Modeling language-learners' errors in understanding casual speech

L. ten Bosch[1], G. Giezenaar[1], L. Boves[1], M. Ernestus[1,2]

[1]Radboud University Nijmegen
[2]Max Planck Institute for Psycholinguistics
{l.tenbosch, g.giezenaar, l.boves, m.ernestus}@let.ru.nl

Abstract

In spontaneous conversations, words are often produced in reduced form compared to formal careful speech. In English, for instance, 'probably' may be pronounced as 'poly' and 'police' as 'plice'. Reduced forms are very common, and native listeners usually do not have any problems with interpreting these reduced forms in context. Non-native listeners, however, have great difficulties in comprehending reduced forms. In order to investigate the problems in comprehension that non-native listeners experience, a dictation experiment was conducted in which sentences were presented auditorily to non-natives either in full (unreduced) or reduced form. The types of errors made by the L2 listeners reveal aspects of the cognitive processes underlying this dictation task. In addition, we compare the errors made by these human participants with the type of word errors made by DIANA, a recently developed computational model of word comprehension.

Index Terms: dictation task, non-native perception, computational modeling, spoken word recognition

1. Introduction

It has been known for a long time that reduction is a phenomenon of everyday speech, affecting a high percentage of the word tokens in spontaneous conversations. In the Buckeye Corpus of Spontaneous American English [1], no less than 40% of all words tokens lack a segment, while about 6% of the word tokens lack a complete syllable [2]. Acoustic reduction is highly frequent also in other Germanic languages, such as Dutch [3] and German [4], and it occurs in non-Germanic language as well, such as French [5] and Finnish [6]; for an overview, see [7]. An English example of a reduced word form is the form /jɛ-ʃej/ for *yesterday*; a French example is /miz/ for *ministre* /ministrə/. Examples of Dutch, the language that is studied in this paper, include /tyrlək/ or even /tyk/ for *natuurlijk* /natyrlək/ *of course*, /fkopə/ for *verkopen* /vərkopən/ *to sell*, /zodɑk/ for *zodat ik* /zo dɑt ɪk/ *such that I*, /xon/ for *gewoon* /xəʋon/ *usual* and /ʋɛs/ for *wedstrijd* /ʋɛtstrɛit/ *game*. Compared to their canonical form, reduced words are always characterized by shorter, weaker or absent segments; even entire syllables may be absent.

For native listeners reduction phenomena go almost unnoticed. For non-native listeners, however, reduced words often present a serious difficulty, especially for the comprehension of spontaneous conversations. There are only a few studies published so far investigating how non-native listeners process reduced pronunciation variants. For example, [8] conducted a dictation task by presenting words in isolation, which showed that there is a large difference in terms of comprehension by Dutch listeners between full and reduced variants of a set of Canadian-French words (92.2 versus 56.1 percent correct, respectively).

In this paper we address the question how reduction affects word comprehension by non-native listeners during a dictation task in which reduced forms occur in natural linguistic contexts (rather than in isolation). Non-native intermediate (level A2-B1 according to the Common European Framework of Reference for Languages) learners of Dutch listened to dictated utterances with the instruction to write down what they heard. 'Dictation' is a regular part in training programs for Dutch as a second language. However, the interpretation of the instruction is left to the learners. Some will try and keep to restrict themselves to conventional spellings of real words, while others may invent words/spellings that match sequences of sounds that they do not recognize as known words. Participants could listen more than once to the same utterance, which means that time pressure was not an issue.

A dictation task tests both listening and writing skills. We therefore created two versions of the dictation tasks: participants either heard all sentences in full or with reduced forms. Comparison of the responses of both variants will show which errors results from the reduced forms and which ones have their origin in participants' problems with correctly spelling known words. All words in the sentences should be known by the participants, since these words were part of the training material taught in previous lessons.

