Word boundaries in French: Evidence from large speech corpora

Rena Nemoto[*⋆], Martine Adda-Decker[*], Jacques Durand[◇]

[*]LIMSI-CNRS (UPR351) BP 133 91403 Orsay Cedex France
{nemoto,madda}@limsi.fr
[⋆]Université Paris-Sud 11, Orsay France
[◇]CLLE-ERSS (UMR5263) CNRS & Université de Toulouse-Le Mirai
5 Allées Antonio-Machado 31058 Toulouse France
jacques.durand@univ-tlse2.fr

Abstract

The goal of this paper is to investigate French word segmentation strategies using phonemic and lexical transcriptions as well as prosodic and part-of-speech annotations. Average fundamental frequency (f_0) profiles and phoneme duration profiles are measured using 13 hours of broadcast news speech to study prosodic regularities of French words. Some influential factors are taken into consideration for f_0 and duration measurements: word syllable length, word-final schwa, part-of-speech. Results from average f_0 profiles confirm word final syllable accentuation and from average duration profiles, we can observe long word final syllable length. Both are common tendencies in French. From noun phrase studies, results of average f_0 profiles illustrate higher noun first syllable after `determiner`. Inter-vocalic duration profile results show long inter-vocalic duration between `determiner` vowel and preceding word vowel. These results reveal measurable cues contributing to word boundary location. Further studies will include more detailed within syllable f_0 patterns, other speaking styles and languages.

1. Introduction

A large body of works in human speech processing addresses the question of whether and how word boundaries may be inferred from the acoustic signal by human listeners. A review of the literature on human word segmentation reveals two main tendencies: (i) the word segmentation problem can be – at least partly – solved by distributional properties of the language (Harris, 1955; McQueen, 1998; Saffran et al., 1996). (ii) the word segmentation problem takes benefit from acoustic cues among which most importantly prosodic information (Cutler and Norris, 1988; Mattys et al., 1999; Bagou et al., 2002). For the time being, automatic speech recognition (ASR) systems hypothesize word boundaries in continuous speech using word and word co-occurrence information, rather than specific acoustic cues. ASR systems can then be viewed as supporters of the first tendency, relying on distributional cues. The distributional information comes from the lexical level here, rather than from the prelexical level in psycholinguistic studies, as ASR systems receive a priori knowledge of a language's lexicon. However, the word segmentation problem remains very tricky due to the combinatorial complexity. Here are two homophone phrase examples illustrating the word segmentation problem in French:

1) /lezar/: *les arts* ; *lézard* (the arts ; lizard)
2) /õsãdegut/: *on s'en dégoûte* ; *on sent des gouttes* (we are disgusted ; we feel drops)

Our belief is that both distributional and prosodic cues are at work to reduce the human word segmentation problem (without neglecting the pragmatics of the situation, which may strongly influence the considered choices).

To highlight potential prosodic cues related to the word segmentation problem, this contribution presents a study of fundamental frequency (f_0) contours of French mono- and polysyllabic words using large speech corpora and automatic processing. The questions addressed are the following: can specific f_0 profiles for French words be measured automatically using large corpora? If so, how do they vary with respect to influential factors, such as word syllable length, the presence of final schwas, syllable duration or part-of-speech (POS) categories? The aim of this study is then to produce empirical evidence from large corpora concerning the raised questions, in order to contribute to our knowledge of prosodic realisations in French words and their potential to contribute to the word segmentation problem. Taking a more long-term perspective, this work aims at improving the acoustic modeling capacities in automatic speech recognition of spontaneous speech.
The speech corpus and the methodology are presented in Section 2. Section 3. deals with f_0 profiles of lexical words (with/without final-schwa) in comparison with noun phrases. A similar study with duration is provided in Section 4. Conclusions are presented in Section 5.

2. Corpus and Methodology

2.1. Corpus

We make use of the manually transcribed French TECHNOLANGUE-ESTER corpus (Galliano et al., 2005), consisting in recordings of broadcast news shows from different Francophone (French and Moroccan) radio stations. We used 13 hours of male speaker audio including 165k word tokens and 14k word types. Most of these broadcast news audio corpora are of prepared speech type. Table 1 shows the corpus composition according to mono-/polysyllabic words.

n	*Syll.class* n_s	*#Words*	*Examples*
0	0_0	12578	l'; d'; de
1	1_0	72249	vingt; reste
2	2_0	36027	beaucoup; journal
3	3_0	15994	notamment; militaire
4	4_0	6053	présidentielle
n	*Syll.class*	*#Words+ /ə/*	*Examples*
0	0_1	12295	de; le; que
1	1_1	3918	reste; test
2	2_1	2087	ministre
3	3_1	698	véritable
4	4_1	174	nationalistes

Table 1: Quantitative description of the corpus according to word tokens of syllable length n (n = 0-4). Separate counts are given for words w/wo realized final schwa(top/bottom). *Syll.class n_s* states *n*: the number of full syllables; *s*: presence(0)/absence(1) of final schwa.

