Rev. Estud. Ling., Belo Horizonte, v. 26, n. 4, p. 1551-1570, 2018

Automatic Speech Segmentation in French

Segmentação automática da fala em francês

Philippe Martin
LLF, UFRL, Université Paris-Diderot, Paris / France
philippe.martin@linguist.univ-paris-diderot.fr

Abstract: Whether we read aloud or silently, we segment speech not in words, but in accent phrases, i.e. sequences containing only one stressed syllable (excluding emphatic stress). In lexically stressed languages such as Italian or English, the location of stress in a noun, an adverb, a verb or an adjective (content words) is defined in the lexicon, and accent phrases include one single content word together with its associated grammatical words. In French, a language deprived from lexical stress, accent phrases are defined by the time it takes to read or pronounce them. Therefore, actual phrasing, i.e. the segmentation into accent phrases, depends strongly on the speech rate chosen by the speaker or the reader, whether in oral or silent reading mode. With a slow speech rate, all content words form accent phrases whose final syllables are stressed, whereas a fast speech rate could merge up to 10 or 11 syllables together in a single accent phrase with more than one content word. Based on this observation, and on other properties of stressed syllables, a computer algorithm for automatic phrasing, operating in a top-down fashion, is presented and applied to two examples of read and spontaneous speech.

Keywords: accent phrase; French; phrasing; stress location; boundary detection.

Resumo: Quando lemos em voz alta ou silenciosamente, segmentamos a fala em palavras, mas em grupos acentuais, i.e., sequências contendo uma única sílaba acentuada (excluindo-se acento enfático). Em línguas lexicalmente acentuadas como o italiano ou o inglês, a localização do acento em um substantivo, um advérbio, um verbo ou em um adjetivo (palavras lexicais) é definida no léxico, e sintagmas acentuais incluem uma única palavra lexical, acompanhada das palavras gramaticais a ela associadas. Em francês, uma língua que não possui acento lexical, sintagmas acentuais são definidos pelo tempo que se leva para lê-los ou pronunciá-los. Assim, os constituintes concretos,

eISSN: 2237-2083
DOI: 10.17851/2237-2083.26.4.1551-1570

i.e., a segmentação em grupos acentuais, depende fortemente da velocidade de fala escolhida pelo falante ou leitor, tanto na fala como na leitura silenciosa. Com uma velocidade de fala baixa, todas as palavras lexicais formam grupos acentuais cujas sílabas finais são acentuadas, enquanto o ritmo de fala rápido poderia juntar de 10 a 11 sílabas em um mesmo grupo acentual contendo mais de uma palavra lexical. Com base nessa observação e em outras propriedades das sílabas acentuadas, um algoritmo computacional para segmentação automática, atuando de maneira top-down é apresentado e aplicado a dois exemplos de leitura e fala espontânea.

Palavras-chave: grupo acentual; francês; segmentação; posição do acento; detecção de fronteira.

Submitted on January 12th, 2018

Accepted on February 27th, 2018

1 Introduction

When we read a text, either aloud or silently, we could proceed word by word, or even syllable by syllable, but if we master the language and identify all the words, we usually proceed by group of words. It is easy to observe in an orthographic transcription where all words would be ended by a final dot that we don't read word by word, as it would be the case in: *In. the. Orthographic. Representation. Of. Speech. Of. Most. Written. Languages. Segmentation. Is. Defined. By. Spaces. Between. Words.* Instead, we normally read a sentence by grouping words in units containing either a noun, an adverb, a verb or an adjective (i.e. a content word), together with the grammatical words (pronoun, conjunction…) associated to them to form an *accent phrase*. The preceding example, segmented in accent phrases, indicated in squared brackets, would be: [*in the orthographic*] [*representation*] [*of speech*] [*of most*] [*written*] [*languages*] [*segmentation*] [*is defined*] [*by spaces*] [*between*] [*words*]. Each of these groups carry a single stressed syllable placed on some syllable of the content words as defined in the lexicon of English: [*in the ortho**gra**phic*] [*represen**ta**tion*] [*of **speech***] [*of **most***] [***wri**tten*] [***lan**guages*] [*segmen**ta**tion*] [*is de**fi**ned*] [*by **spa**ces*] [*bet**ween***] [***words***]. Such groups of words are called in prosodic phonology *accent phrases*, and define the minimal prosodic units, which organized into a hierarchy, constitute the *prosodic structure* of the sentence (MARTIN, 1975, SELKIRK, 1978).

For all fluent speakers of English, the position of stressed syllables in accent phrases is predictable, and results from the acquisition of the lexicon of the language. Other stressed syllables can also occur in speaker's production, but contrary to lexically defined stress, they are not predictable as they result from a specific choice of the speaker to indicate an emphasis, as in ***seg****men****ta****tion in most* ***wri****tten* ***lan****guages* with a stress on the first syllable of ***seg****mentation*. This kind of emphatic stress may occur on a different syllable than the lexically stressed or on the same syllable. In this latter case, the speaker will use a different acoustic realization, as emphatic stress has to be perceived by listeners as different and unpredictable compared to the predictable lexical stress.

