Proceedings of DiSS-LPSS Joint Workshop 2010,
25-26 September 2010, Tokyo, Japan, pp. 43-46.

# Phonological competition in casual speech

*Anne Cutler* [1,2], *Holger Mitterer*[1], *Susanne Brouwer*[1], *Annelie Tuinman*[1]

[1] Max Planck Institute for Psycholinguistics, Nijmegen 6500 AH, The Netherlands
[2] MARCS Auditory Laboratories, University of Western Sydney, NSW 1797, Australia
anne.cutler/holger.mitterer/susanne.brouwer/annelie.tuinman@mpi.nl

## Abstract

The natural processes affecting spontaneous speech production and the natural processes of spoken-word recognition combine to cause significant activation of irrelevant lexical competitors. Using eye-tracking, we show that reduced forms of words that occur in casual speech cause listeners to activate lexical candidates that resemble the reduced form but are quite unlike the canonical form of the intended word. In L2, the problem is worse: casual speech processes that occur in the L2 but not in the L1 lead to activation of irrelevant competitors even where native listeners experience no such competition.
**Index Terms**: word recognition, competition, eyetracking

## 1. Introduction

Spontaneous speech is what, as language users, we mostly hear and produce. However it is not what psycholinguists have mostly in the past conducted experiments on. There is now a detailed psycholinguistic picture of the architecture of word recognition (see [12] for a review). Speech input is processed in a continuous manner, whereby multiple interpretations of the input are concurrently considered and evaluated, with the set of potential candidate words being constantly adjusted as further input arrives. This activation/competition picture of word recognition can be found in all current models of the spoken-word recognition process, no matter how they differ in other architectural aspects.

The word competition refers to the fact that the more candidate words are available for a particular portion of the speech input, the harder it can be to recognise a word [8, 13, 16]. In effect, each word actively competes against the rival candidates for its portion of the speech input. The process of competition assists listeners to arrive at the correct parse for a speech sequence of which parts are potentially ambiguous; thus it plays an important role in segmenting speech. Studies of speech segmentation have manipulated the number of other words potentially activated by the context adjacent to a word; such manipulations indeed affect recognition. For instance, the recognition of *mint* could be compared in *mintayf* (which can call up dozens of words beginning *ta-* such as *table, tail, take* and many more) versus *mintowf* (where *tow-* calls up only four word clusters: *town, towel, tower, tout*). In such a case, the more words the context activates to compete for the final part of a target word, the harder that target is to recognise, both in English [16] and Dutch [27], while the more words the context activates which do not compete for part of the word (such as words beginning *ayf* or *owf*, to re-use this example), the easier it is to segment the target from the context [26].

Models of speech perception are understandably based on a somewhat idealised situation. The mapping of a phoneme or sequence of phonemes to stored word representations can be predicted very well by the perceptual models, but the modelled situation will only arise if the input actually presents an acoustic form corresponding to each proposed segment. As listeners and speech researchers know only too well, however, real speech abounds with casual speech processes such as assimilation, reduction, deletion and intrusion, all of which lead to the realisation of phonetic forms which deviate drastically from the canonical pronunciation of the words intended by the speaker.

In recent years, psycholinguistics has increasingly turned to the study of real speech, and how listeners deal with the non-canonical forms it presents. A simplified summary of the accrued results to date is that listeners are extremely good at exploiting the fine phonetic detail of utterances and identifying intended words even when casual speech processes have altered them from their canonical form, but that the alterations can often (temporarily) mislead listeners, and can often result in word recognition being harder than it would have been for the canonically pronounced versions. The fine differences between intended phonemes and phonemes resulting from a casual speech process have been shown to be exploited by listeners, for example in the case of place of articulation assimilation (e.g., to distinguish the /p/ of English *ripe* in *ripe berries* from the assimilated final phoneme of *right berries*; [7]), in neutralisation (e.g., to distinguish the final /p/ of Dutch *slip* from the devoiced final sound of *slib*; [28]), and in liaison (e.g., to distinguish the word-initial /p/ in French *trop partisan* from the liaison realisation of a word-final /p/ in *trop artisan*; [20]). Listeners are successful at identifying word forms despite assimilation of place or voice [6, 15, 19] and despite reduction [5] or other non-canonical realisation [21].