All utterances were prerecorded. The reduced variants of the sentences were constructed such that they show various different types of reduction, from the shortening of segments to the complete deletion of whole syllables. The use of reduced forms in dictation is not trivial. Listeners do not expect reduced forms in a dictation task, since dictation is usually associated to carefully pronounced utterances. The reduced forms were therefore incorporated in sentences and together the sentences form one story, such that the materials seem to form a spontaneous monologue, in which reduced forms are likely.

We analyzed the errors made by the participants in two ways. First, we provide a qualitative and quantitative description of the errors by comparing these on the full and reduced speech material. Second, we compare the human-made errors with the errors made by DIANA, a computational model of human word recognition, on the exact same task. More specifically, we investigate which assumptions we have to make to have DIANA simulate the non-native listeners' performance.

1.1. Background

In order to classify the errors non-native (L2) listeners make in a dictation task, we take the speech comprehension process as a starting point. The speech comprehension process combines bottom-up and top-down processing. Top-down prediction involves multiple levels of representation such as the morphological, syntactical and contextual level [9, 10]. Predictions are con-

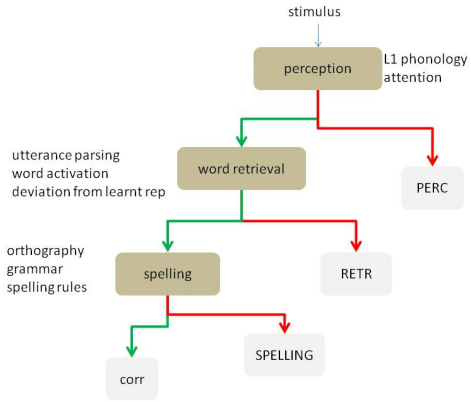

Figure 1: Scheme used for labeling the transcription errors. The scheme shows the following abbreviations: 'rep' for 'representation, 'PERC' for 'perception error', 'RETR' for 'retrieval error', 'corr' for 'correct' and 'spelling' for 'spelling error'. The abbreviation 'rep' refers to 'representation'.

ditioned by the unfolding bottom-up evidence from the speech signal. According to the dual route model of speech processing, speech comprehension in a native language is a process in which all information streams are smoothly calibrated and integrated. It is suggested that the comprehension process of L2 listeners relies more strongly on bottom-up information, and takes more time (e.g. [11]).

In an L2 dictation task transcription errors made by the participants may derive from three different sources, related to different stages in the cascade of the putative cognitive processes (fig. 1). First, early in the cascade, participants might have problems interpreting the speech signal at the phonemic level, because phones occur in the utterance that are absent in their native language (L1). Perceptual integration of L1 and L2 phone patterns due to interference between the L1 and L2 sound systems has been shown by several studies (see e.g. [12] and references therein). Evidently, perception errors may lead to a poor match between the signal and the representations of L2 words in a listener's mental lexicon, which may lead to problems retrieving the intended word and result in the recognition of a different word. The second potential source is insufficient familiarity with a word itself, or insufficient familiarity with a specific acoustic realization of that word. This may happen even if the L2 phones are correctly perceived. This type of retrieval error is likely to play an important role when a word is of a low frequency of occurrence or is spoken in a reduced form. Finally, of course, transcription errors may result from the participant's unfamiliarity with the word's orthography or with spelling rules that have a morphological or grammatical basis. Dutch has many homophone-heterographs, where the correct spelling is determined by morphology and syntax. Unsurprisingly, these forms may also cause substantial problems to native speakers of Dutch.

Although these error sources are different in nature and relate to different cognitive processes, they may result in the same transcriptions. Therefore, it may not be possible to assign all transcription errors to a unique stage of the dictation process.

2. Experiment

2.1. Participants

A group of 58 learners of Dutch, who followed a course which would bring them from level A2 to level B1 (levels according to the Common European Framework of Reference for Languages, CEFR), participated in the dictation task. The learners had 27 different language backgrounds (including, e.g., Chinese and Serbian), and also showed a wide age range (20 to 53 years, average: 36). They performed the task as part of their Dutch class and were not compensated for their participation. In addition, a group of 8 native listeners participated as a control group. They were all undergraduate students of Radboud University and were aged between 18 and 22 years (average: 20). They were paid for their participation.