The predictability of lexical stress suggests that the perception of stressed syllables may not be directly derived from the processing of specific acoustic features of speech, such as vowel duration, fundamental frequency change or intensity modulation, the prosodic parameters often mentioned in the literature as parameters of stress. Instead, the perception of stressed syllables could be considered as the result of an identification mechanism comparing the actual acoustic features of syllables with a predicted position derived from the knowledge of the language. As in silent reading as well as reading aloud, segmentation into accent phrases is inevitable, the same process takes place when we listen to somebody speaking, eventually restoring stress in a position where we would have placed the stressed syllable ourselves.

One can mention on this topic the experiment on the perception of accented syllables of Berber and Hebrew by subjects who had no notion of these languages at all (METTOUCHI *et al*., 2007). The acoustic features are present in the speech signal, but in this experiment the listeners didn't identify any stress locations (except by chance…), positioning stress on syllables belonging to sequences they thought they had identified through the perception grid of their mother tongue (or another they knew). Indeed, no appropriate lexicon allowing the listeners to position an expected stressed syllable and interpret the acoustic data was available, which is not the case for speakers of Berber or Hebrew. Similar observations can be found in Astésano and Bertrand (2016) and Michelas *et al*. (2016).

2 The case of French

French is a language where the position of lexical stress evolved gradually to the last syllable of content words (actually to the last syllable of any word pronounced in isolation) by progressively dropping all post-stressed syllables (VÄÄNÄNEN, 1995). The function of lexical stress as marker of morphological boundary as in lexically stressed languages was gradually lost as redundant. Since its main phonological function was lost, it became then possible for speakers to skip some of the predicted stress locations when speaking or reading. This can be seen in *la ville de Versailles* ("the city of Versailles"); which can be read with one or two stressed syllables, placed on the last syllable of content words *ville* and *Versailles*: *la ville de Ver**saill**es* or *la **vill**e de Ver**saill**es*. Likewise, an example such as *la petite armoire violette* ("the little purple cupboard"), can receive one, two or even three stressed syllables: *la petite armoire vio**lett**e*, *la petite ar**moir**e vio**lett**e*, *la pe**tit**e armoire vio**lett**e* or *la pe**tit**e ar**moir**e vio**lette***. For a French speaker, it is easy to realize that the difference in phrasing of these examples is linked to the speech rate, possibly leading to a different processing of the sentence content. In order to pronounce (or even to read silently) *la petite armoire vio**lett**e* with only one final stressed syllable on *vio**lett**e*, one has to use a (very) fast speech rate, whereas a slower pace would lead to the pronunciation of three stressed syllables as in *la pe**tit**e ar**moir**e vio**lette***. Surprisingly, this dependency of phrasing to the speech rate seems to escape some researchers who are native speakers of French, as it appears in a recent issue of the review *Langue Française* (2016, n. 191), gathering papers devoted to *accentuation et phrasé*. The absence of the time parameter implied in phrasing even lead to the often-mentioned belief that French listeners are ´deaf´ to stress…

We could perhaps then conclude that there is no limit to the number of syllables and thus of words that can be pronounced in French with only one final stressed syllables, and that can be inserted in a single accent phrase. The pronunciation of long words will help discover where the limit stands. Long words such as the well-known *anticonstitutionnelle**ment*** ("against the constitution"), (8 syllables) or *intergouvernementalisa**tion*** ("inter governmentalization") (10 syllables) seem difficult if not impossible to pronounce or read even silently with only one final stressed syllable. Already in the 16th century, the

grammarian Louis Meigret (1550) proposed that the longest word that could be pronounced with only one final stress would have a maximum of 7 syllables. Much later, Martin (2014) showed that it was not the number of syllables that matters, but the time needed to pronounce them, even in silent reading. The data obtained from fast speech rate speakers suggest that the maximum interval between consecutive stressed syllables (in flowing speech) could not exceed some 1,250 ms, depending on the subjects. In *parler jeune* productions in French (the young people speaking style), sequences of up to 10 or 11 syllables with only one final stress have been observed (LEKHA; LE GAC, 2004). This value is close to the theoretical limit, derived from the minimal average duration of syllables that could be perceived in a sequence, about 100 ms (GHITZA; GREENBERG, 2009). These observations would put the maximal duration of accent phrases in French to about 1,250 ms to 1,400 ms or so, with the fastest speech rate reaching about 11 or 12 syllables per second.

If 1,250 ms (the approximative value retained in this paper) is the maximal duration between consecutive stressed syllables in connected speech, there is also a minimal duration that exists between two consecutive stressed syllables. This value will define a minimal duration for accent phrases that would contain only one syllable. Its value is experimentally easy to evaluate, by selecting natural or synthetic occurrences of consecutive stressed syllables, as for example *par le **fait que*** ("by the fact that") or *le travail de **nuit nuit*** ("night work harms") i.e. cases of stress clash with no move or deletion of the first stress. It is often mentioned in the literature that these cases require a kind of acoustic gap between consecutive stressed syllables (e.g. DI CRISTO, 2016), usually but not always implemented by the presence of consonants after the first of before the second stressed syllable (which is the case for the two examples above). However, it is easy to experimentally reduce the gap with a sound editor until the first implied syllable ceased to be perceived as stressed although nothing of its acoustical structure has been modified (i.e. by removing the silent part only). This limit is about 250 ms (depending on the way distances are measured between syllables, from their center or from the two third of their duration), which gives the minimal duration of an accent phrase, since below this value, the word owning the first syllable will become part of the newly formed accent phrase. For example, the perceived desaccentuation of *fait* in [*par le **fait***] [***que***] ("by the fact that") will merge the accent phrase *par le **fait*** with