Despite all this success at dealing with real-speech forms, however, listeners are also often misled. Word recognition response times are slowed by many different types of casual-speech forms [11, 18]. In a phoneme-detection task listeners respond to phonemes that have actually been deleted in a casual pronunciation [10], and they respond to phonemes that are accidentally there, such as /p/ in a casual version of *something* [29].

The above research has been carried out with first-language (L1) listeners. Clearly, an even worse situation may present itself to second-language (L2) listeners. Considerable activation of spurious competitors occurs in L2 listening even with clear "laboratory" speech input [1, 4, 30], and if the phonemes of the L2 cannot be reliably discriminated, then such spurious competition can be particularly hard to get rid of [2]. Although it is known that L2 listeners have difficulty dealing with the kinds of casual-speech phenomena referred to above, little research has addressed the effect of this for word recognition in L2, and in particular for lexical competition.

In the present study we consider the implications of such processes for lexical competition, both in L1 and L2. The evidence we present comes from the eye-tracking paradigm, which is particularly suited to assessing the online availability of activated and competing candidate words during the recognition of spoken language.

## 2. Casual speech and native listeners

The eye-tracking paradigm records where listeners are looking as they hear speech. Typically, a visual display presents a small set of pictures [22] or printed words [14], and listeners are instructed to click on any component of the display that is mentioned in what they hear. For instance, if pictures of candy and candles are in the display, and both of these are looked at as the input *Click on the cand-* is heard, we assume that both are being considered as lexical candidates given that input. Sometimes the display does not actually contain an exact match to anything in the input, and then the looks reveal what among the available options best matches the input [9].

Such experiments have typically been carried out with speech especially recorded for the occasion, but it is also possible to present naturally recorded speech samples. In our Experiment 1, we used real conversational speech extracted from the Corpus of Spoken Dutch [17]. The advantage of using specially recorded input is that it is possible to control the input and, for instance, present minimal pairs. This is not usually feasible with real speech, nor is it possible even with a very large natural corpus to collect a substantial set of critical tokens from a single speaker; but the advantage of using a spontaneous-speech corpus is that we can address in a natural manner the question of how listeners process real speech.

### 2.1. Experiment 1: Methods and Procedure

24 Dutch-native undergraduates from the Radboud University Nijmegen community took part in return for a small monetary compensation. None had visual or hearing problems.

The spoken stimuli included 32 canonical and 32 reduced forms of the same words, extracted from the speech of varying speakers. For example, the word *beneden* 'downwards' is pronounced [bənedə] in its canonical from, and a token was found of such a pronunciation. Another reduced form of the same word was found, in which the phonetic transcription [məneə] shows that the initial sound was assimilated to a nasal, and the onset of the third syllable was deleted. Each of the forms was presented in its full original sentential context.

Concurrent with the presentation of each spoken sentence, the computer screen displayed four printed words. On the experimental trials the display did not include the target word; instead, it included (1) a "canonical form" competitor which overlapped phonologically at onset more with the canonical form than with the reduced form of the spoken word (e.g., for *beneden*, this was *benadelen* 'disadvantage'); (2) a "reduced form" competitor which overlapped phonologically at onset more with the reduced form than with the canonical form of the spoken word (e.g., *meneer* 'mister'); and (3, 4) two phonologically unrelated distractors (e.g., *vakantie* 'holiday'; *juweel* 'jewel'; see Fig. 1).The experiment began with several practice trials, and also included filler items in which one word on the screen matched to a word in the spoken sentence.

