DOI 10.3758/s13428-012-0285-y

Diphones-fr: A French database of diphone positional frequency

Boris New · Elsa Spinelli

Published online: 13 December 2012

Abstract The aim of this article is to describe a database of diphone positional frequencies in French. More specifically, we provide frequencies for word-initial, word-internal, and word-final diphones of all words extracted from a subtitle corpus of 50 million words that come from movie and TV series dialogue. We also provide intra- and intersyllable diphone frequencies, as well as interword diphone frequencies. To our knowledge, no other such tool is available to psycholinguists for the study of French sequential probabilities. This database and its new indicators should help researchers conducting new studies on speech segmentation.

Keywords Speech segmentation · Diphone frequency · Database

In order to recognize spoken words, listeners must match the sound patterns of speech to specific lexical representations. In contrast to written language, in which words are separated by blank spaces, there are no clear word boundaries in spoken language. This means that a given stretch of speech can be consistent with multiple lexical hypotheses, and that these hypotheses can begin at different points in the input. The most common view of spoken-word recognition is that the listener considers the lexical candidates that are consistent with the perceptual input in parallel; in models such as TRACE (McClelland & Elman, 1986) and Shortlist (Norris, 1994), for example, candidate words are activated simultaneously and compete with one another. Processing the speech stream can therefore give rise to transient ambiguities. In the French sequence *l'abricot* [labRiko] "the apricot," segmental information could be compatible with several competing hypotheses, including *l'abri* [labRi] "the shelter," *la brique* [labRik] "the brick," *la brioche* [labRijɔʃ] "the brioche," and so forth. Listeners are routinely confronted with such transient segmentation ambiguities. However, despite the continuity of the speech signal, listeners manage to isolate words in the speech chain and are rarely misled.

A vast body of evidence now shows that listeners use their tacit knowledge of a wide range of patterns in their native language to help them segment speech, including cues from allophonic variation, phonotactic constraints, transitional probabilities, and lexical stress (Cutler & Norris, 1988; McQueen, 1998; Quené, 1992; Saffran, Aslin, & Newport, 1996, inter alia). The present article will focus on one particular cue: that provided by phonotactics.

Languages build their words from a finite set of phonemic units. Moreover, phonological rules guide the ways in which these phonemes can be arranged to form syllables. These rules, called *phonotactic constraint*s, define what sound combinations may and may not occur in a language. They are language-specific: For example, in Spanish, consonant clusters like /st/ are not allowed at the beginning of words, although they are in French (e.g., *structure* "structure" in French, but *estructure* in Spanish). Listeners become sensitive to these phonological regularities during the first year of exposure to their native language. Jusczyk, Friederici, Wessels, Svenkerud, and Jusczyk (1993) demonstrated that 9-month-old infants have knowledge of their native language phonotactic structure. For this study, a Dutch/English bilingual speaker pronounced lists of isolated low-frequency English words that are phonotactically ill-formed in Dutch and

B. New (✉)
Laboratoire Vision Action Cognition, Université Paris Descartes,
Paris Sorbonne Cité, France
e-mail: boris.new@parisdescartes.fr

E. Spinelli
Laboratoire de Psychologie et NeuroCognition,
Université Pierre Mendès France,
Grenoble, France

B. New · E. Spinelli
Institut Universitaire de France,
Paris, France

lists of isolated low-frequency Dutch words that are phonotactically ill-formed in English. The American 9-month-old infants showed a preference for the English words, whereas the Dutch infants showed a preference for the Dutch words.