2.2. Materials

The dictation task consisted of eleven Dutch sentences that formed a story. The sentences contained on average sixteen words. As mentioned above, the participants were supposed to know all words, because these appeared in previous lessons. Every sentence contained minimally two target words or word combinations that were focus points for the error analysis.

Two versions of the story were recorded by a female native speaker of Dutch, who was unknown to all participants. One version had all words produced in full, the other version had the target words reduced. The speech rate in both versions was low (approximately 4.4 and 4.8 syllables per second in the full and reduced version), such that misunderstandings would not be due to an excessive speech rate during the task.

One of the dictation sentences ran as follows:

Ik dacht niet aan mijn studie, maar probeerde
I did not think of my study, but tried
zoveel mogelijk feestjes te bezoeken.
to visit as many parties as possible.

In the full realization, all words were pronounced in canonical form. In the reduced variant, the target word combinations indicated in bold 'ik dacht' /ɪg dɑxt/ *I thought* and 'zoveel mogelijk' /zovel moxələk/ *as many as possible* were reduced to /ɣdɑx/ and /zovel mok/. In addition, in the reduced variant, 'mijn' /mɛin/ *my* was produced as reduced /mə/.

To the native Dutch listeners, this sentence and all other sentences sounded natural and were completely comprehensible in both the full and reduced version.

2.3. Procedure

The dictation task was run at three different locations in the Netherlands (Leiden, Amsterdam, Nijmegen) by means of a web application (webexp2, [13]). This application randomly assigned either the full or the reduced version of the dictation task to a given participant. The participants heard the sentences via headphones; the learners were situated in a class room and the native listeners in sound attenuated booths. They were told that they had to finish the task within 20 minutes (this amounts to about 110 seconds per sentence). By using the web interface, participants themselves could determine how often each utterance was played; it was impossible, however, to listen again to previous utterances. All participants completed the task in less than 20 minutes. Participants could type freely. The web application did not provide any help in spelling. Only the character sequence that was on the screen at the moment when a Carriage

Return was typed was stored; thus, we had no access to revisions –if any– or to the order in which 'words' were typed by the participants.

In total the 58 L2 participants produced 1936 transcriptions of the target words and word combinations, 152 of which were clearly not an attempt to faithfully render the spoken stimulus. However, part of the 'non-word' transcriptions were probably provided as an indication that part of the stimulus was not understood at all.

3. Classification of the errors

Of the 1784 (correct and incorrect) word-like transcriptions, 851 and 933 originated from full and reduced sentences, respectively. This difference results from the random generator in the webex2 application, which more often assigned participants the reduced than the full version of the task. Of the full target words, 620 (73%) were transcribed correctly, whereas for the reduced target words, this was 582 (62%). Thus, it appears that the transcription of the reduced pronunciation forms was more difficult than the transcription of the full forms. Quite a number of target words elicited non-canonical transcriptions, also in the full version of the dictation task. Importantly, the number of transcription variants was always larger for the reduced than for the full forms. This is in line with the assumption that reduced forms are more difficult to decode than full forms.

We have attempted to classify the errors by relating them to the different stages of the dictation task: the representation of the signal with a phonetic-phonological decoding system; followed by a word retrieval stage, followed by the choice of a spelling for that word. Some errors can reliably be attributed to perception errors (misperception of phones or words, 'want het is' – 'want dat is', 'dan het is'), to word retrieval problems ('tegen woorden' instead of 'tegenwoordig'; 'hartstikke' – 'hard sterk'), or to spelling problems (e.g. 'dat betekend' instead of 'dat betekent'). The purpose of the classification was to facilitate the specification of the processes that might result in a given transcription.