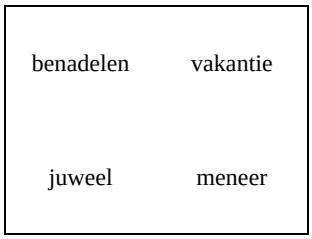


Figure 1: *Example of a visual display (spoken form in the input* beneden*).*

Participants heard the sentences over headphones, and were instructed to click on the visual word matching a word in the input, if it was present (filler trials), and to click in the middle of the screen if none of the words on the visual display matched a word in the spoken sentence (experimental trials). A SMI EyeLink eye-tracking system recorded gaze direction across time, sampling at 250 Hz. We report data from the time period 400-800 ms from onset of the word form in the spoken input (note that there is a lag of about 200 ms between an eye movement's onset and its terminus in a fixation).

### 2.2. Experiment 1: Results and Discussion

As Figure 2 shows, when the spoken sentence contained a word in its canonical pronunciation (e.g., *beneden*), listeners directed significantly more looks to the competitor that began in the same way (Ccomp: *benadelen*) than to the other words on the screen. However, when the input was the reduced form of a word, most looks went to the competitor word which began in the same way as the reduced form (Rcomp: *meneer*). Statistical analysis with linear mixed-effects models revealed that overall, competitors attracted more looks than distractors ($p < .001$), and that when the input contained a canonical pronunciation, the Ccomp attracted more looks than the Rcomp ($p < .001$). This was not the case when a reduced form occurred in the input, although the advantage here for Rcomp over Ccomp did not reach our statistical significance criterion.

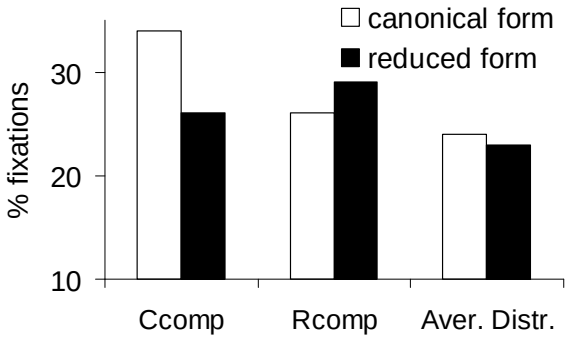


Figure 2: *Fixations by Dutch listeners to the canonical form competitor (Ccomp), reduced form competitor (Rcomp) and the two distractors (averaged) in the time frame 400 to 800 ms from onset of the Dutch spoken form (canonical vs. reduced), Experiment 1.*

The eye movements that this experiment has captured clearly show that natural spontaneous speech induces consideration of lexical competitors that do not sound very much at all like the canonical form of the words in the speech signal, but instead are canonical forms of other words which the reductions occurring in spontaneous speech accidentally resemble. The efficient mechanism for achieving spoken-word recognition via a process of concurrent activation of multiple competing words delivers these forms, quite reasonably, as potential matches to what is actually in the input.

Further exploration of such reductions [3] has revealed that there are in many cases subtle phonetic differences between a given phoneme arising as a reduced form (e.g., the /m/ in the reduced form of *beneden*) and the same phoneme when produced as part of an intended canonical form; listeners are sensitive to these differences and can recover rapidly from the spurious competition. The early eye movement data attests nonetheless that such spurious competition occurs.

## 3. Casual speech and nonnative listeners

Spurious phonological competition is a known problem for L2 listeners, especially when phonetic contrasts in the L2 require distinctions not made in the L1 (such as the /r/-/l/ distinction for many Asian listeners to English [4]). This increase in competition can occur even when the input presents clearly articulated canonical forms.

Casual speech processes in spontaneous speech, however, may add significant further problems for L2 listening. Although some casual speech processes are widespread, so that listeners who come across them in an L2 may already be familiar with them from the L1, others are relatively uncommon. If the L2 presents a type of casual speech transformation with which listeners have had no L1 experience, it is likely that they may be misled into activating irrelevant competitors even when no difficult phonemic distinctions are required.

To test this in our second eye-tracking experiment, we presented Dutch listeners from the same population with input in British English. Some sentences in the input contained an intrusive /r/ at an intervocalic word boundary. This casual speech process is typical of British English; in sequences such as *law and order* or *saw a film* an /r/ will be inserted at the transition between the first and second word. Although British English is the target pronunciation taught in Dutch schools, and British English is widely available in the media in the Netherlands, the process of /r/-insertion is, along with all other casual speech phenomena, not explicitly taught, and the process is completely absent from Dutch. It is therefore a process which appears in the L2 but is unfamiliar from the L1.