Later, adult listeners use their knowledge of phonotactic regularities to decode what they hear. For example, it has been shown that listeners change their phonemic percepts when confronted with phonotactically illegal sequences and tend to reinterpret the acoustic signal in a way that respects phonotactic constraints. Massaro and Cohen (1983) showed that listeners who were asked to identify an ambiguous phoneme taken from an /r/ – /l/ continuum tended to give an interpretation of this phoneme that preserved the phonotactic constraints of their language. Hence, when one edge of the continuum constituted an illegal word-initial sequence (e.g., ***TLI*** in English), whereas the other extremity gave rise to a phonotactically legal sequence (e.g., ***TRI***), listeners interpreted the ambiguous phoneme in favor of the legal sequence (hence, as being an /r/). Conversely, when the same phoneme was presented in a ***SLI*** (legal sequence) / ***SRI*** (illegal sequence) context, subjects identified the phoneme as an /l/. In a similar vein, Hallé, Segui, Frauenfelder, and Meunier (1998) presented nonce sequences beginning with either a legal (/tR/) or an illegal initial cluster (/tl/, /dl/) in French (e.g., *trabdo* or *tlabdo*, *dlabdo*, respectively). In two tasks (phonemic transcription and forced choice), they observed that coronal stops in illegal sequences (/t/, /d/ in *tlabdo*, *dlabdo*) were perceived as being velar stops (/k/ and /g/), although no velar information was present in the signal. This was shown using a gating paradigm in which sequences were presented incrementally. During the first gates, there was no information about the following consonant, and subjects reported coronal responses (/t/, /d/). Later (i.e., when information about the following consonant was available), velar responses (/k/, /g/) replaced coronal ones, suggesting that the perception was modified in favor of a percept that was phonotactically legal. Segui, Frauenfelder, and Hallé (2001) showed similar results in a crossmodal repetition priming study. The initial velar phoneme /g/ of both *glaïeul* "gladiolus" and *groseille* "redcurrant" was replaced by a coronal /d/, hence giving rise to the pseudowords *dlaïeul* (phonotactically illegal) and *droseille* (phonotactically legal). The targets were primed by either their intact (/g/-initial) or altered (/d/-initial) form. In the phonotactically legal set (*groseille* changed to *droseille*), the altered primes produced less facilitation than did the intact ones, whereas in the phonotactically illegal set (*glaïeul* changed to *dlaïeul*), the altered and intact primes facilitated targets to the same extent. This suggests that for illegal primes, the phonemic percept was changed to a legal form of the word, hence allowing a strong repetition-priming effect. Spinelli and Gros-Balthazard (2007) examined the role of phonotactic constraints on the processing of schwa deletion in French. In their study, visual targets (e.g., *renard*, "fox") were auditorily primed by either an intact ([ləRənaR] "the fox") or reduced ([ləRnaR] "the fox") form of the word. When this schwa deletion respected the phonotactic constraints of French (e.g., [lapluz] "the lawn," where /pl/ is a legal word-beginning in French), a processing cost emerged for the targets primed by a reduced form of the word, as compared to intact primes (e.g., [lapəluz] "the lawn"). However, when schwa deletion induced a violation of phonotactic constraints (e.g., [ləRnaR], where /Rn/ is not allowed as a word beginning in French; Dell, 1995), no penalty was found for the targets primed by reduced as compared to intact forms of the word. The authors suggested that phonotactic constraints could help to overcome the penalty caused by schwa deletion in French by restoring the deleted schwa. Taken together, all of these studies suggest that listeners are capable of using their knowledge of the phonological regularities of their language to modify the representations that can be constructed from an acoustical analysis of the speech signal (see also Dupoux, Kakehi, Hirose, Pallier, & Mehler, 1999).

Moreover, adult listeners also make use of their knowledge of phonotactic regularities to segment the speech stream. Since it is often assumed that word boundaries tend to coincide with syllable boundaries, syllable onsets have been proposed as the locations where word boundaries are more likely to occur (Content, Kearns, & Frauenfelder, 2001; Content, Meunier, Kearns, & Frauenfelder, 2001; Cutler & Norris, 1988; Norris, McQueen, Cutler, & Butterfield, 1997; Vroomen & de Gelder, 1997). For example, Vroomen and de Gelder showed, in a cross-modal semantic-priming experiment in Dutch, that *boos* (angry) is activated in *framboos* (raspberry) but that *wijn* (wine) is not activated in *zwijn* (swine). In the latter case, the embedded word *wijn* is misaligned with the beginning of the syllable. These results suggest that embedded words are only activated strongly if their onsets match syllable onsets (but see Spinelli, McQueen, & Cutler, 2003).