2.2. Methodology

How are word boundaries signaled in fluent speech? There are no obligatory cues to signal word boundaries in fluent speech (Cutler et al., 1997). However many studies in psycholinguistics show that language-specific prosodic cues may guide segmentation strategies to postulate word boundaries (Cutler and Norris, 1988; Bagou et al., 2002). Concerning the prosodic level in French, many authors noticed the correlation between accentuation (final and initial), lengthening and word or syntagm boundaries (Vaissière, 1991; Hirst and Cristo, 1998; Lacheret-Dujour and Beaugendre, 1999; Fougeron and Jun, 1998; Gendrot and Adda-Decker, 2006). Whereas authors like Welby (2003; 2007) measure within syllable f_0 variation, in this contribution we will only focus on average cross-syllable variation. In the following, we briefly describe the adopted knowledge representation and the related processing steps on the investigated data (cf. Figure 1).

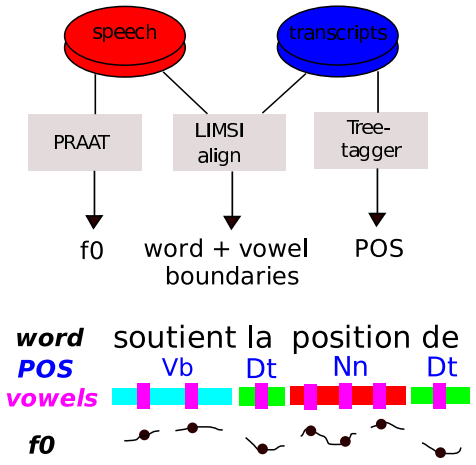

Figure 1: Processing steps to produce f_0, word and vowel boundaries as well as POS tag annotations.

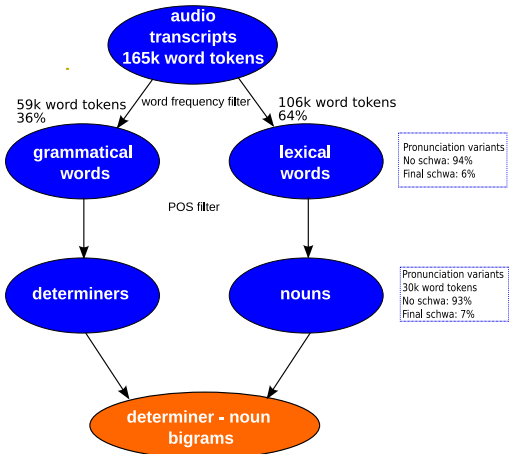

Figure 2: Illustration of the annotated corpus processing and extraction of noun phrases.

2.3. Processing steps

f_0 measurements: Fundamental frequency (f_0) values were measured every 5 milliseconds (ms) using the standard settings of Praat (Boersma and Weenink, 2005) which results in at least six f_0 samples for each segments (a minimum phoneme duration is 30 ms).

Lexical and phonemic alignment: The audio corpus was automatically aligned by the LIMSI speech recognition system (Gauvain et al., 2005) producing word and phoneme segmentations. During the alignment, the pronunciation dictionary allows for optional word-final schwas, if the standard pronunciation ends with a consonant (e.g. word test with standard pronunciation /tɛst/ and variant [tɛstə]). For technical reasons, a phone segment is a minimum 30 ms duration and a boundary location precision of 10 ms.

Part-Of-Speech tagging: The transcribed corpus was POS-tagged by WMATCH, LIMSI word regular expression engine (Galibert, 2009), using TREETAGGER (Schmid, 1994) system, to measure the influence of different POS classes and noun phrases on f_0 realizations.