### 3.1. Experiment 2: Methods and Procedure

24 Dutch-native undergraduates from the Radboud University Nijmegen community took part in return for a small monetary compensation. None had visual or hearing problems and all had high proficiency in English comprehension.

The spoken stimuli included 27 sentence pairs containing sequences such as *My brother likes extra ice...* or *My brother likes extra rice...* produced by a female native British English speaker who produces intrusive /r/ in sequences such as *extra ice*; each such sequence indeed contained intrusive /r/. There were again practice trials and 54 filler trials, some including sequences appropriate for linking /r/ (e.g., *your explanation*).

As in Experiment 1, the computer screen displayed four printed words concurrent with the presentation of each spoken sentence. In this experiment the target word was not absent from the experimental trials. The four words in the display were (1) the r-initial word from the sentence pair in that trial (e.g., *rice*), (2) the vowel-initial word (e.g., *ice*), and (3, 4) two phonologically unrelated distractors which also formed an r-initial/V-initial pair (e.g., *raid*; *aid*).

Except that participants were instructed to click on a word they heard in the sentence, the procedure was as in Experiment 1. Again we analysed looking patterns in the time period 400-800 ms from onset of the word form in the spoken input.

### 3.2. Experiment 2: Results and Discussion

Figure 3 shows the percentage of looks to the actually spoken form (Target: *rice* given *extra rice*, *ice* given *extra ice*), the competitor (Comp: *ice* given *extra rice*, *rice* given *extra ice*) and the two distractors in the crucial early looking phase. Again an asymmetry appears. When the target was indeed *rice* (white bars), listeners looked mostly at the word *rice*. When the target was *ice* (black bars), listeners mostly looked at *rice*.

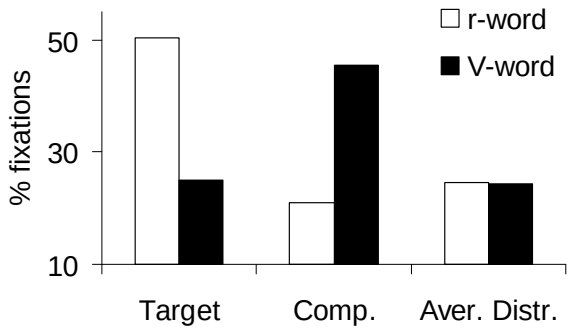


Figure 3: *Fixations by Dutch listeners to the actually spoken target, the competitor (Comp) and the two distractors (averaged) in the time frame 400 to 800 ms from onset of the English spoken form (r-initial word vs. vowel-initial word preceded by intrusive r), Experiment 2.*

Statistical analyses were conducted as for Experiment 1. The preference for looking at the r-initial word over the vowel-initial word was significant for both input types ($p < .001$ in each case).

In other words, the intrusive /r/ in a phrase such as *extra ice* made these listeners activate *rice* as a word candidate to a greater extent than the word that the speaker intended, *ice*. In further studies of this casual speech process, intrusive /r/ was found to be significantly shorter than intended onset /r/. Native speakers of British English were very sensitive to the duration of /r/ in judging its source in a phonetic task [24], and showed no activation of /r/-onset words in a priming task [25]. Dutch listeners, on the other hand, were relatively insensitive to duration in the phonetic task, tending instead to attribute the /r/ to a spelling source, or to base their decision on semantic context. The present findings from eye-tracking show the consequences of these L2 listeners' inability to deal with the /r/-insertion in the way native listeners do; their word recognition processes have to deal with unwanted competitor activation.

## 4. Conclusions

Activation of multiple candidate word forms is not under a listener's control; it is an automatic application of the highly efficient mechanisms underlying the rapid, robust and accurate speech recognition which is one of the greatest achievements of human cognition. Incoming speech signals are efficiently and continuously processed, and the potential words they might consist of are all activated and allowed to compete for the input; a parse that completely accounts for the input, and assigns each part of it to a separate word, should be the result.