Because phonotactic constraints often impose syllabic boundaries, the presence of such specific markers could aid listeners by providing cues for segmentation. A number of studies have indeed shown such a role of phonotactic constraints in speech segmentation. With a word-spotting task, McQueen (1998) showed that detecting a word (e.g., *rok*, "skirt" in Dutch) embedded in a nonword was easier when the word was aligned with a phonotactic boundary (e.g., "fim.rok," in which the syllable boundary is imposed by the phonotactic constraint that /mr/ is an illegal consonant cluster in Dutch) than when it was misaligned with the syllable boundary (e.g., "fi.drok" in which /dr/ is a legal cluster in Dutch). Similarly, /nl/ is an illegal consonant cluster in French. Dumay, Frauenfelder, and Content (2002) showed that detecting *lac* ("lake" in French) is easier in the nonword *zunlac* (syllabified "zun.lac," due to phonotactic constraints) than in the nonword *zuglac* ("zu.glac"). In

English, Weber (2001) showed that detecting *luck* is easier in the nonword *poonluck* ("poon.luck," since /nl/ is an illegal cluster in English) than in the nonword *marfluck* ("mar.fluck"). These studies all suggest that phonotactic constraints provide boundary cues in speech segmentation that are used for word recognition (e.g., of the targets *rok*, *lac*, and *luck*).

Interestingly, if phonotactic constraints often impose a syllabic boundary, they do not always impose lexical boundaries. There has been a recent attempt to distinguish between lexical and syllable boundary cues in speech segmentation in Italian. With a word-spotting task, Tagliapietra, Fanari, De Candia, and Tabossi (2009) tested listeners' sensitivity to phonotactic cues that specifically signaled lexical boundaries, and not just syllabic boundaries. Italian listeners had to detect a word (e.g., *mela* "apple") in either the nonword *ban.mela*, containing the diphone /nm/ at the target boundary, or in *bas.mela*, containing the diphone /sm/ at the target boundary. In Italian, both /sm/ and /nm/ impose a syllable boundary, but /nm/ also imposes a lexical boundary, in the sense that it is illegal in the medial position and can only occur as the last phoneme of one word and the first phoneme of the following word. In fact, Italian listeners were found to be insensitive to these lexical boundary cues, although they exhibited the classical "legality" effect found in other languages (i.e., they detected *lago* faster in *rin.lago* than in *ri.blago*). Note that Italian listeners do not seem to be sensitive to allophones that distinguish syllables (*allargo*) from words (*al largo*), either. This kind of segmentation study should be done in French, another Romance language in which listeners are sensitive to syllabic structure, but also to these kinds of allophonic cues (*l'affiche* vs. *la fiche*; Spinelli, Grimault, Meunier, & Welby, 2010; Spinelli, Welby, & Schaegis, 2007).

More interestingly, it has been shown that listeners are not only sensitive to the phonotactic well-formedness of sequences in their native language, but also to sequential probabilities. This sensitivity to sequential probabilities has been shown with 9-month-old infants (Jusczyk, Luce, & Charles-Luce, 1994; Mattys, Jusczyk, Luce, & Morgan, 1999) and adults (Vitevitch & Luce, 1999). It has also been shown that adult listeners can use the transitional probabilities in the segmentation of an artificial language (Saffran et al., 1996). As regards segmentation of natural language, Van der Lugt (2001) found, in a series of word-spotting experiments, that *kap* was easier to segment in *kap.juif* than was *heup* in *heup.juif*, because the diphone /ap/ is a frequent word-final sequence, whereas [øp] is not. Similarly, he found that *galg* was easier to segment in *pien.galg* than was *geur* in *pien.geur*, because the diphone [xa] is a frequent word-initial sequence, whereas [wø:] is not.

Moreover, such "diphone-based segmentation" has proved valid and learnable in a computational Bayesian model recently developed by Daland and Pierrehumbert (2011). They assessed whether the model could recover word boundaries based on the identity of the surrounding diphone. Their three simulations showed that the diphone-based segmentation (DiBS) model could reach a ceiling of performance after only a small amount of language exposure and that segmentation performance was robust to pronunciation variation. Moreover, the model does not oversegment the speech stream (less than one oversegmentation error per ten words). Diphone-based segmentation could thus be thought of as a reasonable strategy to acquire a lexicon (from a developmental perspective) and to segment speech (from an adult perspective).

