Prominence perception and accent detection in French: from phonetic processing to grammatical analysis

Anne Lacheret-Dujour, Anne-Catherine Simon, Jean-Philippe Goldman, Mathieu Avanzi

Prominence perception and accent detection in French: From phonetic processing to grammatical analysis

Anne Lacheret[1], Anne Catherine Simon[2], Jean-Philippe Goldman[3], Mathieu Avanzi[1, 4]

[1]*Université Paris Ouest Nanterre, France,* [2]*Université catholique de Louvain, Belgium,* [3]*Université de Genève, Switzerland,* [4]*Université de Neuchâtel, Switzerland.*

Anne.Lacheret@u-paris10.fr

Anne-Catherine.Simon@uclouvain.be

Jean-Philippe.Goldman@unige.ch

Mathieu.Avanzi@unine.ch

Key words: auditory perception, automatic labelling, grammatical categories, prominence, prosodic annotation

In the area of large speech corpora, there is a demonstrated need for a common prosodic notation system that would allow easy data exchange, comparison of annotations and analyses, and automatic processing. A single, simple scheme of prosodic transcription could also form the basis of guidelines for non-expert manual annotation, and be used for linguistic teaching and research (Lacheret et al. 2010). Setting up such a system, however, raises several major issues. The first problem that needs to be addressed is which annotation format to select, and which annotation tool. A second concern, discussed by Cresti and Moneglia (2005), is how to establish reference prosodic corpora for different discourse genres. And even assuming that reference corpora can be compiled on these lines, the question still remains whether they will prove adequate to develop corpus-based learning methods for automatic prosodic labelling in spontaneous speech (Buhmann et al. 2002; Tamburini and Caini, 2005; Avanzi et al. 2010b).

At present, many fundamental questions remain unsolved, especially in contemporary French in which the prosodic system has a number of typologically peculiar features, and the answers to these questions are far from obvious. Scholars generally highlight the syncretism between accentuation and intonation in French, or, to put it more accurately, the influence of intonation over accentuation (Rossi, 1979). It is generally argued that accent in French has first of all a demarcative function and pilots the segmentation of the flow of speech into syntactico-semantic units and that accentual prominences lead to the identification of the final boundaries of prosodic groups. In fact, data analysis reveals that there are strong variations in accent distribution: first, accent can also mark the left boundary of a group; second, many internal prominences are observed, linked to rhythmic and/or pragmatic constraints. In view of these findings, hypothetico-deductive theories such as autosegmental and metrical phonology must be questioned (see Avanzi, 2011 for a review). These theories are based on the hypothesis that French prosodic structure is organized around deep categories (accents and boundaries), deep units (accentual phrases and intonational phrases) and various functions (syntactic, rhythmic and informational) marked by unambiguous acoustic cues in the flow of speech. However, many surface prosodic forms in spoken French cannot be explained by this one-to-one processing, thereby calling into question the underlying premises of this theoretical framework.

We therefore propose a usage-based, data-driven processing where syllabic prominence is the first, and the most essential, element in initiating prosodic analysis. A prominent syllable appears perceptually as a figure emerging from its background, relative to its immediate

context. The goal of this paper is to present the different steps of our work which was conducted in order to establish a reference prosodic corpus based on prominence labelling; this corpus is then available to analyse the grammatical constraints which govern prosodic patterns. The analysis will focus on terminal prominences.

1. Prominence annotation in a French speech database: from manual annotation to automatic labelling

In this section, we present 3 pilot experiments which began in 2003 as part of the PFC project (Durand et al. 2002; Eychenne and Mallet 2004; Detey et al. 2010) on prominence annotation. Our goal is twofold: (i) to promote a bottom-up and inductive approach based on a perceptual identification of prominence; and (ii) to discuss the different issues linked to the development of prosodic annotation and automatic tools for prosodic labelling. We will show that while prominences generally have acoustic correlates, this is not always the case: there are situations where no prosodic marker is present. In order to explain these perceptual illusions, we hypothesize that listeners cannot totally ignore the symbolic level which determines in part their prosodic expectations. In other words, grammatical constraints affect the perception of prominences.

1. 1 Pilot experiment 1: prosodic annotation in the PFC database, scope and issues

The earliest empirical approaches to French prominence started in 2003 within the PFC project (Lacheret-Dujour et al. 2004). The first prosodic annotation guideline was defined in order to study correlations between segmental and suprasegmental constructions but also to allow studies in a wide range of prosodic domains (accentuation, intonation, rhythm, role of prosody in the marking of syntactic structure and information flow).

In order to be sharable, the coding procedure had to meet three criteria: (i) it had to be independent of any theoretical framework; (ii) it should rely on perceptual judgments; (iii) it should be reproducible by non-experts (students in linguistics). In other words, the challenge was to propose a set of minimal annotation rules which would allow a network of researchers to produce new prosodic annotated data based on the same format. The coding procedure was based on a bottom-up, data-driven approach, and perceptual processing: the annotator had to identify remarkable syllabic perceptual events in the flow of speech regardless of their acoustic marking and functional constraints (rhythmic, syntactic, or pragmatic). The coding covered 6 lexical fields and was organised as follows (see Lacheret-Dujour et al. 2005 for illustrations): fields 1 and 2 were linked to the domain of coding, while the remaining fields (3 through 6) provided prosodic information. In detail, field 1 indicated the number of syllables of the word which was processed, field 2 provided information on the location, within the lexical word, of the syllable under consideration, field 3 indicated the perception of prominence on the syllable, field 4 gave information about syllabic length, field 5 coded pauses around the processed syllable, and field 6 gave information about the syllabic distribution in the message (initial in a new turn-taking, either initial or internal in a prosodic group).

The data for this study consisted of a sub-corpus of the PFC corpus. Four PFC investigation points were selected: Treize-Vents (a small village in Vendée, in western France), Paris (upper-class speakers), Nyon (a village in French-speaking Switzerland), and Douzens (a small village in southern France). From the 10 speakers recorded per investigation point, we selected 5 and partially coded their productions for prosody. Since for each speaker of each investigation point, the PFC recordings included two reading tasks (a word-list and a short passage) and two conversations (semi-directed and informal), the coding was performed on relevant excerpts of the read text and on one to two minutes of each conversation.

On completion of this first study in 2004, the promoters of the project organized a workshop at Caen University in order to undertake an initial review. The main issues raised during the review were the following: (i) the contradiction between the lexical strategy of annotation and the post-lexical characteristics of French prosody (mainly the word grouping function); (ii) the fact that syllabic processing was used instead of holistic processing, although we know that prominent events are global and continuous rather than local and discrete (Astésano et al. 2004); (iii) lastly, what can be said about the acoustic and linguistic 'anchor points' underlying the annotation? In other words which task is actually carried out by the annotators: is it only a perceptual one or, what is more likely, a task at the interface between perceptual cues and mental linguistic representations about the prosodic system, in which grammatical constraints on prosodic outputs are specified?