Table 1 shows an overview of the results for all 1784 word-like responses to the target words. The labels used to classify the mismatches between the reference transcription and the learner transcription include 'PERC' (transcription difference ascribed to a perception error), 'RETR' (errors ascribed to retrieval problems), 'GRAM' and 'SPELLING' (transcription differences assigned to a spelling problem, possibly due to lack of knowledge of morphology or syntax) and 'CORRECT' (no difference). Perception and retrieval errors manifest themselves in the form of insertions, deletions and substitutions ('INS', 'DEL', 'SUBS').

The most prominent difference between the full and reduced versions is in the number of deletions. Compared to unreduced speech, listeners are over three times more likely to miss a syllable or word when listening to reduced speech (37/851, i.e. 4.2%, versus 135/933, i.e. 14.5%). In the reponses to the reduced words, the proportion of errors that can be attributed to spelling problems is smaller than in the responses to the full stimuli; at the same time the proportion of errors that can be attributed to *earlier* stages in the comprehension process (PERC and RETR) is larger in the responses to the reduced words than to their full counterparts. The number of transcription errors that can be related to grammar (GRAM; e.g. inflection, t/d errors in verb forms) is the same in the responses to reduced and full items.

Here, two observations can be made. First, the data set is small. It is therefore dangerous to draw firm conclusions. However, the error analysis reveals patterns that are consistent with the expectations that we have for L2 listeners who are confronted with reduced speech. Second, our transcription errors labeling leaves room for some ambiguity that is difficult (and maybe impossible) to resolve. Small deviations from the correct orthographic form can almost always be interpreted as spelling errors. When transcriptions substantially deviate from the correct form ('wil ik daarom' – 'wil ik daarenweer'), this may point to problems with retrieval of the correct word or word sequence. Word retrieval problems themselves may interfere with erroneous perception of the speech sounds. Whether a deviant orthographic transcription is based on a perceptual error, or on the lack of available word representations in the passive mental lexicon, or on a spelling problem is difficult to determine.

Table 1: Total numbers of errors for each error type. A response can be labeled with more than one label, and the sum of the classification types may therefore exceed the total number of responses (i.e. 851 and 933 responses to full and reduced target words, respectively).

	unreduced	reduced
total nr of instances	851	933
PERC	51	96
RETR	41	54
GRAM	15	17
INS	3	6
SUBS	13	10
DEL	37	135
SPELLING	151	135
CORRECT	620	582

In the next section we present DIANA, an end-to-end computational model of spoken word recognition. In the past, DIANA has proven successful in simulating the behavior of native participants in word recognition and lexical decision experiments [14, 15, 16]. Here, DIANA is used to compare its errors with errors made by the learners in the dictation task, which is a novel task for DIANA. We will investigate which type(s) of transcription errors can be simulated by DIANA in its present form, and how the model should be extended to be able to simulate the remaining error types. In doing so, we will ignore spelling errors, because these errors are caused in a stage of the transcription process that is not affected by problems caused by reduced pronunciation forms. This can be inferred from the fact that the number of spelling errors in the full version of the dictation task is at least as large as the number of spelling errors in the reduced version (see Table 1). We will focus on errors related to perception and retrieval, which manifest themselves in different ways.

4. DIANA, a computational model of word comprehension

DIANA [14] is an end-to-end computational model of spoken word comprehension, which takes as input the speech signal. Depending on the task, it can provide as output the orthographic transcription of the stimulus, a word/non-word decision and the associated estimated reaction time (RT). The model has been tested for simulating RTs in lexical decision tasks in Dutch [15] and North American English [16]. DIANA consists of three components: an Activation component, which is reminiscent of

models of spoken word recognition such as Shortlist-B [17], a Decision component [18, 19, 20], and an Execution component, which accounts for the time that is needed to externalize the result of cognitive processing in the form of some physical action. Since we will not simulate the participants' RTs, the Execution component is not relevant in this paper.