Every act of speech recognition involves some spuriously activated lexical competitors, simply because in all languages, vocabularies running into hundreds of thousands of words are built from just a handful of phonemes (the mean number of phonemes across languages is around 30). This makes it unavoidable that most longer words contain shorter words accidentally embedded with in them (so, *accidental* contains *axe* and *dent* and *dental*), and these words will become active whenever their carrier words are heard, just as the carrier word will become activated if the corresponding shorter words are heard (*This board will never axe a dental program*). Analyses of (canonical-form) vocabularies (see [12]) reveal that by far the majority of all words contain some other word form. Despite this spurious activation, the competition process delivers highly efficient recognition.

Nonetheless, as described in the Introduction, more competition leads to measurable delays in spoken-word processing [16, 27]. It is interesting to contemplate how this fact is affected by the structure of spontaneous (in comparison to laboratory) speech. Although documenting this is a topic for future research, we suggest that there are probably going to be as many benefits as disadvantages. That is, as often as a reduced form of *beneden* calls up *meneer* or *extra ice* calls up *rice*, there will be cases where a competitor potentially active in a canonical pronunciation will be ruled out by a more casual pronunciation. Consider, for instance, that *fami-* could be the beginning of *family*, *famine*, or *famish*, but if, as so often in casual speech, *family* is pronounced *fam'ly*, then the other two candidates will no longer be active.

Demonstration of such potential benefits must await future investigations. For the present, we have shown that reduced forms in casual speech can definitely lead to momentary activation of unwanted competitors which are not strongly activated by a canonical pronunciation of the same word. Such activation can underlie the delays reported for recognition of reduced words in prior studies using techniques such as lexical decision [11, 18]. In the present research we have used eye-tracking to focus specifically on the early stages of recognition, before the competition process is resolved. Here the added competition due to the casual pronunciation can be clearly seen.

Spurious competition in L2 listening is a well-known problem [1], and it has previously been demonstrated in eye-tracking studies [4, 30]. However, the degree to which L2 listeners can cope with casual speech processes is as yet under-researched. It is reasonable to suppose that a casual speech process that occurs in much the same way in the L1 and in an L2 will be processed by L2 listeners with all the ease that they can bring from their L1 experience (and, indeed, there is recent evidence that this holds for the cross-linguistically common process of /t/-deletion [23]). An unfamiliar process, however, cannot be resolved with L1 resources. Just as L1-biased inability to distinguish phonetic contrasts in L2 leads to spurious competition that is unusually difficult to get rid of [2], so we suggest that the competition caused by an unfamiliar casual speech process such as /r/-insertion may prove quite intractable for L2 listeners. This too must be addressed in a future study. For now, our two sets of eye-tracking results have firmly demonstrated that for both L1 and L2 listeners, the processes that naturally occur in casual speech can induce substantial and potentially disadvantageous phonological competition.

## 5. Acknowledgements

We acknowledge support from Max Planck Society doctoral fellowships (SB, AT), the Deutsche Forschungsgemeinschaft (HM) and NWO-SPINOZA (AC).

## 6. References

[1] Broersma M., "Phonetic and lexical processing in a second language ", Nijmegen: Radboud Universiteit Nijmegen.

[2] Broersma, M. and Cutler A., "Competition dynamics of second-language listening", Quart. J. Exp. Psy., in press.

[3] Brouwer, S., Mitterer, H. and Huettig, F., "Tracking the timecourse of phonological competition during the processing of reduced speech", submitted.

[4] Cutler, A., Weber A. and Otake. T., "Asymmetric mapping from phonetic to lexical representations in second-language listening", J Phon, 34:268-284, 2006.

[5] Ernestus, M., Baayen, H. and Schreuder, R., "The recognition of reduced word forms", Brain Lang, 81:162-173, 2002.

[6] Gaskell M. G. and Marslen-Wilson, W. D., "Phonological variation and inference in lexical access", J. Exp. Psy.: Hum. Perc. Perf., 22:144-158, 1996.