A second experiment was therefore conducted by Poiré (2006) in order to better understand the phenomenon actually involved in annotation and the pitfalls to avoid.

1.2. Pilot experiment 2: but what is prominence?

In this experiment, seven phonetics experts were asked to annotate perceived syllabic prominences in a 3-minute recording of spontaneous speech by a Belgian male speaker, without any other instructions. It was expected that a fairly encouraging degree of agreement would be reached, since prominence has to correspond with accent, and the accentuation rules of French were well-known by the experts. Analysis of the results revealed 3 points: (i) the perception of non-terminal prominences or prominences carried by clitics is weak. This can be explained by the French accentuation system, in that these locations do not correspond to the accent distribution in French (terminal syllable of lexical words, *i.e.* primary stress); (ii) among the 165 syllables which could receive a primary stress, the proportion marked as prominent varied from 19% to 49%, in other words, the inter-rater agreement was poorer than expected. Morel et al. (2006) used the results of this pilot study to conduct experiments on the data. The aim was to evaluate the robustness of two acoustic parameters (f0 and duration) for automatic prominence detection and to suggest measures to evaluate the annotators' performances. The authors concluded that melody was better correlated with inter-rater agreement (the higher the f0 values, the better the inter-rater agreement was), whereas a similar correlation with duration values was not observed. Thus, over a determined duration threshold (between 175 and 200 ms), the proportion is inverted and agreement does not increase, but decreases. This is due to the fact that beyond a certain threshold, lengthening is no longer perceived as a prominence clue, but as a mark of hesitation.

From these initial experiments, we learnt several lessons. The first was that the low rate of agreement came from the lack of accuracy in the coding instructions. To obtain a better inter-annotator agreement, the notion of prominence needed to be carefully defined, and not conflated with the notion of "stress" (which is a phonological notion implying linguistic knowledge). Second, it was necessary to define a context-window for prominence identification, to avoid ending up with large parts of the sound signal without any prominence detection. Furthermore, the above-mentioned authors agreed that visualization of the signal was helpful. Lastly, the study of the acoustic correlates of perceived prominences showed that while f0 was a good cue for automatic identification, so was duration, provided that hesitation marks had a specific annotation, to avoid biasing the relative duration calculations. A third pilot experiment was therefore conducted, involving the construction of a multi-genre and multi-speaker corpus, called C-PROM, annotated for French prominence[1].

[1] See http://sites.google.com/site/corpusprom/.

1.3. Pilot experiment 3: the C-PROM database

The C-PROM corpus was designed with two purposes in mind: (i) to develop an annotation tool in the Praat software (Boersma and Weenink, 2010); and (ii) to build an open data-base to train algorithms for semi-automatic prominence detection in French.

The corpus comprised 70 minutes of speech (28 speakers: 12 females, 16 males), sampled from 7 speaking styles ranging from high to low formality (Read Speech (RS), Political Discourse (PD), Conferences (CF), News Broadcasts (NB), Radio Interviews (RI), Map Tasks (MT) and Narratives (NA); about 3 minutes per sample). Each sample was automatically segmented into phones, syllables and orthographic words using the Easyalign script (Goldman, 2008) and manually checked.

Two expert phoneticians annotated the whole corpus in the following way: each annotator started from an empty annotation tier duplicated from the syllabic tier, and filled each interval with one of the symbols described in Table 1. For each sample, the annotation was conducted by listening no more than three times to stretches of speech of 3 to 5 seconds (over-listening resulting in overdetection). At the end of the labelling, a COMPARE-tier was used to estimate the inter-transcriber agreement (Figure 1).

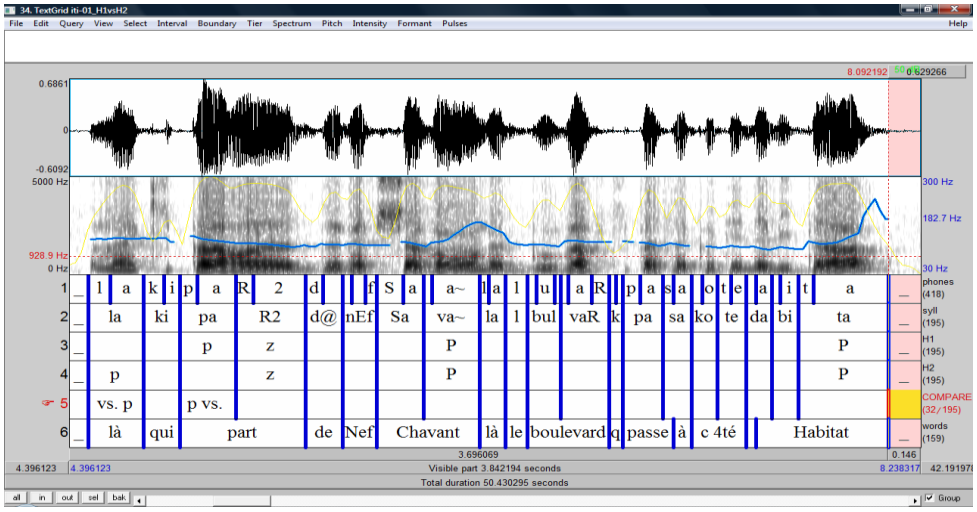


Figure 1. Praat screenshot of the utterance: *là qui part de nef Chavant là le boulevard qui passé à côté d'Habitat* [mp-1]. Annotation tiers are, from top to bottom: phones, syllables (both in SAMPA), delivery (H1 and H2 manual annotations), COMPARE-tier and graphemic words.

The annotation of prominences relies on auditory perception of salience and not on the visual analysis of acoustic parameters (f0 movements, for example). **The first class of symbols** is for annotating prominent syllables: two markers of prominences (weak (p) and strong (P)) can be used. **The delivery labelling** is necessary for singling out syllables which have specific properties likely to hamper automatic prominence identification. The "z" symbol is for extra-lengthened syllables marking hesitation since their length can disturb the calculation of relative duration, as shown by Lacheret-Dujour and Beaugendre (1999). The marking of hesitation also serves to avoid false automatic detection of prominence, since hesitations are often followed by a silent pause, and silent pause is often considered a strong clue for prominence detection; it could therefore introduce false alarms in the automatic detection system. Post-tonic schwa (@) and appendices ($) are considered non prominent, but they are specifically annotated because they introduce irregularity in the final-accent system in French (Dell, 1980; Mertens, 2006). Otherwise, the number of symbols in this "delivery" class can also be explained by the perspective of a semi-automatic identification of these specific prosodic phenomena. **The last category of symbols** contains silent pauses (resulting from the semi-automatic alignment), audible breaths and "junk", *i.e.* parts of the recording

that could not be transcribed (noise, laughter, coughing, overlapping, etc.). These could interfere with the automatic processing of the signal.