2.4. Knowledge representation

Word syllable length; word syllable rank: After speech alignment, a syllable length was associated to each uttered word, which corresponds to its pronunciation vowel count (excluding final schwa). In this way, the word syllable length of population (/pɔpylasjõ/) is 4, as there are 4 full vowels /ɔ/, /y/, /a/ and /õ/. Each vowel of the corpus was annotated by its word syllable rank (e.g. in the former example vowel /y/ has rank 2 of 4). Table 1 shows the corpus composition according to the word syllable length n. Realized final schwas were not used to measure word syllable length, but the corresponding words were registered in separate prosodic classes (see *syll.class* in Table 1). Words with 0 word syllable length according to the adopted representation (0_0), are small function words with elided mute e (schwa), either on the graphemic level (l' pronounced as /l/) or at the aligned pronunciation level (le pronounced as

[l]). Monosyllabic words were the most frequent and word frequency then decreases with syllabic word length.

f_0 profiles; f_0 value by vocalic segment: f_0 profiles were computed for each word class (*syll.class* tags of Table 1). To compute these profiles only vowels with voicing ratio over 70% were used, which resulted in a rejection rate of about 10% to minimize potential segmentation errors due to automatic alignment. For each vowel a mean f_0 value was computed over all voiced frames of the vocalic segment. The values in Hz were converted to semitones (st), with 120 Hz as reference frequency (120 Hz is often considered as average male voice height) ('t Hart, 1981). Perceptual studies show that differences of 3 st play a role in the communicative situations even though weaker differences can contribute to the perception of lexical demarcation. The prepared corpus including orthographic/phonemic transcribed pronunciation was associated to each word with correspondent duration as well as its part-of-speech. Each vowel was thus annotated with its mean f_0 in st, its duration and its word syllable rank. For example, given the `2_0` class of bisyllabic words without final schwa, the corresponding f_0 profile is computed as the average f_0 of the vowels of rank 1 followed by the average f_0 of the vowels of rank 2. This is further developped in the next section.

3. f_0 profiles

To examine potential prosodic cues on word boundaries, mean f_0 profiles were computed for single words, different word classes and noun phrases. As described in Table 1 distinct classes were considered word syllable lengths (distinguishing with/without final schwa). Within each class, the f_0 vowel measurements of the same rank were averaged. In the following we present f_0 profiles for lexical words (excluding function words), nouns and for noun phrases (`Determiner - Noun`). Profiles were computed for function words: they resulted in relatively flat profiles with low average f_0 values.

3.1. Lexical words

Firstly, we present lexical word f_0 profiles, to check whether the known f_0 rise on word final syllables in French can be verified within an adopted representation schema. Grammatical words are not included for the contours in Figure 3. We limited our analyses to syllabic length $n <= 4$ (within more than 6k tokens for $n = 4$). Concerning the words with final schwa, we stopped at $n = 3$ (700 tokens).

Figure 3 shows mean f_0 profile of word classes according to n-syllabic length, without final schwa (top) and with final schwa (bottom). We can observe the following :

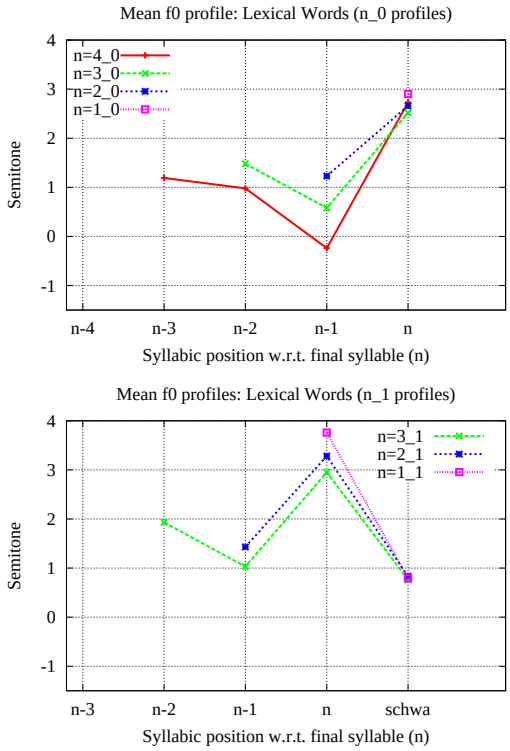

Figure 3: Mean f_0 profiles of n-syllabic lexical words. **Top:** Words without final schwa (1-4 syll.) **Bottom:** Words with final schwa (1-3 syll.).

(i) Mean f_0 is much higher for the final syllable n than for the preceding syllables.
(ii) For the trisyllables or more, the f_0 difference between final and penultimate consecutive vowels is maximal. This difference tends to increase with word syllabic length.
(iii) Mean monosyllabic f_0 is as high as that of the final syllable of longer syllabic word.
(iv) A final schwa (n_1 profiles) globally yields higher mean f_0 than the n_0 profiles, in particular for the final syllable n.
(v) The difference between the final syllable n and the following final schwa correspond to 2-3 st.
(vi) Initial accentuation remains weak on mean f_0 contours.