[7] Gow D. W., "Does English coronal place assimilation create lexical ambiguity?", J. Exp. Psy.: Hum. Perc. Perf., 28:163-179, 2002.

[8] Goldinger, S. D., Luce, P. A. and Pisoni, D. B., "Priming lexical neighbors of spoken words: Effects of competition and inhibition", J. Mem. Lang., 28:501-518, 1989.

[9] Huettig, F. and Altmann, G. T. M., "Word meaning and the control of eye fixation: semantic competitor effects and the visual world paradigm", Cognition, 96:B23–B32, 2005.

[10] Kemps R. J., Ernestus, M., Schreuder, R. and Baayen, H., "Processing reduced word forms: The suffix restoration effect", Brain Lang, 90:117-127, 2004.

[11] LoCasto P. and Connine, C. M. "Rule-governed missing information in spoken word recognition: Schwa vowel deletion", Perc. Psychophys., 64:208-219, 2002.

[12] McQueen, J. M., "Eight questions about spoken-word recognition", in M. G. Gaskell (Ed), The Oxford Handbook of Psycholinguistics, 37-53, Oxford Univ. Press, 2007.

[13] McQueen, J. M., Norris, D. and Cutler, A., "Competition in spoken word recognition: Spotting words in other words", J. Exp. Psychol.: Learn. Mem. Cogn, 20:621-638, 1994.

[14] McQueen, J. M. and Viebahn M. C., "Tracking recognition of spoken words by tracking looks to printed words", 60:661-671, 2007.

[15] Mitterer, H., Csépe, V. and Blomert, L., "The role of perceptual integration in the perception of assimilation word forms", Quart. J. Exp. Psy., 59:1395-1424, 2006.

[16] Norris, D., McQueen, J. M. and Cutler, A., "Competition and segmentation in spoken word recognition", J. Exp. Psychol.: Learn. Mem. Cogn, 21:1209-1228, 1995.

[17] Oostdijk, N., "The Spoken Dutch Corpus Project", ELRA Newsletter, 5:4-8, 2000.

[18] Racine, I. and Grosjean, F., "Influence de l'effacement du schwa sur la reconnaissance des mots en parole continue", L'Année Psychologique, 100:393-417, 2000.

[19] Snoeren, N., Segui, J. and Hallé, P., "Perceptual processing of partially and fully assimilated words in French", J. Exp. Psy.: Hum. Perc. Perf., 34:193-204, 2008.

[20] Spinelli, E., McQueen, J. M. and Cutler, A., "Processing resyllabified words in French", J. Mem. Lang., 48:233-254, 2003.

[21] Sumner, M. and Samuel, A. G., "Perception and representation of regular variation: The case of final /t/", J. Mem. Lang., 52:322-338, 2005.

[22] Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. and Sedivy, J. C., "Integration of visual and linguistic information in spoken language comprehension", Science, 268:1632-1634, 1995.

[23] Tuinman. A. and Mitterer, H., "Transfer of L1 knowledge to L2 processing: The case of /t/-reduction", submitted.

[24] Tuinman, A., Mitterer, H. and Cutler, A., "Cross-language and cross-dialect differences in perception of intrusive /r/ in English", submitted.

[25] Tuinman, A., Mitterer, H. and Cutler, A., "Resolving ambiguity in familiar and unfamiliar casual speech", submitted.

[26] van der Lugt, A., "The use of sequential probabilities in the segmentation of speech", Percept Psychophys, 63:811-823, 2001.

[27] Vroomen, J. and de Gelder, B., "Metrical segmentation and lexical inhibition in spoken word recognition", J. Exp. Psychol.: Hum. Perc. Perf., 21:98-108, 1995.

[28] Warner, N., Jongman, A., Sereno, J. and Kemps, R. J., "Incomplete neutralization and other sub-phonemic durational differences in production and perception of Dutch", J. Phon., 32:251-276, 2004.

[29] Warner, N. and Weber, A., "Perception of epenthetic stops", J. Phon, 29:53-87, 2001.

[30] Weber, A. and Cutler, A., "Lexical competition in non-native spoken-word recognition", J. Mem. Lang., 50:1-25, 2004.