The Activation component in DIANA operates on real acoustic speech signals. Its task is to convert these signals into hypotheses about the sequence of words that best corresponds to the acoustic signal. The sequences, and the individual words that make up a sequence, obtain a probability score. Competing hypotheses are sent to the Decision component. In word recognition and lexical decision experiments, the Decision component determines the point in time at which the score of the leading hypothesis is so much higher than the score of the runner-up that a final winner can be selected. In word recognition, all hypotheses are words in DIANA's vocabulary; thus, the Decision component only establishes the moment at which the leader can be promoted to winner status. In lexical decision, the task of the Decision component is more complex; here it must also decide whether the best-scoring hypothesis consisting of words from the vocabulary compares favorably to a hypothesis that contains one or more non-words. For that purpose DIANA decodes the incoming speech signal in two ways: as a sequence of real words, and as a sequence of sub-word units that do not form words. In the present implementation of DIANA these sub-word units correspond to speech sounds (phones).

For computing probability scores of lexical hypotheses, the Activation component draws upon three different resources that together constitute the language proficiency of the model. The first resource is the lexicon, which lists all word forms that the model knows. Each word form consists of three specifications: a unique orthography, one or more pronunciation variants, and the frequency with which the word occurs in a large corpus. There may be just one frequency count for the entire word or also frequency counts for the different pronunciation variants. The current implementation of DIANA can handle vocabularies of over 30,000 entries.

The second resource consists of statistical models of the acoustic realization of the sub-word units that are used in the specification of the pronunciations in the lexicon. As said before, in the present version of DIANA these sub-word units are phones. Phones are modeled as three-state hidden Markov models. Each state consists of a mixture of 32 Gaussian distributions that together specify the likelihood that an observation is generated by this state. The three states in a phone model must always be traversed from left to right; skipping states is not possible. The set of sub-word unit models is often called the acoustic model (AM).

The third knowledge source in DIANA's Activation component is the language model (LM), which specifies the probability of observing a word after having observed one or more preceding words. Statistical N-gram models capture a large part of the syntactic structure of sentences that is usually specified in quite different ways in linguistic grammars of a language. It is fair to say that N-gram models capture the implicit knowledge of the syntax that one acquires through exposure to a language, while conventional linguistic grammars specify meta-level knowledge about the structure. N-gram models can be learned, while grammars can be taught. As with the occurrence counts in the vocabulary, the transition probabilities in an N-gram model will depend on the corpus from which it learned. From corpus linguistics it is known that the choice of the corpus has a large effect. It is also known that a LM learned from a general purpose corpus is not a very good fit for the language used in a specific situation by a specific person. For this reason it is customary to create LMs that combine data from a general corpus with data collected in a specific situation. The part of the LM learned from a general purpose corpus is usually called a background model, while the part learned from situation-specific data is called a foreground model.

DIANA is implemented using the Hidden Markov Toolkit (HTK) [21]. For the implementation of the model used in this paper we started from a full-fledged automatic speech recognition system [22]. The AM as well as the LM were built using parts of the spoken Dutch corpus. The speaker-independent acoustic models in that system were adapted to the speech of the female speaker who produced the sentences for the dictation task. For this purpose she read 500 phonetically rich sentences that were recorded with the sam equipment as the dictation prompts. The adaptation was performed using the procedure HERest in HTK.

The activation score (the probability) of a word is determined by a weighted combination of the acoustic evidence based on the AM $P(\text{acoustics}|\text{word})$ and the prior probability $P(\text{word}|\text{precontext})$ (the frequency of the words and the language model). To obtain optimal results, the contributions of the AM and LM must be given different weights. For this purpose DIANA contains the parameter γ (cf. Eq. (1)). For values $\gamma \ll 1$ the contribution of prior knowledge about the structure of the language is small; as a consequence, the decoding will be based mainly on the acoustic match. For values $\gamma \gg 1$ the reverse is true: the scores will mainly depend on prior knowledge. Hidden in the LM, as it were, there is an additional parameter, the word insertion penalty (WIP), which specifies the cost of going from one word to the next. In a language such as Dutch, which has many compounds that consist of sequences of words that can also occur independently in the same sequence, a high value of WIP, i.e., a high cost attached to entering a new word, favors decodings of the acoustic signal as short sequences of long words. Low values of WIP favor decodings as long sequences of short words.