the second accent phrase ***que*** to form the new group [*par le fait* ***que***]. The minimal duration between two consecutive stressed syllables is thus about 250 ms (MARTIN, 2014), which implies that a one-syllable accent phrase must include some voiceless or silent segment that precedes it, as the preceding vowel, if exists, is necessarily stressed and ends the preceding accent phrase.

3 Syllables followed by silence are stressed in French

It is equally easy to demonstrate experimentally that any syllable followed by at least 250 ms of silence is perceived as stressed in French. The fact that any final syllable is perceived as stressed is a consequence of the prepositioning of the stressed syllable by the listener, as final syllables are stressed in French, and that a silent gap following the end of an accent phrase is necessarily stressed.

Either by inserting some 250 ms or more acoustic silence on the speech wave, using a sound editor without modifying the acoustic characteristics of the syllable at all, or by simply slowing the speech rate so that the number of syllables reaches a level below some four syllables per second, the final syllables of any word category becomes perceived as stressed, whatever their actual duration or pitch movement. In lexically stressed languages, the perception of an accent phrase final syllable as stressed is preempted by the position of lexical stress (if not in final position). In Italian for example, the lexical stress of the penultimate syllable of *Marco* in *la so**re**lla di **Mar**co è partita* ("Marco's sister left") prevents a listener who knows the language to perceive the last syllable *co* as stressed, although it is followed by more than 250 ms of silence, whereas a speaker of French who does not know Italian will perceive the final syllable of *Marco* as stressed.

The important parameter in these cases pertain to the lack of speech data to be processed by the listener and the actual explanation is linked to the processing of syllables by the brain, and more precisely by the brain oscillations carrying information between neuronal zones (MARTIN, 2015). As mentioned above, it can be shown that the perception of syllables needs at least 100 ms processing time, even if their actual duration is below this value. If given more than some 250 ms, a normally unstressed syllable becomes perceived as stressed, without modification of its acoustic structure. Since two consecutive stressed

syllables must be separated by at least 250 ms, we can conclude that the perception of *stressed* syllables needs at least 250 ms processing time. This is a consequence of the processing of stressed syllables by delta brain waves (MARTIN, 2018).

In summary, a normally unstressed syllable can be perceived as stressed by timing characteristics pertaining to a silent gap following the syllable itself. Likewise, a normally stressed syllable can be perceived unstressed for a similar reason, the gap duration existing between two consecutive syllables.

4 Pronouns

The *pronoms toniques* in French (*moi, toi, lui, elle, nous, vous, eux, elles*) do not belong to the category of content words, but share their characteristics in term of accent phrase stress, in particular in examples with a tonic pronoun placed after the verb. The normal stress pattern of *redonne moi la main*, [*re**donn**e*] [*moi la **main***] ("give me your hand again") leads to the unexpected accent phrase [*moi la **main***], *redonne **moi** la **main*** being emphatic, the stress pattern [*red**onn**e*] [*la **moi***] [*plus **loin***] ("give me it further") is quite possible and leads to consider some tonic pronouns as stressable even if they are not followed by 250 ms of silence. There are cases where tonic pronouns are effectively stressed, although they do not belong to the content word category. In other configurations, as in *moi ma mère le salon c'est de la moquette*, the tonic pronoun *moi* is stressed if followed by 250 ms of silence, ***moi*** # *ma **mèr**e le sa**lon** c'est de la mo**quett**e* ("me my mother the living room is carpet"), but unstressed if there is no sufficient gap after *moi*, as predicted: *moi ma **mèr**e*... The same configuration can be observed in well-known examples such as *mon manège à **moi** c'est **toi*** ("my ride to you is me"), from a famous Edith Piaf song, or ***Je** est un **autr**e* ("I am another"), Arthur Rimbaud.

Likewise, demonstrative pronouns are also stressable although they don't belong to the content word category. In *...pour tous **ceux** et toutes **cell**es*... (NS) "for all those…", both demonstrative pronouns are stressable and stressed. The same observation applies to possessive pronouns such as *le **mien**, le **tien**, la **leur**, les **leurs***.... "mine, yours, their, theirs".

Finally, relative pronouns (*qui, que, quoi, dont, où, lequel…)* are also stressable, but become stressed mainly if followed by a 250 ms silence.