To sum up, a growing body of evidence now shows that sequential probabilities in speech are strong cues to segmentation (Cairns, Shillcock, Chater, & Levy, 1997) and that the recognition system makes use of them. The aim of this article is to describe a new database of diphone positional frequencies in French. To our knowledge, no such tool has previously been available to psycholinguists for the study of French sequential probabilities.

Diphones-fr database

In order to compute diphone frequencies, we used the word frequencies identified in Lexique 3.80 (New, Pallier, Ferrand, & Matos, 2001) coming from a subtitle corpus of 50 million words (New, Brysbaert, Veronis, & Pallier, New et al. 2007). The latter corpus essentially consists of dialogue coming from movies or TV series. These dialogues could come from four subcorpora: French films, English films, English television, and non-English films. On the basis of extensive testing, it seemed to us that the best frequency measure to derive from the subtitle corpus was one in which we gave equal weights to each of the four subcorpora. In this way, the frequency estimates were based on the largest possible corpus, and we avoided the estimates being overly dependent on (American) movies. Therefore, we first calculated the frequencies per million words for the French films, the English films, the English television series, and the non-English films. Then, the average was taken of these four measures. Finally, a great advantage of this corpus is that it consists of more than 16 million words, which is the benchmark size that we have recommended as being very important for calculating reliable frequencies of low-frequency words (Brysbaert & New, 2009). We chose the subtitle corpus for three reasons. First, we found in previous studies that it was the best measure for predicting reaction times in a French lexical decision task (New et al., 2007). Second, this corpus consists of spoken interactions between people; indeed, a big part of this corpus looks like spoken interactions in real life. Third, for many people, television and TV series are an important source of spoken language input.

Before starting to compute the diphone statistics, we corrected more than 10,000 phonological forms from Lexique 3.50, which will shortly give rise to a new of version of Lexique: Lexique 3.80. Then we computed, for the 108,803 different words of our 50-million-word corpus, each diphone frequency. This diphone frequency was computed according to its position. For instance, a word such as *casque* /kask/ "helmet" is made of the following diphones: /ka/, /as/, and /sk/. The diphone /ka/ here would be counted as a diphone in initial position, while /sk/ would be counted as a diphone in final position. We also calculated type (number of words having this diphone) and token (number of occurrences of this diphone per million words) frequencies. We computed a diphone's frequency within words and also for continuous speech (from our subtitle corpus)—that is, between words (interword frequencies). We were able to compute interword frequencies because we have the entire corpus from which the lexical frequencies from Lexique 3.80 were computed. One potential problem in continuous French speech is that phonological variations can occur. For example, schwas [ə] are sometimes deleted in French casual speech (e.g., both *la pelouse* [lapeluz] and *la p'louse* [lapluz], "the lawn," are attested). We did not take into account potential schwa deletion—that is, we considered that when a schwa could be produced, it was counted. Another phonological variation that occurs in French is "liaison," which constitutes an extremely frequent phonological alternation in French. This involves the production of a silent (latent) consonant (/n/, /z/, or /t/ in 99.7 % of cases; Boë & Tubach, 1992) between two words (word1 and word2). For example, the liaison /z/ appears in *deux ours* [døzurs] "two bears" between *deux* and *ours*. For the liaison to appear, word2 must begin with a vowel (e.g., [urs] "bear"). The phonetic nature of the liaison depends on word1: /n/ after *un* "a/one" or *aucun* "no/not any"; /z/ after *les* "the + plural" or *deux* "two"; and /t/ after *petit* "little" or *grand* "big/tall." As a consequence, the latent consonants that are normally realized orally do not appear in the subtitles. We therefore provided two different versions of the interword frequency data: one without any coded liaison, and one for which liaisons had been inferred on the basis of our knowledge of the linguistic functioning of liaison (phonetic and semantic constraints). We applied the following rules in order to infer a liaison in French:

1. Infer a liaison if an adjective ends with one of the written consonants *s*, *x*, *d*, *t*, and *n* and is followed by a noun that begins with a vowel or with "h" (except for the aspirated ones).
2. Infer a liaison between *un*, *mon*, *ton*, *son*, *aucun*, *des*, *les*, *ces*, *mes*, *tes*, *ses*, *nos*, *vos*, *leurs*, *aux*, *quels*, *quelles*, *quelques*, *deux*, *trois*, *tout*, *quant*, *nous*, *vous*, *ils*, *elles*, *on*, *suis*, *es*, *êtes*, *est*, *après*, *plus*, *très*, *dans*, *chez*, or *sans* and a word that begins with a vowel or with "h" (except for the aspirated ones).
3. Infer a liaison between *quand* and *est*.