$$\begin{aligned} \text{score} &= \text{AM} \\ &+ \gamma * \overbrace{(\text{foregroundLM} + \text{backgroundLM})}^{\text{LM}} \\ &+ \text{WIP} \end{aligned} \tag{1}$$

The Decision component in the present version of DIANA was designed to determine the time it takes to make a decision, and to distinguish between real words and pseudo-words. For simulating the dictation task, in which we have no timing information, the former capability is not relevant. However, the capability to distinguish between real words and non-words allows DIANA to determine that part of a spoken utterance does not correspond to real words. The present version of DIANA will identify that stretch of speech as a pseudo-word. The pseudo-word decision is based on the fact that the probability of a decoding of that stretch of speech in terms of a sequence of sub-word units (phones) is higher than the best competing decoding in the form of real words. It would be simple to convert the best-scoring sequence of sub-word units into the spelling that one would expect if the sub-word units would make up a real word.

The present version of the Decision component does not reorder the rank of the hypotheses delivered by the Activation

component. This is because there are no knowledge sources that could be used for this purpose. It may well be that L2 learners who do a dictation task eventually settle for a hypothesis that is less probable in terms of the combined information in the acoustic model and N-gram language model, but that would be more probable on the basis of semantic or pragmatic knowledge. It might even happen that a learner revises a transcription to make it fit with explicit grammatical knowledge.

5. Understanding L2 transcription through simulation

We used a slightly extended version of DIANA to investigate the output that was produced with different implementations of the vocabulary and language model, and different values of the parameters γ and WIP. We focused on retrieval problems, i.e., the type of problems that is expected to depend least on the participant's exact L1 backgroud. Perception errors, in contrast, are almost by definition due to L1-L2 interference. To simulate perception errors, we would have to build multiple acoustic models, one for each (group of) L1. While this is theoretically possible, it is very costly and cumbersome in practice. Moreover, it is not evident that simulations with L1-specific acoustic models would provide knowledge and insights over and above what is already available from research dedicated to L1-L2 interference at the phonemic-phonetic level [12].

5.1. Issues related to the vocabulary

While the issues involved in designing the vocabulary in an automatic speech recognition system are fairly well understood, this is not true for the lexicon that should be used in the simulation of an L2-learner in a dictation task. Investigating the impact of reduced pronunciation variants in the dictation sentences complicates things even further. Even if we ignore difficult issues, such as potential interactions between the L1 and L2 vocabularies, many open questions remain. Should the vocabulary include all words that occurred in previous classes and lessons? Should it contain only the words encountered in the lessons, or should one also add words and expressions that the learners may have encountered outside the classroom? Is it possible to compute reliable, occurrence counts for the words that are included? Should reduced pronunciation variants be included, or would that render an experiment with reduced variants trivial? But then again, if there is no knowledge about reduced variants, how would a learner be able to correctly transcribe reduced variants? Only by revising 'incorrect' decodings using semantic or syntactic knowledge?

Another question is what counts as a 'word'. There is a large literature on multi-word expressions, which come in different types, from fixed sequences of words with a meaning that cannot be deduced from the meaning of the individual words, to sequences whose only claim to a special status is their frequency of use, perhaps combined with the fact that some parts are often substantially reduced. In Dutch L2 courses, special attention is paid to the fact that a few sequences such as *ik heb het* 'I have it' are often reduced to two or one syllables. It is not known how native speakers process these forms. For L2 learners who have been taught these forms explicitly, it might be argued that they should go into the vocabulary, possibly with only a fraction of the occurrence count of the expression.

5.2. Issues related to the language model

For the construction of an L2-learner's language model the same question arise as for the L2 vocabulary: should the model be based only on the texts in the course material, or should it be extended with texts in the L2 that (most) learners are likely to be familiar with? In many L2 courses, in any case in courses of Dutch as L2, explicit attention is paid to syntax and morphology. This raises the question whether the L2 language model should not only be based on statistical N-gram models, but rather on some mix of N-grams and a formal grammar. The algorithms for building and using such a mix are available, but little is known about the link with the simulation of cognitive processes in L2-learners.