5 Eurhythmy

The eurhythmicity observed for both read and spontaneous speech may also be taken into account in a top-down approach for prosodic segmentation, i.e. an approach not processing from acoustic data to phonological conclusions, but rather from the general properties of stressed syllables to eventually validate acoustic data in the speech signal. As a general observation (WIOLAND, 1985), spontaneous speech eurhythmy proceeds by adjusting the average duration of accent phrases syllables to reach comparable duration of successive accent phrases. Read speech uses more often a strategy aiming to balance the number of syllables of successive accent phrases, at the possible expense of congruence with the syntactic structure. A classic example is given by a sentence such as *Marie adore les chocolats* ("Mary loves chocolates") in which spontaneous speech subjects would have a tendency to realize a phrasing congruent with syntax **[***Ma**rie***] [***adore les choco**lat**s***]** and possibly aim for eurhythmy by slowing the syllabic rate of [*Marie*] and going faster on [*adore les chocolats*]. On the contrary, readers of this sentence show a tendency to group the words to balance the number of syllables in consecutive accent phrases, at the expense of congruence with syntax [*Marie a**dor**e*] [*les choco**lat**s*].

To implement eurhythmicity in a segmentation algorithm, an average duration of accent phrases can be estimated in a running window containing some 3 or 4 consecutive accent phrases. This value should be between about 250 ms (each syllable is followed by 250 ms silence, a production style where all syllables are pronounced detached) to about 1250 ms, characteristic of the *parler jeune*. Assuming the speaker or reader rhythm does not vary too much in a given amount of time, a more or less reliable duration value is obtained from two or three consecutive accent phrases duration values. Experimental data obtained from spontaneous speech show that the average accent phrase duration is about 500 to 700 ms (MARTIN, 2018).

To summarize the properties and observations on accent phrase stress in French:

1. Duration of accent phrases (in French): between 250 ms and 1,250 ms;
2. Accent phrases may contain 1 to 11 syllables ;
3. The minimal and maximal speech rates are between 4 and 10 syllables per second (in continuous speech);
4. Any syllable followed by more than 250 ms silence is perceived as stressed;
5. Eurhythmicity aims to balance the duration of successive accent phrases.

6 Virtual and actual stress, stressable and stressed syllables

Phrasing determines an essential step in the comprehension of speech. The segmentation into accent phrases constitutes the first phase to rebuild the prosodic structure intended by a speaker, which is essential and unavoidable to access the syntactic structure when we read. The resulting prosodic structure will not necessarily match the prosodic structure intended by the writer of the text we read, which leads to consider reading resulting from our own segmentation of the text, as the phrasing depends on the reading speed selected, and this is true in both reading aloud or silently. The only limits to these variations are given by the minimal and maximal duration of accent phrases.

The simple fact that we can restore stress locations when we read aloud or silently tells us that we may not really need any acoustical input to perceive stressed syllables (again non-emphatic). Not only reading aloud or silently of the same text could lead to different phrasings, but while listening to speech, we cannot prevent to have expectations towards the location of stressed syllables different from the one actually realized by the speaker. In other words, we can "hear" stressed syllables that actually may not be present acoustically. This apparent illusion is a direct consequence of many perception processes in speech (ARNAL; GIRAUD, 2017) involving not a direct processing of some physical input, but rather the validation of an expected input by comparison between what's expected and what is actually physically realized. In the case of accent phrase defined by a final stress, we can predict from our lexicon the location of a stressed syllable in a group of words, which will depend on the speech rate selected in this operation. Considering again the former example *la petite armoire violette*, the speaker could have stressed *armoire* and *violette*, *la petite ar**moir**e vio**lett**e*, but we may

have expected a slower speech rate and mentally also stressed *petit*: *la pe**tit**e ar**moir**e vio**lett**e*. Consequently, we could then hear three stressed syllables although the speaker had realized only two. The only way to avoid this perception of this *virtual* stress (for syllables that would be stressed), opposed to the *actual* stress (for effectively stressed syllables) present in the acoustic wave, would be to constantly adapt our speech rate to the one used by the speaker, or the one assumed to be used by the scripter.

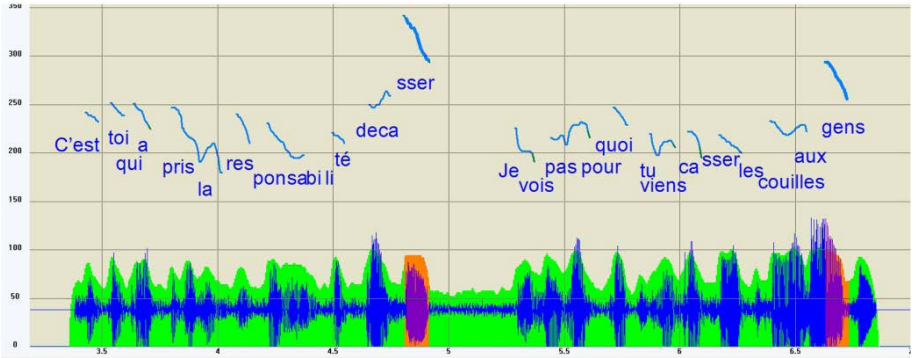
This adaptation is not always easy or even possible. The examples provided by the *parler jeune* with a very fast speech rate exceeding 7 or 8 syllables per second are hard to match for most listeners, to the point that some will have trouble to understand such speech tempo. Therefore, some listeners will have a tendency to hear stressed syllables where they do not exist acoustically. In the example illustrated in Fig. 1 displaying a speech wave and the corresponding fundamental frequency curve, the actual accent phrases acoustically realized by the speaker are