1. Prominence labeling	
P	strongly prominent syllable
p	weakly prominent syllable
	non prominent syllable
2. Delivery labeling	
z	lengthening connected with a hesitation
@	post-tonic syllabic schwa (as in "c'est dinguE" [sEde~g@])
$	unaccented post-tonic syllables (appendice)
3. Others	
%	junk (noise, laugh, cough, etc.)
*	Breath
‾	silence

Table 1. Annotation symbols. Among the 17.778 syllables of the corpus, 805 (4.5%) were annotated with a delivery symbol, 4,570 as prominent (25.7%) and 12,403 (69.7%) as non prominent.

This third pilot experiment was a turning point in our work plan and led to two major directions. First, it was clear that other experiments involving more speaking styles, more transcribers, both experts and non experts, needed to be conducted on larger corpora. Indeed, better results in the estimation of the inter-annotator agreement constitute strong evidence against the idea that French prominence transcription is more an art than a scientific practice (Martin, 2006). The second positive outcome of this experiment is that sub-parts of the corpus have already been used to train different automatic prominence detection algorithms (Avanzi et al. 2007; Goldman et al. 2007; Avanzi et al. 2008; Obin et al. 2008a). It also resulted in studies on the automatic classification of speaking styles according to the prosodic features (Obin et al. 2008b; Simon et al. 2008). Finally, it was used to study in detail the question initiated in pilot experiment 2 (see section 1.2) regarding the interplay between grammatical constraints and phonetic cues involved in the perception of accentuation in French (Goldman et al. 2010).

1.4. New experiments

Nevertheless some fundamental issues remain about the C-PROM methodology. In this experiment, only two experts performed the annotation. The distinction between “p” and “P” was used to help the transcribers to develop a more accurate listening strategy and avoid annotating only the strongest prominences. During comparison of the manual annotations, however, these two categories were merged. In fact, it is not certain whether the p/P distinction is only a heuristic tool, or whether it refers to a relevant distinction between weak and strong prominences. Regarding terminal stress, at the phonological level, this opposition might be a cue to distinguish accentual from intonational units. At the phonetic level, the question remains open whether we can detect different degrees of activation of the acoustic parameters which contribute to those two kinds of prominences. The merging of the P and p categories precludes both studies. Finally, on the pragmatic level, this opposition may be a clue towards a better understanding of information packaging in speech (Lacheret et al. 2011).

These issues led to two further experiments. First, in the Rhapsodie program[2], 3 hours of speech in different situations (private, professional and public, 58 samples from 1 to 15

[2] See http://rhapsodie.risc.cnrs.fr/fr/index.html.

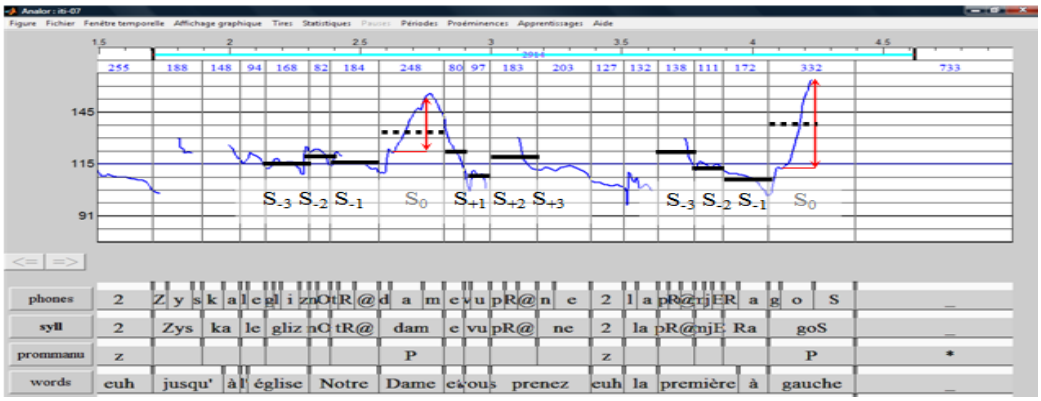
minutes) were labelled. The annotation, conducted by 4 novice annotators and controlled by 3 experts, led to the detection of two kinds of prominences (p vs. P) and disfluencies on two separate tiers. This experiment provided a new annotated database which could be used to compare the performance of the two types of annotation (discrete where the p/P markers were merged vs. semi-continuous where the p/P opposition was kept). Secondly, the continuous prominence detection system ANALOR was developed. This means that we now have all the material needed to understand which annotation strategy allows the best comprehension of phenomena involved in prominence perception (acoustic cues vs. perceptual illusions derived from syntactic and pragmatic expectations). More specifically, we can evaluate in what way a continuous prominence model may or may not help to understand the interaction between the different levels in the perceptual processing of prosodic structure, the goal being to find the best compromise between performance and computational complexity.

2. A corpus-based learning method for continuous prominence detection in continuous speech implemented in the Analor software

This section presents the algorithm implemented within the ANALOR software[3]. First, from a phonemic alignment, our system conducts a detection of prominent syllables (§2.1). Then, on the basis of this detection, it estimates the degree of prominence (§2.2).

2.1. Prosodic Prominence location

The prominent syllables detection procedure is implemented under Matlab in an interface called ANALOR. The detection relies on the calculation of four prosodic parameters: (i) syllabic duration; (ii) syllabic height average (following House, 1990, only the f0 points of the vowels are taken into account); (iii) amplitude of the rising tone on vocalic nucleus; (iv) presence of a silent pause not connected with a hesitation or a false start. As shown in Figure 2, the algorithm calculates, for the current syllable (S_0), (i) its relative height and duration average compared with the f0 and averages of the three preceding syllables (S_{-3}; S_{-2} and S_{-1}) and the three following ones S_{+1}; S_{+2} and S_{+3}); (ii) the presence of a rise if there is a positive movement of f0 on the syllabic nucleus, and (iii) the presence of an adjacent silent pause. f0 measures are given in semi-tones, while duration measures are calculated without any unit. Note that contextual relativization is blocked if there is a syllable marked as excluded in the labeling tier (based on the pre-manual annotation of the corpus) or a silent pause in the immediate context of the current syllable. In Figure 2, the last syllable of the utterance is followed by a pause. Duration and f0 measures are thus calculated only with reference to the three preceding syllabic intervals.



[3] Since the procedure is described in detail in Avanzi et al. (2008, 2010b, 2011a-b), we just give in this section a brief summary of the main steps.

Figure 2. ANALOR screenshot of the utterance: *euh jusqu'à l'église Notre Dame vous prenez la première à gauche* [MT]. On the abscissa, temporal values are given in milliseconds; on the ordinate, the values of f0 in a logarithmic scale can be seen. Duration labels are given in milliseconds. Annotation tiers are, from top to bottom: phones, syllables (both in SAMPA), manual annotation ("prommanu", indicating prominence syllables (P), excluded syllables (z), silent pause (_) and breath (*)) and graphemic words.