General diphone statistics The following overall statistics were calculated for each diphone in the database:

FreqToutTyp: This is the general diphone type frequency, without any position information.

FreqToutTypDeb: Diphone frequency when the diphone's first phoneme is also the word's first phoneme. For instance, in *amiante* /a-mj@t/, /am/ is counted (see Table 1 for the phonemic transcriptions used in the database).

FreqToutTypMil: Diphone frequency when the diphone is not at the beginning or at the end of the word. For instance, in *amiante* /a-mj@t/, /mj/ is counted.

FreqToutTypFin: Diphone frequency when the diphone is at the end of the word. For instance, *amiante* /a-mj@t/, /@t/ is counted.

Between-syllable positional statistics We also computed each diphone's statistics between syllables. The general idea is the following: Freq(am) is the diphone frequency when /a/ and /m/ belong to two different syllables (as in *amener* /a-m2-ne/).

FreqInterTyp: This is the general between-syllable statistic, without any position information. For instance, in *cascade* /kas-kad/, /sk/ is counted.

FreqInterTypDeb: Diphone frequency when the diphone's first phoneme is also the word's first phoneme. For instance, in *amiante* /a-mj@t/, /am/ is counted.

FreqInterTypFin: Diphone frequency when the diphone's last phoneme is also the word's last phoneme. For instance, in *laureat* /lO-Re-a/, /ea/ is counted.

Table 1 Phonetic transcriptions in the database

[b] = b	[f] = f	[ʁ] = R	[ɥ] = 8	[i] = i
[d] = d	[v] = v	[l] = l	[a] = a	[u] = u
[g] = g	[z] = z	[m] = m	[o] = o	[y] = y
[p] = p	[ʒ] = Z	[n] = n	[ɔ] = O	[ø] = 2
[t] = t	[s] = s	[j] = j	[e] = e	[œ] = 9
[k] = k	[ʃ] = S	[w] = w	[ɛ] = E	[ə] = °
[ɑ̃] = @	[ɔ̃] = §	[ɛ̃] = 5	[ŋ] = G	[ɳ] = N
[œ̃] = 1				

FreqInterTypMil: Between-syllable diphone frequency when the diphone is not counted in FreqInterTypDeb or FreqInterTypFin. For instance, in *laureat* /lO-Re-a/, /OR/ is counted.

Within-syllable positional statistics We used the same principles for a diphone's statistics within a given syllable:

IntraMotDeb: Diphone frequency when the diphone's first phoneme is also the word's first phoneme. For instance, in *laureat* /lO-Re-a/, /lO/ is counted.

IntraMotFin: Diphone frequency when the diphone's last phoneme is also the word's last phoneme. For instance, in *amiante* / a-mj@t/, /@t/ is counted.

IntraMotMil: Within-syllable diphone frequency when the diphone is not counted in IntraMotDeb or IntraMotFin. For instance, in *laureat* /lO-Re-a/, /Re/ is counted.

It is still debated whether syllable onset cues to segmentation could be different from word onset cues to segmentation. Hence, as the question remains, we also give these diphone statistics applied to the syllable unit: For instance, in *laureat* /lO-Re-a/, the diphone /Re/ occurs in the middle of the word (*IntraMotMil*) but is also syllable initial (*IntraSylDeb, see below).*

IntraSylDeb: Diphone frequency when the diphone's first phoneme is also a syllable's first phoneme. For instance, in *amiante* /a-mj@t/, /mj/ is counted.

IntraSylFin: Diphone frequency when the diphone's last phoneme is also the syllable's last phoneme. For instance, in *amiante* / a-mj@t/, /@t/ is counted.

IntraSylMil: Within-syllable diphone frequency when the diphone isn't counted in IntraSylDeb or IntraSylFin. For instance, in *amiante* /a-mj@t /, /j@/ is counted.