A more detailed examination of our sample profiles revealed that large proportions of the word class populations correspond to the average profiles. These average profiles do not arise from a small number of samples with extreme values, neither from a heterogeneous and aleatory population, but most of word proportions follow the tendencies.

3.2. Noun phrase

In this subsection, mean f_0 profiles were measured for noun phrases, limited to the `determiner noun` bigram (cf. Figure 2). Is the mean f_0 profile of a n length noun phrase different from the one of a n length noun? Figure 4 (top) shows the mean f_0 profiles of `Noun` words (30 853 occ.), very similar to Figure 3 (top). The bottom figure exhibits the mean f_0 profiles of noun phrases (12 888 occ.). From

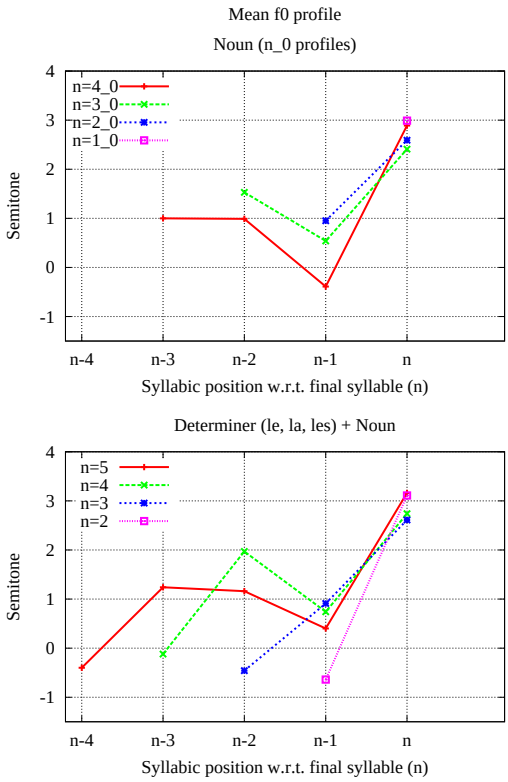

Figure 4: Mean f_0 profiles for n-syllabic length. **Top:** Nouns (without final schwa) **Bottom:** Noun phrase `determiner noun` (without final schwa) with n variant from 2 to 5.

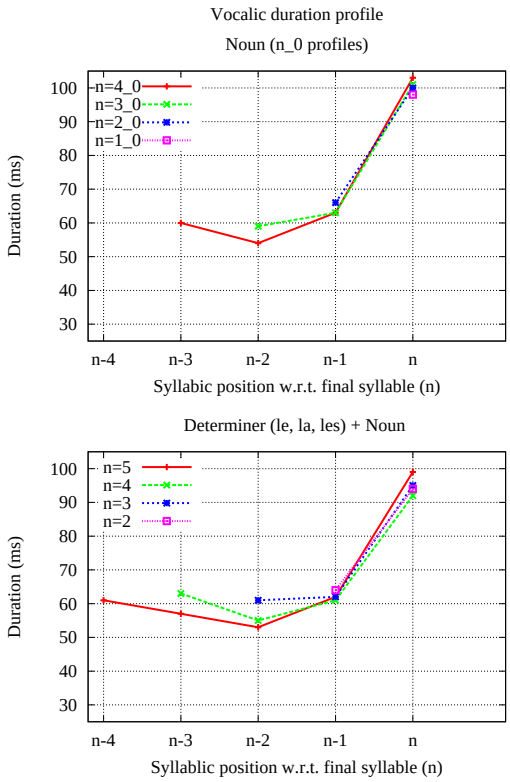

Figure 5: Mean vocalic duration profiles for n-syllabic length. **Top:** Nouns (without final schwa) **Bottom:** Noun phrase `determiner noun` (without final schwa) with n variant from 2 to 5.

these two figures, we can observe that the height difference of f_0 values is maximal between the first syllable of the noun phrase (here monosyllabic determiner) and the last syllable (last syllable of the noun). The f_0 profile differences are about 3 st. These results suggest that the f_0 (within a temporal window of some syllables) allows locating syntagm boundaries, at least for the noun phrase case (`determiner noun`).

