Journal Article

Iris Hanique, Mirjam Ernestus*, and Lou Boves

Choice and pronunciation of words: Individual differences within a homogeneous group of speakers

Abstract: This paper investigates whether individual speakers forming a homogeneous group differ in their choice and pronunciation of words when engaged in casual conversation, and if so, how they differ. More specifically, it examines whether the Balanced Winnow classifier is able to distinguish between the twenty speakers of the Ernestus Corpus of Spontaneous Dutch, who all have the same social background. To examine differences in choice and pronunciation of words, instead of characteristics of the speech signal itself, classification was based on lexical and pronunciation features extracted from hand-made orthographic and automatically generated broad phonetic transcriptions. The lexical features consisted of words and two-word combinations. The pronunciation features represented pronunciation variations at the word and phone level that are typical for casual speech. The best classifier achieved a performance of 79.9% and was based on the lexical features and on the pronunciation features representing single phones and triphones. The speakers must thus differ from each other in these features. Inspection of the relevant features indicated that, among other things, the words relevant for classification generally do not contain much semantic content, and that speakers differ not only from each other in the use of these words but also in their pronunciation.

Keywords: casual speech, acoustic reduction, individual differences, speaker classification

Iris Hanique: Radboud University Nijmegen & Max Planck Institute for Psycholinguistics
***Corresponding Author: Mirjam Ernestus:** Radboud University Nijmegen & Max Planck Institute for Psycholinguistics
Lou Boves: Radboud University Nijmegen

1 Introduction

Language users have a multitude of different words at their disposal, and individuals may differ in their choice of words. For instance, some people may prefer the word *start* to the word *begin* or may use *big* instead of *large*. In speech, an additional type of variation is the exact pronunciation of words. Many words produced in casual speech show a range of possible pronunciations from the full pronunciation variant to highly reduced ones, in which phones are replaced by others or are completely missing. For instance, *probably* may be pronounced as [prɑbəbli], [prɑbli], [prɑli], and [prɑ]. Deviations in one phone from the full form occur in over 60% of the word tokens in casual American English, and two or more phones deviate in 28% of the tokens (Johnson, 2004). Similar numbers have been found for other languages (Ernestus and Warner, 2011). In this paper, we focus on a socially homogeneous group of speakers and investigate whether these speakers differ in their choice and pronunciation of words in casual conversations, and if so, how they differ. Research on individual differences in conversational speech will improve our understanding of the speech production process, and may help us improve psycholinguistic models of speech processing.

Previous research on individual differences in word choice has focused on written text and function words (e.g., Ebrahimpour et al., 2013; Koppel et al., 2009; Stamatatos, 2009). Content words, such as *table* and *sleeping*, and word combinations, such as *old tree*, are very context dependent. In contrast, function words, such as *that* and *but*, are not likely to vary greatly with the topic of the text and can consequently more easily reveal topic independent individual differences in word choice. Nevertheless, Barlow (2010) showed that under specific circumstances the use of content words may distinguish between speakers, as he reported differences in the frequencies with which five white house press secretaries use content words and function words as well as word combinations. In the present study, we investigate the roles of both function words and content words, henceforth *unigrams*, and also of combinations of two words, henceforth *bigrams*.

Differences in acoustic reduction have been shown between groups of speakers. Several studies have reported effects of gender; for example, in Dutch, men tend to reduce words ending in the suffix /lək/ *-lijk* more often than women (e.g., in /moxələk/ *mogelijk* ‘possible’; Keune et al., 2005), and, in American English, they more often delete word-final /d/ and /t/ (Guy, 1980) and glides (Phillips, 1994). Further, younger speakers tend to reduce more than older speakers. This has been demonstrated, for instance, for the

absence of word-final /d/ and /t/ in American English (Guy, 1980) and for the absence of segments in spontaneous Dutch (Strik et al., 2008). Finally, speakers of Dutch in Flanders tend to reduce less than speakers of Dutch in the Netherlands (Keune et al., 2005).

There is also some evidence that individual speakers may differ from each other in their reduction of words, even if they are members of the same social group. Ernestus (2000:143) studied a group of twenty speakers who were all highly educated men aged between 21 and 55, and who were all born and raised in the western part of the Netherlands. She observed differences in the pronunciation of the Dutch word /natyrlək/ *natuurlijk* "of course". Whereas most speakers only produced the extremely reduced variant [tyk] in the middle of Intonational Phrases, one speaker also pronounced [tyk] in the initial and final positions of the Intonational Phrase and even in isolation. This raises the question whether differences between individual speakers can also be observed for other reduction phenomena that are typical for casual conversations.

Similar to Van Bael and Van Halteren (2007), to study differences in the choice and reduction of words between individual speakers, we applied a classification algorithm, in which speech fragments are attributed to their speakers on the basis of lexical and pronunciation patterns. If classification results in high performance scores, this would indicate that speakers differ in their speech habits. To examine how speakers differ, we inspected which words and pronunciation variants were important for distinguishing a speaker from others.