Positional phoneme frequencies and transitional probabilities We also computed the phoneme frequency depending on position (initial, medial, or final). Finally, we computed the transitional probability for each diphone, which is the diphone frequency divided by the initial phoneme frequency. This computation was made both depending on the position (initial, medial, or final) and not depending on the position.

Between-word statistics We computed each diphone's statistics between words. The general idea was the following: Freq(am) is the diphone frequency when /a/ and /m/ belong to two different words (as in "fer**a m**on"). We give both the type (number of two-word combinations having this diphone) and token (number of occurrence of this interword diphone per million words) frequencies.

FreqTout: This is the general between-word statistics.

FreqQdLiaison: This is the between-word statistic after the phenomenon of liaison is taken into account, when it is likely to occur. For instance, in *deux amis*, /za/ is counted.

FreqSansLiaison: This is the between-word statistic when the phenomenon of liaison is not taken into account. For instance, in *deux amis*, /2a/ is counted.

Comparison with Tubach and Boë (1990)'s computation

The database provided here is not the first attempt at counting French phonemes and phoneme combinations. Tubach and Boë (1990) published statistics on phoneme, diphone, and triphone frequencies taken from a corpus of 300,000 phonemes arising from several conferences and conversations. There are several differences between the two databases: First of all, the Tubach and Boë database is based on a spoken corpus. The advantage of this approach is that it keeps track of errors or hesitations. In return, they do not make a distinction between diphones that come from only one word and diphones from the boundaries between two different words (e.g., the diphone /ta/ in *petitavion* comes from two different words *petit* and *avion*). Second, there is a difference in corpus size: Our database is based on a corpus of 50 million words, while Tubach and Boë's corpus is based on 300,000 phonemes, or around 100,000 words, if we consider that a French word's average length is 2.9 phonemes. Recently, Brysbaert and New (2009) showed that corpus size is particularly important when estimating word frequency. For instance, they recommend a corpus size of at least 16 million words for estimating the word frequencies of rare words. Of course, this has to be put in perspective, as a diphone unit is much smaller than the word unit. Finally, the two corpora differ in their language registers. Our corpus consists mainly of dialogue, while the Tubach and Boë corpus is composed mainly of conferences and conversations on various subjects among members of the French intelligentsia. Despite these differences, we observed a significant correlation between the 50 most frequent diphones in our corpus and in Tubach and Boë's (r = .47, p < .001). A scattergram plotting the correlations between the diphone frequencies in our database and in Tubach and Boë's is presented in Fig. 1.

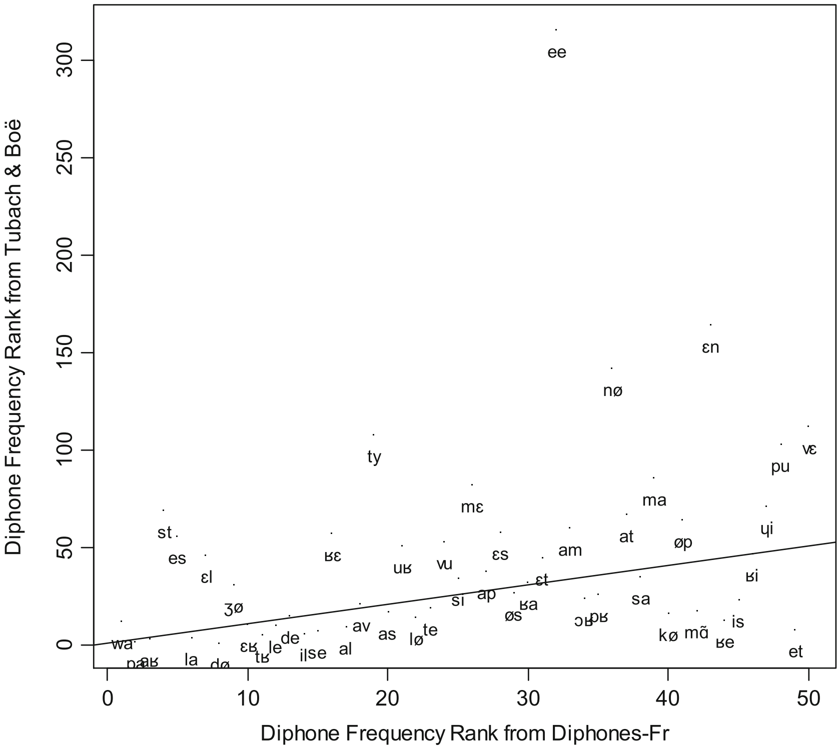

Fig. 1 Diphone frequency ranks in Tubach and Boë's (1990) database and in Diphones-Fr