4. Duration profiles

In the previous sections, 3.1. and 3.2., revealed the mean f_0 profiles of noun phrases can differentiate from the same n-syllabic nouns. But these f_0 contours were not considered duration which is an important part of prosodic components. In this section, duration profiles were computed in a way similar to the f_0 profiles. Firstly we measured mean vocalic duration of each vowel rank of noun words (Figure 5 top). And then we also measured mean vocalic duration of each vocalic rank of a noun phrase (Figure 5 bottom). These two figures present longer final vowel duration that are characteristic of French. But other vowel duration do not show a remarkable difference between them. Even we do not observe a difference vowel duration between `determiner` and `noun` (excluding final vowels). Each vowel duration did not show the distinctive difference between word boundaries. We hypothesized that inter-vocalic duration could differentiate word boundaries. We measured inter-vocalic duration as illustrated in Figure 6. For a given vowel of rank n, its inter-vocalic duration measures the time span between the centers of the given vowel and its preceding vowel. For `determiner` vowels, the preceding vowel corresponds to the last vowel of the preceding word event though there is a breath, silence, hesitation, etc. But in this calculation, we excluded the inter-vocalic duration more than 3 seconds for not taking an extreme variation.

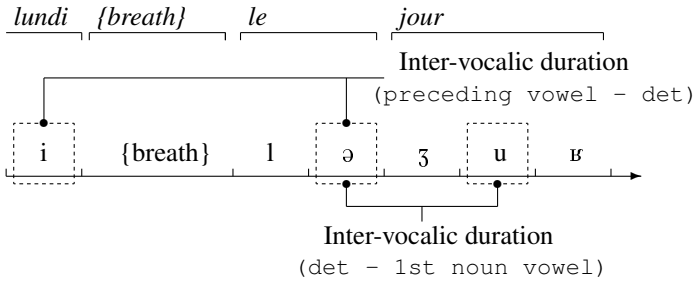

Figure 6: Inter-vocalic duration measurements.

The results for this measurement are illustrated in Figure 7. Figure 7 (top) presents the mean inter-vocalic duration profiles of `Noun` words (30 853 occ.) and Figure 7 (bottom) shows noun phrase (`determiner noun`) ones (12 888 occ.). In Figure 7 (top), we can observe inter-vocalic duration of last position n are longer. We also recognize that the first inter-vocalic duration of each syllable class are as much as long than last inter-vocalic duration. In Figure 7 (bottom), we can notice rapidly longest inter-vocalic dura-

tion (between preceding vowel - `determiner` vowel) for each n syllabic word class. The results from Figure 7 (bottom) are expectable because noun phrases can be uttered after breath, silence, a short pause. These factors produce longer inter-vocalic duration between `determiner` and preceding vowel. These results demonstrate that the inter-vocalic duration profiles, specially noun phrase case, allow locating word boundaries.

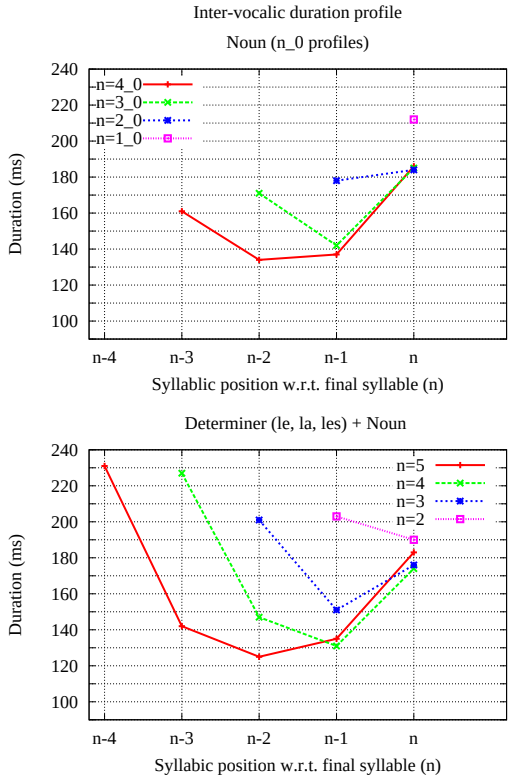

Figure 7: Mean inter-vocalic duration profiles for n-syllabic length. **Top:** Nouns (without final schwa) **Bottom:** Noun phrase `determiner noun` (without final schwa) with n variant from 2 to 5.