A syllable is considered as prominent if (i) one of the first three parameters reaches a certain threshold and/or (ii) a silent pause (whatever its duration) follows the current syllable (the annotation of silent pauses was made during the semi-automatic phone alignment step; "false" pauses such as pre-occlusive silences were automatically excluded). The method we decided to follow in order to determine the best parameters for automatic prominence identification consisted in the development of supervised corpus-based learning. The aim was to hone the F-measure performance by systematically comparing the results with the manual annotations. We used the C-PROM corpus to train the algorithm and compare its performance with the manual annotation (§ 1.3).

Automatic learning is based on a random local search, in decreasing steps, in the parameter space from a relevant value. More precisely, if we denote V a vector of the space (a 3-D space, the vector V taking S_D, S_H and S_R as components), the algorithm can be described as follows:

Call δ_i the browse step, V_i the value of the parameters, and F_i the F-measure at step i of the procedure. We make a random search to find a new value of V which improves the F-measure by looking in the neighborhood of V_i defined by step δ_i. That is to say we try the V values of the form:

$$V = V_i + \delta_i \cdot \Phi \cdot V_i$$

where Φ is a regular distributed random vector in the hypercube unit.

As long as we do not find a better value for V, we continue by replacing V_i by this value. If we do N_{max} searches without finding a better value, we move on to step i+1 of the procedure with a step $\delta_{i+1} = \delta_i /2$. The procedure stops when the step becomes smaller than the given ordinate value δ_{min}. The results shown below were obtained with $N_{max} = 250$, $\delta_1 = 0.4$ and $\delta_{min} = 0.01$.

The description of the corpus-based learning method shows that this algorithm is efficient if and only if the initial values of the parameters are sufficiently close to the optimal value. In other words, the initial values were fixed on the basis of a linguistic analysis, and drew on specific linguistic knowledge. For this study, we considered the following initial value: $S_D = 2$; $S_H = 2$ and $S_R = 3$. Justification for the value of these thresholds which were fixed *a priori* can be found in Rossi (1972) and D'Alessandro and Mertens (1995).

On this basis, we trained the initially fixed intuitive thresholds, for each discourse genre[4]:

Corpus	S_D	S_H	S_R
NB	1.49	2.68	1.71
RS	1.61	1.43	2.07
PD	1.55	2.46	2.1
CF	1.48	2.29	2.73
RI	1.76	2.67	3.67

[4] News Broadcasts (NB), Read Speech (RS), Political Discourse (PD), Conferences (CF), Radio Interviews (RI), Map Tasks (MT) and Narratives (NA).

MT	1.71	2.6	2.47
NA	1.54	1.38	2.48

Table 2. Values of the optimal thresholds obtained for relative duration (S_D), relative height (S_H) and intra-vocalic amplitude rise (S_R) for each discourse genre of the C-PROM. S_H and S_R values are given in semi-tones, while S_D values are calculated without any unit.

The measure chosen for estimating the agreement between manual annotation and automatic identification is the F-measure, *i.e.* the harmonic average between precision and recall (van Rijsbergen, 1979). Table 3 gives the performance of our tool for each discourse genre.

Genre	**initial performance**			**trained performance**		
	Prec.	Rec.	F-ms	Prec.	Rec.	F-ms
RS	79.86	71.7	75.56	76.41	77.87	77.13
PD	75.07	83.39	79.01	82.35	81.86	82.16
NB	74.57	73.58	74.07	75.7	82.3	78.86
CF	76.11	73.23	74.64	79.18	79.95	79.56
RI	71.88	82.6	76.87	79.3	80.89	80
MT	75.31	76.51	75.9	79.86	79	79.43
NA	83.27	61.75	70.91	73.44	80.73	76.91
TOTAL	**76.58**	**74.68**	**75.28**	**78.03**	**80.37**	**79.15**

Table 3. % of F-measure for each discourse genre, before and after training. The average for all the discourse genres is given in the grey columns.

As can be seen, the corpus-based learning enabled the results to be honed by about 3.85% of F-measure: the performance before training is 75.3%, compared to 79.15% after. The best progression is observed for the CF discourse-genre (5.14%) and the worst for RS (1.64%). Concerning the rate of agreement between manual annotation and automatic detection, we can see that the best score is for PD, while the worst is for NA recordings. Overall, the performance reached by our tool (79.15%) is fairly close to the inter-annotator agreement found by Avanzi et al. (2010c) (estimated at 82.8% of F-measure), which is quite encouraging.

2.2. Prominence Degree Categorization

In order to estimate the degree of prominence of the syllables detected as prominent, we adopted the following hypothesis: the greater the number of acoustic parameters involved in the identification of prominence, the more the fixed thresholds are exceeded, and the more the prominence is perceived as strong.

First, for each of the first three criteria used to detect the location of stressed syllables in the preceding step (relative duration, relative height and rising tone), we attributed a score between 0 and 10. This score was determined according to the difference with the optimal threshold fixed during the corpus-based learning procedure. A value equal to the threshold gives a score of 5; a value of 0 (i.e. 100% lower than the threshold) gives a score close to 0, and a value of twice the threshold (i.e. 100% above the threshold) gives a score close to 10. The exact formula used is:

$$f(x) = 10 \cdot \left(\tfrac{1}{2} + \tfrac{1}{2} \cdot \tanh(2 \cdot \lambda \cdot \frac{x-t}{t}) \right)$$

where x is the value of the current syllable for the given criterion, t the threshold, and λ the slope of the function (changing this makes the slope more or less steep; by default its value is 1.5).

Concerning the silent pause criterion, the score is 0 or 10 since it is a binary criterion.

Finally, the strength of the prominent syllable is obtained by computing the weighted average of the four scores:

$$strength = \frac{f_D(x_D).wght_D + f_H(x_H).wght_H + f_R(x_R).wght_R + f_P(x_P).wght_P}{wght_D + wght_H + wght_R + wght_P}$$

where D is the duration value, H the height value, R the rise value and P the silent pause value. The weight (wght) for the three continuous criteria is 1, while that for silent pause is 0.5.

2.3. Illustration

The result of the automatic identification of the location of prominences and of their strength is visualized on Figure 3. Syllables detected as prominent are marked “p” or “P” in a dedicated tier (named “Pauto”), and the score of prominence (rounded to the nearest unit) in the tier just below (named “Strength”).