5.3. Simulations

To make a first shot at investigating the theoretical and practical issues discussed above we conducted a number of simulation experiments that were mainly aimed at better understanding word deletions and the observation that L2 learners tend to break up long words into sequences of shorter words, e.g. *mogelijkheid* 'possibility' – *mogelijk tijd* 'possible (incorrectly uninflected) time'; *openbaar* 'public' – *open haar* 'open here'; *ontbijtje* 'small breakfast' – *ontbijt je* 'you have breakfast', *telkens* 'every time' – *te elke* 'to every' (which is ungrammatical) , etc. We created two vocabularies: V_1, which only contains the 118 words in the dictated story, and V_2, which contains all 4576 words that occur in the course material. The vocabularies only contained canonical (full) pronunciation variants.

A foreground bigram LM was created on the basis of the dictation sentences. We merged this foreground LM (by merging word token counts) with a general background unigram language model based on the 4576 words from the language course material, since these words were assumed to form the background mental lexicon available for the participants. The prior probabilities of these words were set according to their frequency counts in Spoken Dutch Corpus [23]. These 'systems' were used with a range of values of the parameters γ and WIP.

The breaking-up of long words can be simulated by reducing γ and WIP at the same time. This combination strengthens the contribution of the acoustic evidence and lowers the cost of a decoding of the signal as a sequence of many short words. However, the story is more complex than this. For small values of WIP, we observe many insertions of short words, such as *op* 'on', *in* 'in' that were not observed in the L2 dictation output. Because these function words are frequent, their insertion is likely, especially when they are phonetically close to word final codas and/or word initial cohorts. We expect that these spurious insertion errors can be reduced with a more appropriate LM, either by increasing the N in the N-gram model, or by creating a mix of N-grams and a grammar. Here, again, semantics does not play any role and so cannot prohibit these insertions in DIANA's recognition result. In addition, insertion errors may also disappear for larger values of WIP. For very large values of the WIP, we see many deletions of short words (e.g. determiners), especially if they are adjacent to a long word. These results suggest that non-natives have a preference for short words and have problems predicting words from the previous words.

6. Conclusion

L2 learners of Dutch participating in a dictation task show more, and more varied transcription errors for reduced utterances than for unreduced utterances. Evidently, listeners have problems converting what they hear into real Dutch words. Sometimes listeners invent words such as *mook*, *mok*, *mog* for *mogelijk*, which shows that they are able to faithfully represent speech sounds into pseudo-words but not able to retrieve the corresponding word *mogelijk*. This may be due to the way in which such highly reducible words are presented during the language course, that is, only their full acoustic forms.

We developed a taxonomy of errors that makes it possible to determine the most likely causes of the errors: in perception, in lexical access or in spelling. Since we are interested in the effect of reduction, we ignored spelling errors, as these occur equally often in the full and reduced versions.

We discussed ways in which DIANA, a computational model of spoken word comprehension that has been successful in simulating word recognition and lexical decision, could be extended to simulate lexical retrieval errors. This resulted in a fairly long list of desiderata. This list will guide future research into the effect of reduction on L2 comprehension.

7. Acknowledgement

This work was funded by an ERC starting grant (284108) and an NWO VICI grant awarded to Mirjam Ernestus. Karolien Kamma, Freek Bakker and Johanneke Caspers are gratefully acknowledged for coordinating the dictation tasks in Amsterdam and Leiden.

8. References

[1] M. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, "Buckeye corpus of conversational speech (2nd release)," www.buckeyecorpus.osu.edu, 2007, columbus, OH: Department of Psychology, Ohio State University (Distributor).

[2] K. Johnson, Massive reduction in conversational American English, K. Yoneyama and K. Maekawa, Eds. The National International Institute for Japanese Language, 2004.