> [*C'est toi qui a pris la responsabilité de ca**sser***] pronounced with 13 syllables
>
> [*Je vois pas pourquoi tu viens casser les couilles aux **gens***] 12 syllables
>
> ("you took the responsability to break up. I do not see why you come to break people's balls"). (corpus *l'Esquive*)

The first accent phrase contains 13 syllables, which is unusual for the average speaker of French. Therefore, any listener not practicing the *parler jeune* will have a strong tendency to restore mentally a stressed syllable on *toi* leading to a different phrasing [*C'est **toi***] 2 syllables [*qui a pris la responsabilité de ca**sser***] 11 syllables (elision of [i] in *qui a*), or even also on *pris*, resulting in a four accent phrases phrasing [*C'est **toi***] 2 syllables [*qui a **pris***] 2 syllables pronounced [kapri] [*la responsabilit**é***] 7 syllables, [*de ca**sser***] 3 syllables. Still, as shown on Fig. 1, the only obvious acoustical marker of stress is on the final syllable of *ca**sser***.

Likewise, the second accent phrase with 12 syllables, could be mentally segmented into 2 or 3 accent phrases, depending on the speech rate adopted mentally or in oral production: [*Je vois pas pour**quoi***] [*tu viens casser les couilles aux **gens***] (the [ə] of the pronoun *je* is deleted here)*, or* [*Je vois pas pour**quoi***] [*tu viens ca**sser***] [*les couilles aux **gens***].

FIGURE 1 – An example of fast speech rate with 13 and 12 syllables in the accent phrases: *C'est toi qui a pris la responsabilité de casser. Je vois pas pourquoi tu viens casser les couilles aux gens* («you took the responsability to break up. I do not see why you come to break people's balls").



The possible discrepancy between perceived and realized stressed syllables in French leads to differentiating virtual from actual stressed syllables. Virtual syllables correspond to what Paul Garde (1968, 2013) called *stressable*, whereas the syllable effectively marked by acoustic parameters are then *stressed*. The number of stressable syllables is necessarily equal or superior to the number of actually stressed ones, whose number depends on the speech rate.

7 Stress annotation: *mission impossible ?*

The problem for an annotator of stressed syllables (outside emphasis) in French is to adapt to the speech rate of the recording when accented syllables are annotated. The perception of stress will be influenced by the annotator's own prediction process, thereby tending to detect stressed syllables where they would have been placed by reading or speaking not at the speaker's speech rate but at the annotator's own pace.

Most often, implemented automatic detection of stressed syllables in French operates in a bottom-up fashion from the speech recording, looking for significant variations between consecutive syllables in duration, fundamental frequency and intensity (for recent examples, see GOLDMAN *et al*., 2013; MERTENS; SIMON, 2013). Vowel quality does not appear as a significant parameter for stress detection in French. Still, top-down approaches do exist, essentially applied to English (ARNOLD; WAGNER, 2008), operating from the word category to detect syllabic stress.

In a paper published in 2013, M. Avanzi, faced with the uncertainty of annotating stressed syllables in French, describes in detail a complex procedure involving two experts, possibly helped by a third in case of disagreement between the first two. Even with this protocol, agreement between annotators varies between 60 % and 80 %. In Martin (2006), some explanations were already proposed for the lack of convergence observed in the perception of stressed syllables in French by experts. These explanations pertain to the expectations of stress placement by various annotators, experts or non-experts.

In another paper on the same topic, Christodoulides and Avanzi (2014) implemented an automatic detector of prominence (i.e. not just accent phrase stressed syllables) by machine learning methods applied to a large corpus (11 hours) which included two different speech styles. They use a comprehensive set of acoustic parameters that they hoped would be appropriate to differentiate prominent syllables from others (syllabic duration, minimal, maximal average fundamental frequency, pitch movement, peak intensity, spectral balance, part of speech tag, presence and duration of subsequent pause, syllabic structure, position of the syllable in the word). Their best results, evaluated against manual placement by experts in syllabic prominence (therefore subject to the limitations evoked above), reaches a 90% correct identification level.

Considering these difficulties, it appears that stress detection should proceed not from the speech wave analysis, but rather from the knowledge a system could have access to beforehand, the location of potential positions as final syllables of content words among others, i.e. proceed in a top-down fashion.

Indeed, as we have seen above, the perception of stressed syllables by listeners proceeds not by direct evaluation of actual acoustic parameters in the speech wave, but rather by comparison of listener expected stress locations with perceived acoustic parameters. In this process, the evaluation of an expected stress position and actual realization by the speaker precedes the actual validation process comparing expectation and reality. This explains why even expert listeners may perceive as stressed syllables not carrying specific acoustic features differentiating from surrounding syllables, and how we restore stressed syllables in silent reading without any actual acoustic information.

To attain a reasonable chance of success, a computer implementation dealing with speech wave should then adopt a comparable

strategy, and not infer results starting from acoustical analysis of the speech wave but rather from expectation of stressed syllable locations. The availability of transcribed and segmented speech data, down to the syllabic and phone level, should be a prerequisite towards automatic stress detection, as the candidates for syllabic stress can be directly inferred from the aligned transcription.