This figure shows several outlier diphones, such as /ty/ ([ty]), /nø/ ([nə]), /En/ ([εn], or /ee/ ([ee]), that are more frequent in our database than in Tubach and Boë (1990). This could be due to the different language registers of the two corpora. Hence, *tu* (/ty/ "you"), *ne* (/nə/ "not"), *semaine* (/səmεn/ "week"), and *et* (/e/ "and") or *est* (/e/ "is") are frequently used in common dialogue. Another difference possibly comes from different codings, such as the closed or open e/ε.

From a descriptive point of view, there is also a crucial difference between our database and that of Tubach and Boë (1990): We provide many different frequencies. For instance, we give positional frequencies, both within and between syllables and within and between words, but also type and token frequencies.

Availability of the database

The database, as well as the Perl scripts and the subtitle corpus used to compute the diphone frequencies, can be downloaded as supplemental information from the Lexique website (www.lexique.org/projets/Diphones).

Conclusion

A cluster that is illegal when beginning a word, such as /gm/ or /Rb/, may not necessarily be illegal when ending a word (e.g., *he**rb**e* "grass" [εRb]) or when used within a word (*ma**gm**a*). Phonotactic constraints (defined in terms of "legality") would then help speech segmentation in sequences such as *gan**g** **m**ystérieux* ([gãgmisteRjø] "mysterious gang") or *ca**r** **b**ondé* ([KaRbɔ̃de] "crowded bus"), but would mislead segmentation in cases like *ma**gm**a* or *he**rb**e*. It is thus highly probable that listeners exploit more subtle regularities involving the positional frequencies of phoneme sequences. The aim of the present article was to provide statistical data on diphone positional frequencies in French.

Exploring gradient effects of diphone frequency, and not just "legality effects," will be made possible thanks to the database that we provide. We are confident that this new tool will allow researchers to explore more refined cues in French speech segmentation. First of all, the database will provide a way to find experimental stimuli to test whether listeners are sensitive to diphone frequency, and not only to diphone legality in speech segmentation. Second, the database will allow testing of the role of a diphone's positional frequency (e.g., is word-initial diphone frequency a stronger cue to segmentation than word-final diphone frequency?). Finally, this database will allow testing of the relative contributions of lexical boundary as opposed to syllable boundary cues. This is an important debate, especially for Romance languages such as French, in that there is abundant evidence that syllable boundaries play a role in the segmentation of spoken French (Content, Kearns, & Frauenfelder, 2001; Content, Meunier, et al., 2001; Cutler, McQueen, Norris, & Somejuan, 2001; Cutler, Mehler, Norris, & Segui, 1986;

Kolinsky, Morais, & Cluytens, 1995; Mehler, Dommergues, Frauenfelder, & Segui, 1981).

In conclusion, we have provided a database of diphone positional frequencies in French. This database and its new indicators should help researchers to conduct new studies on segmentation.

References

Boë, L. J., & Tubach, J. P. (1992). *De A à Zut: Dictionnaire du français parlé*. Grenoble, France: ELLUG.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41,* 977–990. doi:10.3758/BRM.41.4.977

Cairns, P., Shillcock, R. C., Chater, N., & Levy, J. (1997). Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology, 33,* 111–153.

Content, A., Kearns, R. K., & Frauenfelder, U. H. (2001a). Boundaries versus onsets in syllabic segmentation. *Journal of Memory and Language, 45,* 177–199.

Content, A., Meunier, C., Kearns, R. K., & Frauenfelder, U. H. (2001b). Sequence detection in pseudowords in French: Where is the syllable effect? *Language and Cognitive Processes, 16,* 609–636.