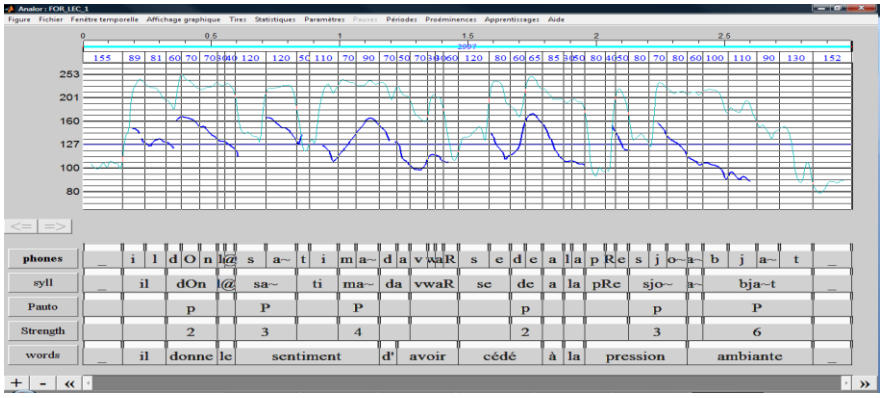


Figure 3. ANALOR screenshot, illustrating the automatic identification of prominence. Analysis of the utterance: *il donne le sentiment d’avoir cédé à la pression ambiante* [RP]. In the top part, the evolution of F0 can be measured in hertz (values are on the left) or in semi-tones (the interval between two horizontal lines is one semi-tone). In the bottom part, the transcription tiers are, from top to bottom: phones, syllables (both in SAMPA), prominent syllables, strength of the prominence and words.

The details of the calculations are given in dedicated windows that the user can call up by clicking on the syllable. Figure 4 below gives the detail for the last syllables of the words *sentiment* and *ambiante.*

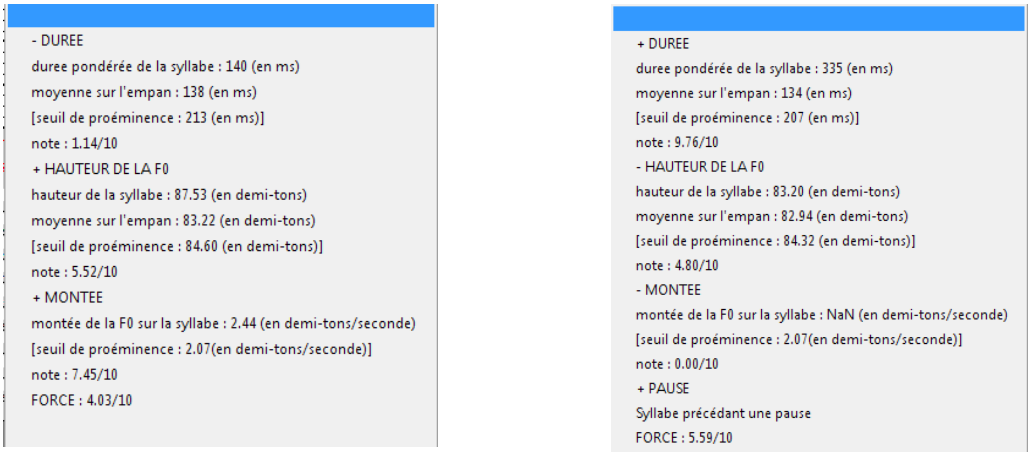


Figure 4. Values and scores for the syllables *sentiment* (on the left) and *ambiante* (on the right) of the sentence analyzed in figure 3.

3. Prominence, Accentuation and Grammatical Categories

The research reported in section 2 does not make any explicit link between syllabic prominence and stress or accentuation. The very notion of prominence is indeed twofold: the phonetic side considers the acoustic correlates of prominence, and their relative contribution to prominence perception (see Terken and Hermes, 2000 for a review). We will focus on the phonological side, which envisages the functions of prominent syllables in relation to the accentual system of the language under consideration. The remainder of this paper aims at understanding better why human perception of prominence can diverge from automatic detection and particularly the impact of grammatical categories (basically the distinction between content words and function words) on prominence perception. In other words, we try to measure whether the listener's expectation for a word to be accented has any effect on his/her perception of that word as prominent.

3.1. Part-of-speech Annotation

The data-set used for automatic prominence detection and prominence degree categorization was described in section 2.1. The same corpus was annotated for part-of-speech, using an automatic procedure and manual validation (see Goldman et al. 2010). Table 4 gives an overview of the grammatical categories and sub-categories, and the number of tokens per category in the corpus.

Macro-categories	**Categories**	**Subcategories**
CONTENT WORDS	NOUN (2126)	nouns and proper names
	ADJ (556)	adjectives
	VERB (588)	verbs (finite verbs, participles, infinitives)
	ADV (693)	adverbs of manner, degree, negation, comparison and interrogation
FUNCTION WORDS	CONJ (417)	coordination and subordination conjunctions
	PRON(806)	pronouns (includ. 12 different classes)
	AUX (305)	verbal auxiliaries (144) and predicative use of "être" (741)
	PREP (939)	prepositions
	DET (1287)	determiners (definite, indef., interrogative , multiple words, prepositional)

Table 4. Grammatical categories for POS tagging in C-PROM

In order to focus on primary stress in French (see Lacheret-Dujour and Beaugendre, 1999, p. 49), we restricted ourselves to observing the last syllable of the words (or the single syllable in case of monosyllabic words).

3.2. Overall Distribution of Syllables' Prominence Degree

According to the reference experts' annotation reported by Avanzi et al. (2010c), the perceptually prominent syllables amount to 33.1%. However, prominent syllables are far from all having the same degree of acoustic prominence. Figure 5 shows the overall distribution of all the syllables in the corpus, according to their degree of prominence; syllables perceived as prominent or non prominent are grouped in this chart.

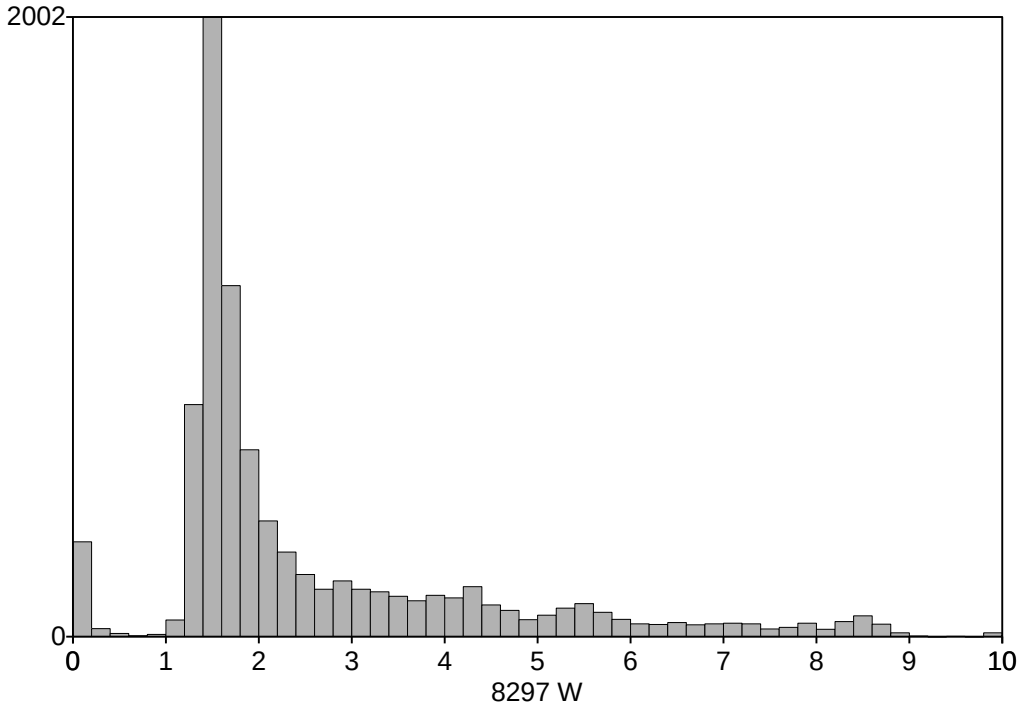


Figure 5. Overall distribution by degree of acoustic prominence (ranging from 1 to 10), for the final syllables of the words.