[3] M. Ernestus, "Voice assimilation and segment reduction in casual dutch, a corpus-based study of the phonology-phonetics interface." LOT, Utrecht, the Netherlands, 2000.

[4] K. Kohler, "Segmental reduction in connected speech in german: Phonological facts and phonetic explanations." In: Hardcastle, W.J., Marchal, A. (Eds), Speech production and speech modelling. Kluwer Academic Publishers, Dordrecht, pp. 21–33, 1990.

[5] M. Adda-Decker, P. Boula de Mareüil, G. Adda, and L. Lamel, "Investigating syllabic structures and their variation in spontaneous french," Speech Communication, vol. 46, pp. 119–139, 2005.

[6] M. Lennes, N. Alaroty, and M. Vainio, "Is the phonetic quality of unaccented words unpredictable? an example from spontaneous finnish," Journal of the International Phonetic Association, vol. 31, pp. 127–138, 2001.

[7] M. Ernestus and N. Warner, "An introduction to reduced pronunciation variants," Journal of Phonetics, vol. 39, pp. 253–260, 2011.

[8] D. Nouveau, "Limites perspectives de l'e caduc chez des apprenants néerlandophones," Revue Canadienne de Linguistique Appliquée, vol. 15, pp. 60–78, 2012.

[9] M. J. Pickering and S. Garrod, "Do people use language production to make predictions during comprehension?" Trends in cognitive sciences, vol. 11(3), pp. 105–110, 2007.

[10] ——, "An integrated theory of language production and comprehension." Behavioral and Brain Sciences, vol. 36, pp. 329–347, 2013.

[11] M. H. Christiansen and N. Chater, "The now-or-never bottleneck: A fundamental constraint on language (author's version)," Behavioral and Brain Sciences, vol. 37, pp. 1–52, 2015.

[12] J. Flege, C. Schirru, and I. MacKay, "Interaction between the native and second language phonetic subsystems," Speech Communication, vol. 40, pp. 467–491, 2003.

[13] webexp2, "Web application for conducting linguistic experiments," URL: groups.inf.ed.ac.uk/webexp/docs-/UserManual.pdf, 2013.

[14] L. ten Bosch, L. Boves, and M. Ernestus, "Towards an end-to-end computational model of speech comprehension: Simulating a lexical decision task," in Proceedings of Interspeech, Lyon, France, 2013.

[15] ——, "Comparing reaction time sequences from human participants and computational models," in Proceedings of Interspeech, Singapore, 2014.

[16] L. ten Bosch, L. Boves, B. Tucker, and M. Ernestus, "DIANA: towards computational modeling reaction times in lexical decision in North American English," in Proceedings of Interspeech, Dresden, 2015.

[17] D. Norris and J. McQueen, "Shortlist B: A Bayesian model of continuous speech recognition," Psychological Review, vol. 115, pp. 357 – 395, 2008.

[18] R. Ratcliff, "Continuous versus discrete information processing: Modeling the accumulation of partial information," Psychological Review, vol. 95, pp. 238 – 255, 1988.

[19] R. Bogacz, E. Brown, Moehlis, P. E., Holmes, and J. D. Cohen, "The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced choice tasks," Psychological Review, vol. 113, pp. 700–765, 2006.

[20] M. Usher, Z. Olami, and J. L. McClelland, "Hick's law in a stochastic race model with speed-accuracy trade-off," Journal of Mathematical Psychology, vol. 46, pp. 704 – 715, 2002.

[21] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK book (for HTK version 3.4)," Cambridge University Engineering Department, Cambridge, UK, Tech. Rep., 2009.

[22] A. Hämäläinen, M. Gubian, L. ten Bosch, and L. Boves, "Analysis of acoustic reduction using spectral similarity measures," Journal Acoustical Society of America, vol. 126, pp. 3227–3235, 2009.

[23] N. Oostdijk, "The spoken dutch corpus: Overview and first evaluation," Proceedings LREC, pp. 887–893, 2000.