Our study is based on human-made orthographic and automatically generated broad phonetic transcriptions. These show the words that were used and how these words were pronounced at the phone level. By using broad phonetic transcriptions, we ignore all detailed information in the spectro-temporal representation of the speech. We do so because this spectro-temporal representation not only contains linguistically relevant information about how words were exactly articulated, but also paralinguistic information including voice quality, and these two types of information cannot easily be separated.

We are not the first to apply speaker classification to phonetic transcriptions. Van Bael and Van Halteren (2007) studied the effects of the speaker's age, gender, regional background, and level of education on word choice and pronunciation variation by classifying speakers belonging to groups differing in these characteristics. Using automatically generated broad phonetic transcriptions of the telephone dialogues of the Spoken Dutch Corpus (Oostdijk, 2002), the authors generated two sets of classification features: one set of approximately 150,000 lexical features, including average utterance length, part-of-speech tags, and uni-, bi-, and trigram counts, and another set of 94

pronunciation features representing phone differences between full pronunciations of words and their actual phonetic transcriptions. The classification algorithm that they used was able to classify speakers according to their age, gender, and regional background on the basis of lexical features. Interestingly, classification was hardly effective on the basis of the pronunciation features. The authors suggested that this may be due to the broadness of the phonetic transcriptions, the limited set of pronunciation features, or the heterogeneity within their speaker groups.

This paper builds on the work by Van Bael and Van Halteren (2007) by also using a classifier to investigate pronunciation differences between speakers and by applying the classifier on phonetic transcriptions of conversations instead of the speech signal itself. However, we address related but different research questions, since we focus on Dutch speakers who have the same regional background, gender, and educational level; that is, a homogeneous set of speakers. We investigate whether these speakers show individual differences in their choice and pronunciation of words in casual conversations, and if so, how these speakers differ. In addition, our study differs from Van Bael and Van Halteren (2007) in that we use a different classifier and different features, and we developed our own research approach.

2 Method

2.1 Speech data

For our study, we used the Ernestus Corpus of Spontaneous Dutch (ECSD; Ernestus, 2000), which consists of 15 hours of casual dialogues produced by ten pairs of speakers. These twenty speakers together uttered 155,294 word tokens representing 9044 word types. On average, each speaker produced 7765 word tokens (ranging from 5419 to 10,936 tokens). The speakers form a very homogeneous group: they are all males who hold academic degrees. Further, they are all native speakers of standard Dutch born and raised in the western part of the Netherlands. The main characteristic in which these speakers vary is their age, which ranges from 21 to 55 years.

Schuppler et al. (2011) generated broad phonetic transcriptions for the ECSD using an automatic speech recognition (ASR) system based on the Hidden Markov Model Toolkit (Young et al., 2002). An ASR system uses speech fragments and orthographic transcriptions of these fragments as input. In addition, it requires a pronunciation lexicon containing the full form and

possible pronunciation variants for each word in the corpus (e.g., for the Dutch word *gewoon* 'just' the lexicon contained the full form /xəʊon/ and the variants /xʊon/ and /xon/). These pronunciation variants were created with 32 rules that had been formulated on the basis of earlier observations of pronunciation variation and that insert, alter, or delete phones. Finally, the ASR system uses 37 monophone acoustic models consisting of three states with 32 Gaussians per state (Hämäläinen et al., 2009). On the basis of these phone models, which had been trained on the read speech component of the Spoken Dutch Corpus, the ASR system determined for each word in the orthographic transcriptions which variant from the pronunciation lexicon best matched the speech signal.

Schuppler et al. (2011) validated this transcription procedure by comparing its output for the IFA corpus (Van Son et al., 2001) with manual transcriptions of this corpus. They calculated how often phones in the automatic transcriptions deviated from those in the manual transcriptions in terms of insertions, replacements, and deletions. They observed an overall agreement of 86.0%, which is similar to agreements among human transcribers reported in the literature (e.g., Kipp et al., 1997, reported agreements between human-made transcriptions of spontaneous German of 78.8%, 79.9%, and 82.6%; for more information on agreements typically obtained for phonetic transcriptions see Ernestus and Baayen, 2011). Hanique et al. (2013) validated the automatically generated transcriptions of the ECSD with human-made transcriptions on the basis of 148 schwas in the initial syllables of past participles (as in /xəmɪst/ *gemist* 'missed'). Two human transcribers agreed on the presence versus absence of schwa in 82.4% of the tokens, while they agreed with the ASR system in 75.7% and 77.0% of the tokens. These agreements did not differ significantly from each other. Given these evaluations, and since obtaining better transcriptions for such a large corpus is difficult, we accepted these automatic transcriptions as being valid.

As automatic phonetic transcriptions can only be created for uninterrupted speech (i.e., without, for instance, overlapping speech or laughter), the number of transcribed words is lower than the number of words in the entire corpus. Our transcriptions contain 95,173 word tokens and 6965 word types, ranging from 1 to 3459 word tokens per word type. The most frequent word types were *ik* 'I' and *dat* 'that' with 3459 and 3402 tokens respectively. On average, for each speaker 4759 word tokens were transcribed, representing 944 word types.