The median value of the distribution is 1.7 degrees of acoustic prominence[5] on the degree of prominence scale presented in section 2.2. Observing the shape of this distribution, we may assume that non-prominent syllables are fairly alike and grouped at the 1.7 peak and that the prominent syllables are dispatched over a larger scale of acoustic degree.

3.3. Distribution of Syllables' Prominence Degree by Grammatical Category

The analysis of the distribution of prominence degree for each grammatical category led us to group our 9 grammatical categories into 2 macro-categories, content words and function words (see Figure 6), which have the following properties:

- Content words (Nouns, Adjectives, Verbs and Adverbs) may be described as non clitic frequently accented words. Acoustically, their prominence degree has a median value of 2.5 (mean=3.26 dev=2.09) and shows a wider spread than function words. They are perceived (by experts) as bearing a final primary accent in 54.1% of the cases. Amongst content words, it is interesting to note that Nouns and Adjectives have an even higher proportion of syllables with a high degree of acoustic prominence (median = 3.1).
- Function words (Conjunctions, Prepositions, Auxiliaries and Determiners) are clearly clitic in the sense that they are hardly ever perceived as prominent (7.7%). Acoustically, most of the tokens have a very low degree of acoustic prominence (median=1.5; mean=1.67 dev=0.93). Amongst them, Conjunctions behave in a slightly different way since they are perceptually detected as prominent in 14.9% of the cases and have a lesser proportion of syllables ranging from degree 0 to 2.

[5] Mean = 2.54 and standard deviation = 1.85.

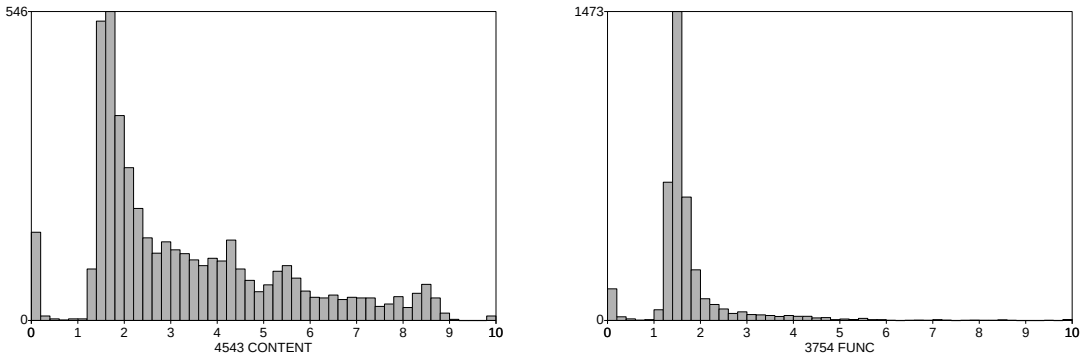


Figure 6. Distribution of Prominence Degree for Content Words (left) and Function Words (right)

The analysis of both distributions confirms that, from an acoustic point of view, prominence is a matter of degree[6] even if the perception may be quite clear-cut and even categorical. Phonologically, prominence is intrinsically related to stress and accentuation, and may be derived from the morphological and syntactic structure of a constituent.

3.4. Distribution of Syllables' Prominence Degree compared to Auditory Perception

We now compare the distribution of prominence degree of non-prominent vs. prominent perceived syllables. 2749 out of 8297 syllables were *perceived* as prominent by the two annotators. This amounts to 33.1%. The dashed line in Figure 7 corresponds to the percentile 66.9 of the distribution, *i.e.* the degree of acoustic prominence under which there are 66.9% of syllables and above which there are 33.1% of syllables. Theoretically, we might say that the $Q_{66.9}$=2.4 degree of prominence separates non prominent from prominent syllables.

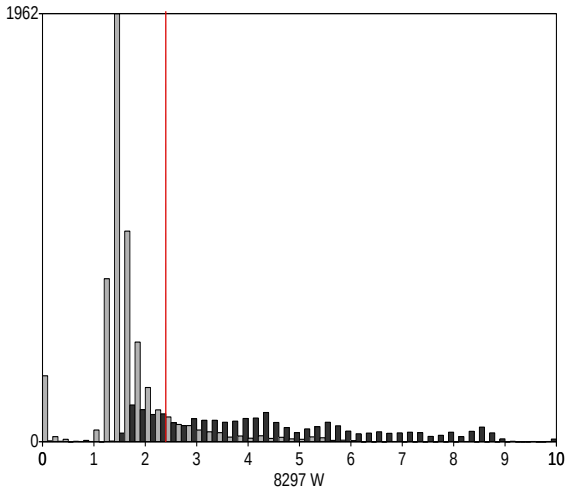


Figure 7. Overall distribution of Syllables' Acoustic Prominence Degree with syllables perceived as non-prominent (in grey) and syllables perceived as prominent (in black). The dashed line indicates degree 2.4 that theoretically separates the 66.9% of non prominent syllables from the 33.1% of syllables perceived as prominent.

The cumulative distribution interestingly shows that, for prominence degree from 2 to 3, a syllable can be perceived as either prominent or not prominent by the listener. On the other hand, syllables having an acoustic prominence degree above 4 are likely to be perceived as prominent.

This means that black bars on the left correspond to acoustically weak but perceptually prominent syllables ("missed" prominence), while grey bars on the right correspond to

[6] The acoustic prominence degree gradually increases from determiners (mean = 1.4) to nouns (mean = 3.1).

acoustically strong but perceptually non prominent syllables ("perceptual illusion") (see Table 5).

Acoustic Degree	Non-prominent perceived syllable (66.9%)	Prominent perceived syllable (33.1%)
<2.4	Correct rejection 4926 (59.4%)	Perceptual illusion (false alarm) 622 (7.5%)
≥2.4	Miss 620 (7.5%)	Hit 2129 (25.6%)

Table 5. Matching of perception of prominence and automatic detection of prominence

3.5. Distribution of Syllables' Prominence Degree as compared to Auditory Perception, by Grammatical Macro-Categories

Does grammar have an impact on the perception of prominence? We make the following hypotheses:

- Perceived prominence with a low degree of acoustic prominence (the so-called perceptual illusion) will be more frequent for Content words (Nouns, Adjectives, Verbs and Adverbs).
- Function words perceived as prominent will have a higher mean degree of acoustic prominence. The listener does not expect them to be prominent; consequently, they have to stand out from their context with more force to be perceived as prominent.

