirms that the experts' phonological models significantly affect annotation and questions the reliability of expert annotation. However further investigation is needed and a comparison with non-expert annotators must be conducted before drawing any definitive conclusions.

4 Conclusion

Given the results of pilot experiments 1 and 2, we conclude that neither the static concept of **prosodic boundary**, nor its dynamic substitute **prosodic packaging** leads to a high inter-annotator consensus. In other words, these two concepts are probably too dependent on different levels of processing (syntactic, phonological, and rhythmic) and each annotator, depending on his own definition of the notion (formal or functional) will focus on one aspect or another. Con-

sequently, even if precise instructions are given for annotation, the labelled data still remain heterogeneous. Therefore, these two concepts should not be used as the basis for the development of a shared prosodic annotation method aiming to establish a reference prosodic corpus and annotation software, which are essential tools in handling large volumes of speech data. In contrast, we hypothesize that prominence annotation based on perceptual criteria represents the cornerstone of speech prosodic segmentation, as prosodic structure will be generated from prominence labelling. Although the results of the first pilot experiment are rather poor 0.68), recent experiments have shown that the scores rise (0.86) after training sessions (Avanzi *et al* 2010b). We have therefore decided to focus our annotation guideline on the labelling of prominences (two levels of prominence: strong or weak) and disfluencies (hesitations, false starts, speaker overlaps, post-tonic schwas, etc.). The method does not depend on some abstract property of words or groups of words, as in the case of lexical stress (Martin, 2006; Poiré, 2006; Post *et al*. 2006), but is based on a neutral phonetic definition of prominence, associated with perceptual salience in the context of the speech background. This approach has the advantage of being consensual, whatever the theoretical framework adopted. Based on these criteria, a one day training session has been organized for 5 novice annotators (students in linguistics) in order to annotate 3.30 hours of different speech genres (private, public, professional), over 2 months (from February to April 2010). For each genre a monologal and an interactional sample of around 5 minutes (42 speech files altogether) have to be labelled. Prominences and disfluencies are coded on two independent tiers.

The annotation deliverable will be processed during the spring by five experts who will have to perform four tasks: (i) compute the inter-annotator scores applying the statistical measures used in the two pilot experiments; (ii) diagnose the distributions with the poorest scores for all the samples; (iii) diagnose the genres with the worst scores and (iv) make explicit decisions to provide an output prosodic reference annotation and to enhance automatic prominence detection software (see for French: Avanzi *et al.*, 2010a; Martin 2010; Obin *et al.* 2008a, 2008b, 2009; Simon *et al*. 2008).

Acknowledgements

This research is supported by the Agence Nationale de la Recherche (ANR-07-CORP-030-01, “Rhapsodie – Corpus prosodique de référence du français parlé”). We would like to thanks our colleagues from LPL (Aix-en-Provence), ERSS (Toulouse), Ircam (Paris), MODYCO (Paris) and also University of Geneve (Switzerland), Louvain-la-Neuve, Leuven (Belgium) to have conduct this work for *Rhapsodie*.

References

Mathieu Avanzi, Anne Lacheret and Anne-Catherine Simon. 2010. *Proceedings of Prosodic Prominence, Speech Prosody 2010 Satellite Workshop, Chicago, May 10th*.

Mathieu Avanzi, Anne Lacheret and Bernard Victorri. 2010a. A Corpus-Based Learning Method for Prominences Detection in Spontaneous Speech. *Proceedings of Prosodic Prominence, Speech Prosody 2010 Satellite Workshop, Chicago, May 10th*.

Mathieu Avanzi, Anne-Catherine Simon, Jean-Philippe Goldman and Antoine Auchlin, 2010b. C-PROM. An annotated corpus for French prominence studies. *Proceedings of Prosodic Prominence, Speech Prosody 2010 Satellite Workshop, Chicago, May 10th*.

Frédéric Beaugendre, Christophe d’Alessandro, Anne Lacheret-Dujour and Jacques Terken. 1992. A Perceptual Study of French Intonation. *Proceedings of the International Conference on Spoken Language Processing*, J. Ohala (ed.), Canada, 739-742.

Paul Boersma and David Weenink. 2010. Praat: doing phonetics by computer (version 5.1), www.praat.org.

Jeska Buhmann, Johanneke Caspers, Vincent J. van Heuven, Heleen Hoekstra, Jean-Pierre Mertens and Marc Swerts. 2002. Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the Spoken Dutch Corpus. *Proceedings of LREC2002*, Las Palmas, 779-785.

Jean Carletta, 1996. Assessing agreement on classification tasks: the Kappa statistic. *Computational Linguistics*, 22(2):249-254.

Wallace Chafe. 1998. Language and the Flow of Thought. *New Psychology of language*, M. Tomasello (ed.), New Jersey, Lawrence Erlbraum Publishers, 93-111.

Emmanuela Cresti and Massimo Moneglia, eds. 2005. *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*, Studies in Corpus Linguistics 15. Amsterdam, Benjamins.

Elisabeth Delais-Roussarie. 2005. *Phonologie et grammaire, étude et modélisation des interfaces prosodiques*. Mémoire d'habilitation à diriger des recherches, Toulouse.