8 A top-down algorithm

To apply the definition given in lexically stressed languages to French, we can assign a virtual stress to final syllables of all words belonging to the category of noun, verb, adverb, adjective and pronoun. To help select actual stressed syllables among the list of stressable ones, we can in a first step use the constrains described above, i.e. the minimal and maximal duration of accent phrases (respectively 250 ms and 1,250 ms), the minimal separation of 250 ms between two consecutive stressed syllables, and the presence of at least 250 ms of silence following a virtual stress. The application of these constraints would make some virtual candidate stressed syllables actually stressed (as a unique stressable syllable) in a time window of 1,250 ms for example), and eliminate some from the list of possible actual stressed syllables (the first stressable syllable cannot be actually stressed if closer to the next stressed syllable by less than 250 ms).

As stated above, the next step to select stressed syllables effectively without even starting accent phrase looking at the speech wave would be to look at the speech rate, i.e. the number of syllables per second actually observed on the transcription of the speech wave. Linked to an average number of syllables per accent phrase, we can then have an approximation of the phrasing realized in a given recording, validated by an assumed eurhythmicity.

To finally exploit the actual acoustic data, and innovate from the existing list of traditional parameters, i.e. changes / contrasts in syllabic duration, fundamental frequency and frequency, we could refer to the function of syllabic stress to define accent phrases as minimal units of the sentence prosodic structure. According to the model of Martin (1975, 1987), the prosodic structure results from a dynamically built hierarchical organization of accent phrases. From the presence of an expected terminal conclusive contour, perceived as a marker of non-

continuation of the sentence, two other melodic contours, one rising, the other falling, indicate respectively a major and a minor continuity.

The interesting characteristics of the continuity contours (always located on the vowel of the stressed syllable) is that they indicate a dependency relation, minor continuation towards major continuation and major continuation towards the terminal conclusive contour, by a contrast of melodic slope, where a falling contour indicates a dependency toward a rising contour. Of course, this model implies that the falling and rising melodic slope are effectively perceived, i.e. that the speed of melodic change in time is above what is called the glissando threshold. The glissando threshold is evaluated as the difference from the beginning to the end in semitones referred to the duration of the contour (assuming a linear variation, cf. ROSSI, 1971).

According to this definition of accent phrases as minimal units of prosody whose hierarchy constitute the sentence prosodic structure, we can designate any stressable syllable whose change in fundamental frequency on its vowel exceeds the glissando threshold as effectively stressed. Although this step assumes the validity of the glissando threshold (which in fact implies an adjustment parameter), as well as the linearity of the fundamental frequency change of the syllable vowel used for the evaluation of the glissando value, we have enough tools to implement an innovative algorithm for automatic selection of stressed syllables from a list of stressable syllables.

9 Automatic detection of stressed syllables in French

From these various observations and considerations, the following rules for a computer implementation can be applied:

1. Any syllable followed by more than 250 ms silence is stressed;
2. Any final syllable of a noun, adjective, verb, adverb or pronoun is stressable (from accent phrase definition);
3. If two consecutive stressed syllables are separated by less than 250 ms, the first one is unstressed (accent phrase minimum duration from the minimum spacing between consecutive stressed syllables);
4. Any stressable syllable with change of fundamental frequency over the glissando threshold is stressed;

5. If two consecutive stressed syllables are separated by more than 1,250 ms in continuous speech, at least one stressable syllable in this interval is stressed (accent phrase maximum duration). Make stressed the one with the highest glissando value;
6. One stressable syllable must exist in any time window duration equal to the accent phrase average duration (eurhythmy).

The eurhythmic aspect is implemented by evaluating the first accent phrases realizations and the number of syllables they contain. This starting accent phrase duration will then be used to define a sliding time window, in which most prominent syllables in value of glissando will be retained as stressed. The size of this sliding window defines a speech rate assumed to be constant in the whole recording.

10 An example of read speech

A first read example: *il était une fois un pauvre escargot qui souffrait beaucoup à chaque fois qu'il partait en randonnée car il avait du mal à suivre le rythme de ses compagnons* ("Once upon a time, there was a poor snail who suffered a lot every time he went on a hike because he had trouble keeping pace with his companions").

In the steps detailed below, stressable syllables are underlined, and stressed syllables are underlined and bold.