The first hypothesis is verified in Table 6, which clearly demonstrates that perceptual illusions (non-prominent syllables perceived as prominent by expert listeners) predominantly affect Content words (10.2 %) and seldom Function words (3.1%). This comes from the fact that content words are expected to be accented; a smaller acoustic salience may trigger the perception of a final accent.

Grammatical Category	N	Experts' Detection	Hit	Correct rejection	Miss	Percept. Illusion
CONTENT	4543	54.1%	1995 (43.9%)	1607 (35.4%)	476 (10.5%)	**465 (10.2%)**
FUNC	3754	7.7%	173 (4.6%)	3349 (89.2%)	116 (3.1%)	**116 (3.1%)**

Table 6. Numbers and percentages of hits, correct rejections, misses and perceptual illusions (see Table 5 for a definition) for content words vs. function words.

The second hypothesis is also verified since the theoretical degree threshold of function words perceived as prominent ($Q_{92.3}$=2.75)[7] is higher than for content words ($Q_{45.9}$=2.3). This degree is indicated by the dashed line in Figure 8.

[7] Percentile 92.3 separates syllables perceived as non -prominent (92.3%) from those perceived as prominent (7.7%) for function words. Percentile 45.9 separates perceptually non prominent (45.9%) from prominent (54.1%) syllables for content words.

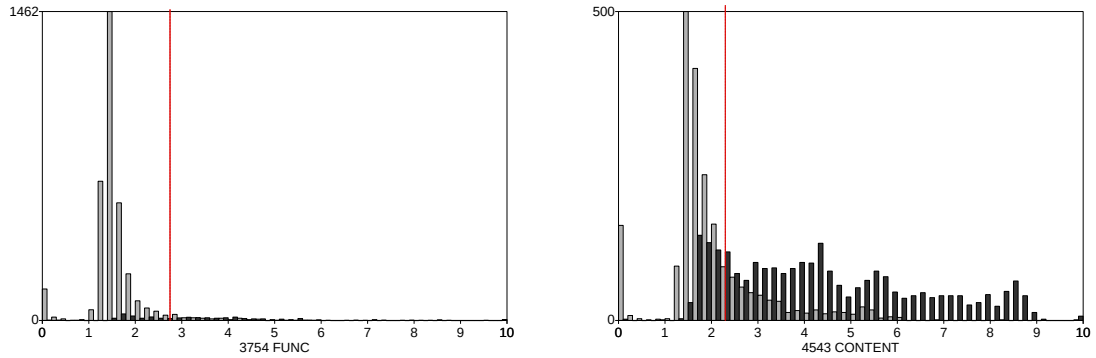


Figure 8. Cumulative distribution (perceptually prominent (in black) and non prominent (in grey) syllables) by macro-category (Noun-Adj and Verb-Adv on the left; Function words on the right)

In sum, we have shown that, acoustically speaking, the clitic vs. non-clitic distinction is a matter of degree, since we observe overlapping distributions indicating that clitic words have a globally lower degree of acoustic prominence. Perceptually, it appears that clitic words are perceived as prominent in 7.7% of the cases (vs. 54.1% for content words). In addition, we have shown that the expectation for a word to be accented will foster the perception of prominence on this word: perceptual illusion (*i.e.* perception of a prominence where the acoustics show a low degree of prominence) is far more frequent for content words than for function words. However, we do not claim that a ten-degree prominence distinction (see section 2.2) could be perceptually relevant, i.e. that it could be transcribed at all reliably. Nevertheless, any prominence detection or annotation procedure should take into account the part played by grammar in perception.

Conclusion

The goal of the first part of this paper was to: (i) promote a data-driven study in order to investigate the question of relationships between syllabic prominences, their acoustic cues and underlying syntactic constraints, and (ii) present the different steps of our research program devoted to the prosodic annotation of speech corpora in French and its linguistic analysis, which began in 2004 as part of the PFC project. We pointed out how the interplay between manual and automatic data processing is necessary to provide valuable insight into the bias inevitably associated with manual annotation.

In the second section, we showed, with the presentation of the ANALOR system based on a continuous prominence detection in continuous speech, how questions linked to the annotation of syllabic prominences, far from being restricted to practical goals, also serve as a lever to stimulate theoretical proposals that involve new modes of data representation and processing (in this case, a continuous processing of prominences).

In the last section, we showed how such a system can be profitably used to study the interplay between grammatical constraints and phonetic cues involved in the perception of prosodic structure in speech processing (see also for English the opposition formulated by Cole et al. 2010 between *signal-driven prominence perception* and *expectation-driven hypotheses*). Although the domain investigated here is restricted to prosody, it raises crucial issues of more general scope in corpus linguistics: do we have a robust definition of the object to be annotated, *i.e.* an annotation model on which one can build guidelines and comprehensive tutorials for annotation? How should an annotation campaign (minimal number of annotators, level of expertise of the annotators (expert vs. novice), training before the annotation process, computing of the inter-annotator agreement, etc.) be conducted? These questions and certainly many others highlight the fact that it is necessary to dissociate

reference annotation (a stable and consensual model) and experimental annotation used for research.

References

Astésano, C, Magne, C., Morel, M., Coquillon, A., Espesser, R., Besson, M, Lacheret, A., 2004. Marquage acoustique du focus contrastif non codé syntaxiquement en français. Actes des XXVèmes Journées d'étude sur la parole, Fès, Maroc.

Avanzi, M., 2011. L'interface prosodie/syntaxe en français. Dislocations, incises et asyndètes. Peter Lang, Bruxelles.

Avanzi, M., Goldman, J.Ph. Lacheret-Dujour, A., Simon, A.-C., Auchlin, A., 2007. Méthodologie et algorithmes pour la détection automatique des syllabes proéminentes dans les corpus de français parlé . Cahiers of French Language Studies 13 (2), 2–30.

Avanzi, M., Lacheret-Dujour, A., Obin, N., Victorri, B., 2011a. Vers une modélisation continue de la structure prosodique : le cas des proéminences syllabiques. Journal of French Language Studies 21 (1), 53-71.

Avanzi, M., Lacheret-Dujour, A., Victorri, B., 2008. ANALOR. A Tool for Semi-Automatic Annotation of French Prosodic Structure. Proceedings of Speech Prosody 2008, pp. 119-122.

Avanzi, M., Lacheret-Dujour, A., Simon, A.C., 2010a. Proceedings of Prosodic Prominence. Speech Prosody 2010 Satellite Workshop, Chicago, May 10th, http://speechprosody2010.illinois.edu/ .