François Dell. 1984. L'accentuation dans les phrases françaises. *Forme sonore du langage, structure des représentations en phonologie*, F. Dell *et al*. Paris, Hermann, 65-122.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Joseph L. Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33:613-619.

Jean-Philippe Goldman. 2008. EasyAlign: a semi-automatic phonetic alignment tool under Praat, http://latlcui.unige.ch/phonetique.

Trevor Hastie, Robert Tibshirani and Jerome Friedman, 2009. Hierarchical clustering. *The Elements of Statistical Learning* (2nd ed.). New York: Springer, 520-528.

Daniel Hirst and Albert Di Cristo. 1998. *Intonation Systems: A Survey of Twenty Languages*, Cambridge, Cambridge University Press.

David House.1990. *Tonal perception in Speech*, Lund University Press.

Sun Ah Jun and Cécile Fougeron. 2002. The Realizations of the Accentual Phrase for French Intonation", *Probus*, 14:147-172.

Anne Lacheret. 2007. Prosodie du discours, une interface à multiples facettes. *Nouveaux Cahiers de linguistique française*, 28:7-40.

Philippe Martin. 2003. ToBI : l'illusion scientifique? *Actes du colloque international Journées Prosodie 2001. Université de Grenoble*, 109-113.

Philippe Martin. 2006. La transcription des proéminences accentuelles : mission impossible?, *Bulletin de phonologie du français contemporain*, 6:81-88.

Philippe Martin. 2010. Prominence Detection without Syllabic Segmentation, *Proceedings of Prosodic Prominence, Speech Prosody 2010 Satellite Workshop, Chicago, May 10th*.

Piet Mertens. 1993. Intonational Grouping, boundaries, and syntactic structure in French. *Working Papers Lund University*, 41:156-159.

Piet Mertens. 2004. The Prosogram: Semi-Automatic Transcription of prosody based on a Tonal Perception Model. *Proceedings of Speech Prosody 2004, Nara, Japan*. 549-552.

Michel Morel, Anne Lacheret-Dujour, Chantal Lyche, Morel M. and François Poiré. 2006. "Vous avez dit proéminence?, *Actes des 26èmes journées d'étude sur la parole, Dinard, France*. 183-186.

Marina Nespor and Irene Vogel. 1986. *Prosodic Phonology*, Foris, Dordrecht

Nicolas Obin, Xavier Rodet and Anne Lacheret-Dujour. 2008a. French prominence: a probabilistic framework. *Proceedings of ICASSP'08*, Las Vegas, U.S.A.

Nicolas Obin, Jean-Philippe Goldman, Mathieu Avanzi and Anne Lacheret. 2008b. Comparaison de trois outils de détection automatique de proéminences en français parlé. *Actes des 27èmes journées d'étude sur la parole*, Avignon, France.

Nicolas Obin, Xavier Rodet and Anne Lacheret-Dujour. 2009. A Syllable-Based Prominence Detection Model Based on Discriminant Analysis and Context-Dependency, *Proceedings of SPECOM'09*, St-Petersburg, Russia.

Valérie Pasdeloup. 1990. *Modèle de règles rythmiques du français appliqué à la synthèse de la parole*, PhD, Université de Provence.

François Poiré. 2006. La perception des proéminences et le codage prosodique, *Bulletin de phonologie du français contemporain*, 6:69-79.

Brechtje Post. 2000. *Tonal and phrasal structures in French intonation*. The Hague, Thesus.

Brechtje Post, Elisabeth Delais-Roussarie and Anne Catherine Simon. 2006. IVTS, un système de transcription pour la variation prosodique, *Bulletin de phonologie du français contemporain*, 6:51-68.

Elisabeth Selkirk. 1984. *Phonology and Syntax: the Relation between Sounds and Structure*. Cambridge, Cambridge MIT Press.

Kim Silverman, Mary Beckman, John Pitrelli, Mary Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert and Julia Hirschberg. 1992. ToBI: A standard for Labeling English prosody, *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*. 867-870.

Anne Catherine Simon, Mathieu Avanzi, Jean-Philippe Goldman, 2008. La détection des proéminences syllabiques. Un aller-retour entre l'annotation manuelle et le traitement automatique. *Actes du 1er Congrès Mondial de Linguistique Française, Paris*.1673-1686.

Caroline L. Smith. 2009. Naïve listeners' perceptions of French prosody compared to the predictions of theoretical models. *Proceedings of the third symposium Prosody/discourse interfaces, Paris, September 2009*.

Ann K. Syrdal and Julia McGory. 2000. Inter-transcribers Reliability of ToBI Prosodic Labelling. *Proceedings of the International Conference on*

Spoken Language Processing, Beijing, China. Vol. 3, 235-238.

Fabrizio Tamburini and Carlo Caini. 2005. An automatic System for Detecting Prosodic Prominence in American English Continuous Speech. *International Journal of Speech technology*, 8:33-44.

Colin W. Wightman. 2002. ToBI or not ToBI?. *Proceedings of Speech Prosody*, Aix-en-Provence, France, 25-29.