Cutler, A., McQueen, J. M., Norris, D., & Somejuan, A. (2001). The roll of the silly ball. In E. Dupoux (Ed.), *Language, brain, and cognitive development: Essays in honor of Jacques Mehler* (pp. 181–194). Cambridge, MA: MIT Press.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language, 25,* 385–400.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 113–121. doi:10.1037/0096-1523.14.1.113

Daland, R., & Pierrehumbert, J. B. (2011). Learnability of diphone-based segmentation. *Cognitive Science, 35,* 119–155.

Dell, F. (1995). Consonant clusters and phonological syllables in French. *Lingua, 95,* 5–26.

Dumay, N., Frauenfelder, U. H., & Content, A. (2002). The role of the syllable in lexical segmentation in French: Word-spotting data. *Brain and Language, 81,* 144–161.

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance, 25,* 1568–1578.

Hallé, P., Segui, J., Frauenfelder, U., & Meunier, C. (1998). The processing of illegal consonant clusters: A case of perceptual assimilation? *Journal of Experimental Psychology: Human Perception and Performance, 24,* 592–608.

Jusczyk, P. W., Friederici, A. D., Wessels, J., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language, 32,* 402–420.

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language, 33,* 630–645.

Kolinsky, R., Morais, J., & Cluytens, M. (1995). Intermediate representations in spoken word recognition: Evidence from word illusions. *Journal of Memory and Language, 34,* 19–40.

Massaro, D. W., & Cohen, M. M. (1983). Phonological constraints in speech perception. *Perception & Psychophysics, 34,* 338–348.

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

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 1–86. doi:10.1016/0010-0285(86)90015-0

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

Mehler, J., Dommergues, J., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior, 20,* 298–305.

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics, 28,* 661–677.

New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur Internet: LEXIQUE. *L'Année Psychologique, 101,* 447–462.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 52,* 189–234. doi:10.1016/0010-0277(94)90043-4

Norris, D. G., McQueen, J. M., Cutler, A., & Butterfield, S. (1997). The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology, 34,* 191–243.

Quené, H. (1992). Durational cues for word segmentation in Dutch. *Journal of Phonetics, 20,* 331–350.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274,* 1926–1928. doi:10.1126/science.274.5294.1926

Segui, J., Frauenfelder, U., & Hallé, P. (2001). Phonotactic constraints shape speech perception: Implications for sublexical and lexical processing. In E. Dupoux (Ed.), *Language, brain, and cognitive development: Essays in Honor of Jacques Mehler* (pp. 195–208). Cambridge, MA: MIT Press.

Spinelli, E., Grimault, N., Meunier, F., & Welby, P. (2010). An intonational cue to segmentation in phonemically identical sequences. *Attention, Perception, & Psychophysics, 72,* 775–787. doi:10.3758/APP.72.3.775

Spinelli, E., & Gros-Balthazard, F. (2007). Phonotactics constraints help to overcome effects of schwa deletion in French. *Cognition, 104,* 397–406. doi:10.1016/j.cognition.2006.07.002

Spinelli, E., McQueen, J., & Cutler, A. (2003). Processing resyllabified words in French. *Journal of Memory and Language, 48,* 233–254.

Spinelli, E., Welby, P., & Schaegis, A. L. (2007). Fine-grained access to targets and competitors in phonemically ambiguous spoken sequences: The case of French elision. *Language and Cognitive Processes, 22,* 828–859.

Tagliapietra, L., Fanari, R., De Candia, C., & Tabossi, P. (2009). Phonotactic regularities in the segmentation of spoken Italian. *Quarterly Journal of Experimental Psychology, 62,* 392–415.

Tubach, J. P., & Boë, L. J. (1990). *Un corpus de transcription phonétique*. Paris, France: Telecom.

Van der Lugt, A. (2001). The use of sequential probabilities in the segmentation of speech. *Perception & Psychophysics, 63,* 811–823.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language, 40,* 374–408.

Vroomen, J., & de Gelder, B. (1997). Activation of embedded words in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 710–720. doi:10.1037/0096-1523.23.3.710

Weber, A. C. (2001). *Language-specific listening: The case of phonetic sequences (MPI Series in Psycholinguistics* (Vol. 16). Nijmegen, The Netherlands: University of Nijmegen.