Avanzi, M., Lacheret-Dujour, A., Victorri, B., 2010b. A corpus-based learning method for prominence detection in spontaneous speech. Prosodic Prominence: Perceptual and Automatic Identification. Proceedings of Speech Prosody 2010 Workshop, Chicago, http://speechprosody2010.illinois.edu/.

Avanzi, M., Obin, N., Lacheret-Dujour, A., Victorri, B., 2011b. Toward a Continuous Modeling of French Prosodic Structure: Using Acoustic Features to Predict Prominence Location and Prominence Degree. Proceedings of Interspeech, Firenze, Italy, pp. 28-31.

Avanzi, M., Simon, A. C., Goldman, J.Ph., Auchlin, A. 2010c., C-PROM. An annotated corpus for French prominence studies. Proceedings of Speech Prosody 2010 Workshop, Chicago, http://speechprosody2010.illinois.edu/.

Boersma, P. & Weenink, D., 2010. Praat: doing phonetics by computer (Version 5.3). www.praat.org.

Buhmann, J., Caspers, J., van Heuven, V.J., Hoekstra, H., Mertens, J.P., Swerts, M., 2002. Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the Spoken Dutch Corpus. Proceedings of LREC2002, Las Palmas, pp, 779-785.

Cole, J., Mo, Y., Hasegawa-Johnson, M., 2010. Signal-based and expectation-based factors in the perception of prosodic prominence. Laboratory Phonology 1, 425–452.

Cresti, E., Moneglia, M., (eds.) 2005. C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages. Studies in Corpus Linguistics 15. Benjamins, Amsterdam.

D'Alessandro, C., Mertens, P., 1995. Automatic pitch contour stylization using a model of tonal perception. Computer Speech and Language 9/3, 257-288.

Dell, F., 1980. Generative Phonology and French Phonology. Cambridge University Press, Cambridge.

Detey, S., Durand, J., Laks, B., Lyche, Ch., (eds.) 2010. Les variétés du français parlé dans l'espace francophone. Ressources pour l'enseignement. Paris, Ophrys, Paris.

Durand, J., Laks, B., Lyche, Ch., (eds.) 2002. Protocole, Conventions et directions d'analyse. Bulletin PFC n°1, http://www.projet-pfc.net/bulletins-et-colloques.html.

Eychenne, J., Mallet, G., (eds.) 2004. Du segmental au prosodique : protocoles, outils, extensions et travaux en cours, http://www.projet-pfc.net/bulletins-et-colloques.html.

Fonagy, I. Léon, P., (eds.) 1979. L'accent en français contemporain. Studia Phonetica 15, Didier, Paris.

Goldman, J.Ph.; Avanzi, M.; Lacheret-Dujour, A.; Simon, A.-C.; Auchlin, A, 2007. A Methodology for the Automatic Detection of Perceived Prominent Syllables in Spoken French. Proceedings of Interspeech'07, Antwerp, Belgium, August 27-31.

Goldman, J.Ph., 2008. EasyAlign: a semi-automatic phonetic alignment tool under Praat, http://latlntic.unige.ch/phonetique.

Goldman, J.Ph., Auchlin, A., Roekhaut, S., Simon, A.C., Avanzi M., 2010. Prominence perception and accent detection in French. A corpus-based account. Proceedings of Speech Prosody 2010, Chicago, Illinois, http://speechprosody2010.illinois.edu/.

House, D., 1990. Tonal Perception in speech. Lund University Press, Lund.

Lacheret-Dujour, A., Beaugendre, F., 1999. La prosodie du français. CNRS, Paris.

Lacheret-Dujour, A., Lyche, Ch., Morel, M., 2004. Pour une transcription prosodique normalisée au sein du projet PFC (phonologie du français contemporain): champ d'action et limites. Actes des 25èmes journées d'étude sur la parole, Fès, Maroc.

Lacheret-Dujour, A., Lyche, Ch., Morel, M., 2005. Phonological Analysis of Schwa and liaison within the PFC Project (Phonologie du français contemporain) : how Determinant are the Prosodic Factors*?*. Eurospeech 2005, Lisbonne.

Lacheret, A., Kahane, S., Pietrandrea, P, Avanzi, M., Victorri, B., 2011. Oui mais elle est où la coupure là ? Quand syntaxe et prosodie s'entraident ou se complètent. Langue française 170, 61-80.

Lacheret, A., Obin, N., Avanzi, M., 2010. Design and evaluation of shared prosodic annotation for spontanesous French speech: from expert knowledge to non-expert annotation. 48th Annual Meeting of the Association for Computational Linguistics, UPPSALA, pp. 265-273.

Martin, Ph., 2006. La transcription des proéminences accentuelles : mission impossible ? Bulletin PFC 6, 81-87.

Mertens, P., 2006. A Predictive Approach to the Analysis of Intonation in Discourse in French. In: Kawaguchi, Y. et al. (Eds), Prosody and Syntax. Benjamins, Amsterdam, pp.64-101.

Morel, M., Lacheret-Dujour, A., Lyche, Ch., 2006. Vous avez dit proéminence ? Actes des 26èmes journées d'étude sur la parole, Dinard.

Obin, N., Goldman, J.Ph., Avanzi, M., Lacheret-Dujour, A., 2008a. Comparaison de trois outils de détection semi-automatique des proéminences dans les corpus de français parlé. Actes des 22èmes Journées d'étude sur la parole, Avignon, pp. 333-336.

Obin, N., Lacheret-Dujour, A., Veaux, Ch., Rodet, X, Simon, A.C., 2008b. A Method for Automatic and Dynamic Estimation of Discourse Genre Typology with Prosodic Features. Proceedings of Interspeech, Brisbane, pp. 1204-1207.

Poiré, P., 2006. La perception des proéminences et le codage prosodique. Bulletin PFC 6, 69-79.

Rossi, M., 1972. Le seuil différentiel de durée. In: Valdman, A. (Ed.), Papers in Linguistics and Phonetics to the Memory of Pierre Delattre (54), A. Collection Janua Linguarium, Mouton, The Hague, Indiana University.

Rossi, M., 1979. Le français, langue sans accent ? In Fonagy, I., Léon, P. (Eds.), L'accent en français contemporain. Studia Phonetica 15, Didier, Paris, pp. 13-51.

Simon, A.C., Auchlin, A., Avanzi, M., Goldman, J.-Ph., 2008. Les phonostyles. Une description prosodique des styles de parole en français. Actes du colloque Les voix du français : usages et représentations, Oxford.

Tamburini, F., Caini C., 2005. An automatic System for Detecting Prosodic Prominence in American English Continuous Speech. International Journal of Speech technology (8), 33-44.

Terken, J., Hermes, A., 2000. The perception of prosodic prominence. In: M. Horne (Ed.), Prosody, Theory and Experiment. Studies Presented to Gösta Bruce. Kluwer Academic Publisher, The Netherlands, pp.89-217.

van Rijsbergen, C.J. 1979. Information Retrieval. Butterworths, London.