Step 1: Any syllable followed by more than 250 ms silence is stressed:

*Il était une fois un pauvre escargot qui souffrait beaucoup à chaque fois qu'il partait en rando**nnée***

Step 2: Any final syllable of a noun, adjective, verb, adverb or tonic pronoun is stressable:

*Il était une fois un pauvre escargot qui souffrait beaucoup à chaque fois qu'il partait en rando**nnée***

Step 3: If two consecutive stressed syllables are separated by less than 250 ms, the first one is unstressed: the gap between *chaque* and *fois* is 180 ms, below the 250 ms limit:

Il était une fois un pauvre escargot qui souffrait beaucoup à chaque |180 ms| *fois qu'il partait en rando**nnée***

Step 4: Any stressable syllable with F0 change over the glissando threshold is stressed {glissando value/glissando threshold with coefficient 0.16}. The stressable syllables below the threshold are unstressed:

Il était {35/76} *une fois* {36/17} *un pauvre* {44/66} *escargot* {32/12} *qui souffrait* {54/144} *beaucoup* {79/66} *à chaque fois* {46/106} *qu'il partait* {32/51} *en randonnée*

Step 5: Two consecutive stressed syllables separated by more than 1,250 ms, as in the case of the last accent phrase:

[*à chaque fois qu'il partait en randonnée*]1367 ms

We can select the highest glissando value, on *fois*:

[*à chaque fois qu'il partait en randonnée*] 1367 ms

or both stressable syllables on *fois* and *partait*:

[*à chaque fois qu'il partait en randonnée*] 1367 ms

Step 6: Apply eurhythmicity to retain the latter possibility:

726 ms 5 syl. 145 ms/syl. *Il était une fois*

687 ms 5 syl. 137 ms/syl. *un pauvre escargot*

765 ms 5 syl. 153 ms/syl. *qui souffrait beaucoup*

407 ms 3 syl. 135 ms/syl. *à chaque fois*

487 ms 3 syl. 162 ms/syl. *qu'il partait*

546 ms 4 syl. 136 ms/syl.*en randonnée*

The average accent phrase duration is about 709 ms.

11 An example of spontaneous speech

The second example belongs to the category of *parler jeune*: *Juste pour une carte d'identité t'as pas ta carte tu fais tes vingt-quatre heures tu ressors t'as la haine encore plus ça augmente* ("Just for an identity card you do not have your card you make your twenty-four hours you come out you hate even more it increases").

Step 1: The last syllable is followed by more than 250 ms of silence:

Juste pour une carte d'identité t'as pas ta carte tu fais tes vingt-quatre heures tu ressors t'as la haine encore plus ça augmente

Step 2: Any final syllable of a noun, adjective, verb, adverb or tonic pronoun is stressable:

Juste pour une carte d'identité t'as pas ta carte tu fais tes vingt-quatre heures tu ressors t'as la haine encore plus ça augmente

Step 3: If two consecutive stressed syllables are separated by less than 250 ms, the first one is unstressed: the gap between *encore* and *plus* is 240 ms, below the 250 ms limit:

Juste pour une carte d'identité t'as pas ta carte tu fais tes vingt-quatre |230 ms| *heures tu ressors t'as la haine encore* |240 ms| *plus ça augmente*

Step 4: Any stressable syllable with F0 change over the glissando threshold is stressed. The stressable syllables below the threshold are unstressed:

Juste {64/36} *pour une carte* {44/38} *d'identité* {54/45} *t'as pas ta carte* {44/38} *tu fais* {54/142} *tes vingt-quatre* |230 ms| *heures* {49/37} *tu ressors* {38/32} *t'as la haine* {25/22} *encore* |240 ms| *plus* {38/23} *ça augmente*

Steps 5 and 6 do not apply:

227 ms 1 syl. 227 ms/syl *Juste*

356 ms 3 syl. 118 ms/syl. *pour une carte*

537 ms 4 syl. 134 ms/syl. *d'identité*

569 ms 4 syl. 142 ms/syl. *t'as pas ta carte*

945 ms 6 syl. 157 ms/syl. *tu fais tes vingt-quatre heures*

486 ms 3 syl. 162 ms/syl. *tu ressors*

431 ms 3 syl. 143 ms/syl. *t'as la haine*

496 ms 3 syl. 165 ms/syl. *encore plus*

592 ms 3 syl. 197 ms/syl. *ça augmente*

The average accent phrase duration is 515 ms.

12 Conclusion

Contrary to lexically stressed languages such as English or Italian, in which accent phrases contain one stressed syllable usually carried by a

content word, French segmentation into accent phrases depends strongly on the speaking or reading rate used. In fact, the only limitation for the number of words contained in a single accent phrase in French is the time taken to pronounce them, which cannot exceed some 1,250 ms, even in silent reading or speaking to oneself.

In view of this property, and of the fact that the perception of stressed syllable results from a validation process comparing the predicted position with the actual acoustic parameters, a top-down automatic phrasing segmentation in French is briefly described. The algorithm incorporates the following observations: 1) Speakers and readers of French are capable to restore accent phrase stressed syllables even without any acoustic input; 2) The minimum duration of accent phrases is 250 ms, and the maximum about 1,250 ms; 3) The actual duration of accent phrases depends on the speech rate selected by the speaker or the reader; 4) The actual syllabic stress defining phrasing carries a melodic movement above the glissando threshold.

References

ARNAL, L. ; GIRAUD, A-L. Neurophysiologie de la perception de la parole et multisensorialité. In : PINTO, Serge ; SATO, Marc (Éd.). *Traité de neurolinguistique*. Louvain-la-Neuve : De Boeck, 2017. p. 97-108.