5. Conclusion

In this paper, 13 hours of broadcast news speech including a total of 165k words from male speakers were used to globally investigate lexical f_0 realisations via average f_0 profiles. As a working hypothesis, we considered that syllabic word length, presence/absence of word-final schwa and syntactic information are influential factors for prosodic structuring. Following this, speech data were first organized into subsets of words of equal syllabic lengths regardless of POS information, and then considering subsets of nouns and determiners. Words with realized final schwa were put into different subsets than words without final schwa. For each subset average f_0 and duration profiles were computed for word classes of given syllabic length, word final-schwa and noun phrases to reveal potential regularities of their prosodic contours which contributed for locating the word boundaries in audio corpus. Word boundary information was evidenced via average f_0 and duration profiles, namely word final syllable f_0 rises and long word final syllable lengths. Both are known tendencies in French. From noun phrase studies, results of average f_0 profiles illustrate lowest average f_0 values on `determiners` and a local maximum on the noun's first syllable. Inter-vocalic duration profile results show long inter-vocalic duration between `determiner` vowel and preceding word vowel highlighting a phrase boundary. These average results indicate that measurable cues contributing to word boundary location can be found in large speech corpora. Future studies will include other POS sequences, more detailed within syllable f_0 patterns, other speaking styles (especially spontaneous speech) and languages. The current findings will be implemented in an ASR post-processing step for improved word boundary location.

6. Acknowledgements

This work was partially funded by the research cluster Digiteo through a Région Ile-de-France doctoral grant to the first author and by OSEO under the Quaero program.

7. References

O. Bagou, C. Fougeron, and U. H. Frauenfelder. 2002. Contribution of prosody to the segmentation and storage of "words" in the acquisition of a new mini-language. In Bernard Bel and Isabelle Marlien, editors, *Proceedings of Speech Prosody*, pages 59–62, Aix-en-Provence, France.

P. Boersma and D. Weenink. 2005. Praat: doing phonetics by computer [computer program], from http://www.praat.org/. Technical report.

A. Cutler and D. Norris. 1988. The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, 14:113–121.

A. Cutler, D. Dahan, and W. Van Donselaar. 1997. Prosody in the comprehension of spoken language: a literature review. *Language and Speech*, 40(2):141–201.

C. Fougeron and S.-A. Jun. 1998. Rate Effects on French Intonation: Prosodic Organization and Phonetic Realization. *Journal of Phonetics*, 26:45–69.

O. Galibert. 2009. *Approches et méthodologies pour la réponse automatique à des questions adaptées à un cadre interactif en domaine ouvert*. Ph.D. thesis, Université Paris-Sud 11.

S. Galliano, G. Edouard, M. Djamel, C. Khalid, J.-F. Bonastre, and G. Guillaume. 2005. The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News. In *Proc. Interspeech*, Lisbonne, septembre.

J.-L. Gauvain, G. Adda, M. Adda-Decker, A. Allauzen, V. Gendner, L. Lamel, and H. Schwenk. 2005. Where Are We In Transcribing French Broadcast News? In *Proc. Interspeech*, Lisbonne, septembre.

C. Gendrot and M. Adda-Decker. 2006. Analyses formantiques automatiques en français : périphéralité des voyelles orales en fonction de la position prosodique. In *Proc. Journées d'Étude sur la Parole*, Dinard, juin.

Z. Harris. 1955. From phoneme to morpheme. *Language*, 31:190–222.

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

A. Lacheret-Dujour and F. Beaugendre. 1999. *La Prosodie du Français*. CNRS Éditions, Paris.

S. L. Mattys, P. W. Jusczyk, P. A. Luce, and J. L. Morgan. 1999. Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, 38(4):465–494.

J. M. McQueen. 1998. Segmentation of continuous speech using phonotactics. *Journal of Memory and Language*, 39:21–46.

J. R. Saffran, E. L. Newport, and R. N. Aslin. 1996. Word segmentation: The role of distributional cues. *Journal of Memory & Language*, 35:606–621.

H. Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester.

J. 't Hart. 1981. Differential sensitivity to pitch distance, particularly in speech. *Journal of Acoustical Society of America*, 69(3):811–21, March.

J. Vaissière. 1991. Rhythm, accentuation and final lengthening in French. *Music, Language, Speech and Brain. In Sundberg, J. et al. (eds.)*, pages 108–121.

P. Welby. 2003. French intonational rises and their role in speech seg mentation. In *Proceedings of Eurospeech: The 8th Annual Conference on Speech Communication and Technology*, pages 2125–2128, Geneva, Switzerland.

P. Welby. 2007. The role of early fundamental frequency rises and elbows in French word segmentation. *Speech Communication*, 49:28–48.