ARNOLD, D.; WAGNER, P. The influence of top-down expectations on the perception of syllable prominence. In: TUTORIAL AND RESEARCH WORKSHOP ON EXPERIMENTAL LINGUISTICS (ISCA), 2., Athens, Greece, 2008. *Proceedings*… Athens: University of Athens, 2008. p. 25-28,

ASTÉSANO, C. ; BERTRAND, R. Accentuation et niveaux de constituance en français : enjeux phonologiques et psycholinguistiques. *Langue Française*, Paris, v. 191, n. 3, p. 11-30, 2016. Doi: 10.3917/lf.191.0011

AVANZI, M. Note de recherche sur l'accentuation et le phrasé à la lumière des corpus du français. *Tranel*, Neuchâtel, Suisse, v. 58, p. 5-24, 2013.

CHRISTODOULIDES, G.; AVANZI, M. An Evaluation of Machine Learning Methods for Prominence Detection in French. In: ANNUAL CONFERENCE OF THE INTERNATIONAL SPEECH

COMMUNICATION ASSOCIATION, 15., 2014, Singapore. *Proceedings*... Singapore : International Speech Communication Association (ISCA), 2014. p. 116-119.

DI CRISTO, A. *Les musiques du français parlé*. Berlin : De Gruyter Mouton, 2016.

GARDE, P. *L'accent*. Paris : Presses Universitaires de France, 1968. (Collection SUP « Le linguiste », n. 5)

GARDE, P. *L'accent*. Paris : Lambert-Lucas, 2013.

GOLDMAN, J-P. ; AUCHLIN, A. ; ROEKHAUT, S. ; SIMON, A-C. ; AVANZI, M. Prominence perception and accent detection in French. A corpus-based account. *Language Science*, Elsevier, v. 39, p. 95-106, 2013.

GHITZA, O.; GREENBERG, S., On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, Basel, Suisse, v. 66, n. 1-2, p. 113-126, 2009.

LEHKA, I. ; LE GAC, D. Etude d'un marqueur prosodique de l'accent de banlieue. In : JOURNÉES D'ÉTUDES SUR LA PAROLE, XXV., Fès, Maroc, avril 2004. *Actes*... Fès, Maroc : Association Francophone de la Communication Parlée, 2004. Disponible sur : <http://www.afcp-parole.org/doc/Archives_JEP/2004_XXVe_JEP_Fes/actes/jep2004/Lehka-LeGac.pdf>. Accessed on January 12th, 2018.

LANGUE FRANÇAISE. *La prosodie du français : accentuation et phrasé*. Paris , v. 191, n. 3, 2016. 150 p. Doi: 10.3917/lf.191.0005

MARTIN, Ph. Analyse phonologique de la phrase française. *Linguistics*, v. 146, p. 3568, Fév. 1975.

MARTIN, Ph. La transcription des proéminences accentuelles : mission impossible ? *Bulletin PFC*, Toulouse, n. 6, p. 81-87, Sept. 2006.

MARTIN, Ph. Spontaneous speech corpus data validates prosodic constraints. In: INTERNATIONAL CONFERENCE ON SPEECH PROSODY, 6., 2012, Shangai, China. *Proceedings*... Shangai: Tongji University Press, 2014. p. 525-529.

MARTIN, Ph. *The Structure of Spoken Language. Intonation in Romance*. Cambridge: Cambridge University Press, 2015.

MARTIN, Ph. *Intonation, structure prosodique et ondes cérébrales*. London: ISTE Editions, 2018.

MEIGRET, L. *Le tretté de la grammére françoéze*. Paris : C. Wechel, 1550. Disponible sur : <http://gallica.bnf.fr/ark:/12148/bpt6k507854/f1.image>. Accessed on January 12th, 2018.

MERTENS, P.; SIMON, A-C. Towards automatic detection of prosodic boundaries in spoken French. In: PROSODY-DISCOURSE INTERFACE CONFERENCE (IDP-2013), 2013, Leuven. *Proceedings*… Leuven: KU Leuven, 2013. p. 81-87.

METTOUCHI, A.; LACHERET-DUJOUR, A.; SILBER-VAROD, V.; IZRE, Shlomo El. Only Prosody ? Perception of speech segmentation in Kabyle and Hebrew. *Cahiers de Linguistique Française*, Genève, v. 28, p. 207-218, 2007.

MICHELAS, A.; FRAUENFELDER, U. H.; SCHÖN, D.; DUFOUR, S. How deaf are French speakers to stress? *Journal of the Acoustical Society of America*, [s.l.], v. 139, n. 3, p. 1333-1342, 2016. Doi: https://doi.org/10.1121/1.4944574

ROSSI, M. Le seuil de glissando ou seuil de perception des variations tonales pour la parole. *Phonetica*, Aix-en-Provence, n. 23, p. 1-33, 1971. Doi:10.1159/000259328

SELKIRK, E. O. On prosodic structure and its relation to syntactic structure. In: FRETHEIM, T. (Ed.). *Nordic Prosody II*. Trondheim: TAPIR, 1978. p. 111-140.

VÄÄNÄNEN, V. *Introducción al latin vulgar*. Madrid: Editorial Gredos, 1995.

WIOLAND, F. *Les structures rythmiques du français*. Paris : Slatkine-Champion, 1985.