For our classification tests (see below), we used the automatically generated phonetic transcriptions and the corresponding orthographic transcriptions. Moreover, we divided the transcriptions of each speaker into ten equally

sized fragments. The size of these fragments was different for each speaker: it varied between 375 and 742 word tokens, and had an average of 479 tokens.

2.2 Classification Features

Classification algorithms distinguish between classes (in our case speakers) on the basis of features which represent properties of these classes (e.g., single words such as *window* or the absence of a phone such as /t/). We represented each of the 200 fragments in our dataset (10 fragments per speaker) as a list of features that are based on the fragment's orthographic and phonetic transcription. We designed a number of lexical and pronunciation features.

To investigate word choice, we extracted all unigrams and bigrams from the transcriptions. If a word was not preceded or followed by another word, a bigram was created including a silence before or after the word. We then selected those unigrams and bigrams that occurred more than twenty times in the entire corpus and that were produced by at least two speakers. This resulted in 403 unigrams, each of which on average occurred 195 times in the entire corpus (range: 21 to 3459) and in 61 fragments (range: 12 to 200). The total number of bigrams was 642, each of which on average occurred 75 times in the corpus (range: 21 to 1931) and in 44 fragments (range: 12 to 199). Following Van der Sijs (2002), we considered prepositions, conjunctions, determiners, pronouns, and numerals as function words, and nouns, verbs, adjectives, adverbs, and interjections as content words. The selected unigrams consisted of 158 function word types and 245 content word types.

To study individual differences in the pronunciation of words, we designed pronunciation features. Since we focused on how speakers reduce words in casual speech, we ignored three well-known types of variation in Dutch pronunciation, because these occur as often in read speech as in casual speech. First, we ignored the variation in the pronunciation of Dutch word-final *-en*, which is mostly pronounced as [ə] and sometimes as [ən] (e.g., Booij, 1995, p. 139) except in the east of the Netherlands. Second, we also ignored the insertion of phones, for example, the pronunciation of /mɛlk/ *melk* 'milk' as [mɛlək], as this is not reduction.

Third, we ignored the variation in the pronunciation of obstruents as voiced or voiceless (e.g., /v/ versus /f/). The voicing of obstruents is highly variable in Dutch, especially as spoken in the western part of the Netherlands. Like most speakers from this area, our speakers often replaced voiced fricatives by their voiceless counterparts (Schuppler et al., 2011) and frequently applied regressive and progressive voice assimilation (Ernestus et al., 2006)

Moreover, the automatic transcriptions that were used in our study have been generated by means of acoustic models that may not reliably encode voicing for obstruents, since they had been trained with speech for which the voicing of obstruents may not have been reliably transcribed. We therefore collapsed the members of the obstruent pairs /z,s/, /v,f/, /d,t/, /g,k/, /b,p/, /ʒ,ʃ/, and /ɣ,x/.

We examined pronunciation variation at the phone level, henceforth *phone features*, and at the word level, henceforth *word pronunciation features*. We designed phone features providing information whether a phone was produced as it would be if the word had been produced in careful speech. For each phone in a word token, we determined whether it was unreduced, i.e., produced as in the full pronunciation of the word, was replaced by another phone, or was completely absent. For example, the Dutch word /heləmal/ *helemaal* 'completely' may be pronounced as [həlmal], in which we considered all consonants and the final vowel as being unreduced, the first vowel as being replaced by schwa, and the second vowel as being absent.

For each phone, we defined four types of possible features. One type was the phone itself without any neighboring phone, henceforth *uniphones*; for example, /e/ replaced by schwa. Two possible feature types included either the preceding or following phone from the full form, henceforth *biphones*; for example, /e/ replaced by schwa and followed by /l/. The final possible feature type represented the phone and both neighboring phones from the full form, henceforth *triphones*; for example, /e/ replaced by schwa in the sequence /hel/. If the phone was positioned at a word boundary, we took the neighboring phone from the neighboring word, for example, in [həlmalni] *helemaal niet*, the phone following the final [l] of *helemaal* was /n/.

Each possible feature was only used if it met the following two criteria. First, it had to occur more than twenty times in the entire corpus. Second, there had to be at least one other pronunciation of that phone (sequence) that occurs at least twice in the entire corpus; for example, the feature *replacement of* /e/ *by schwa followed by* /l/ was only used if another variant, that is a *present* or *absent* /e/ *followed by* /l/, occurred at least twice. In total, we used 955411 phone features that represent 2394 phone feature types. Table 1 presents the numbers of the phone features split to uniphones, biphones, and triphones, and whether the phone was unreduced, replaced, or absent. On average, a phone feature occurred 399 times (ranging from 21 to 41799 times) and in 87 fragments (ranging from 4 to 200 fragments).

Finally, we designed word pronunciation features representing pronunciation variation at the word level. For instance, the word *mensen* 'people' can be produced in the unreduced form [mɛnsə] or in a reduced variant, such as

Table 1. Number of tokens and types (in brackets) of the uniphone, biphone, and triphone features split into phones that are unreduced, replaced, or absent. Biphones with the preceding or following phone are indicated by *biphones prec.* and *biphones foll.*, respectively.

	Unreduced	Replaced	Absent
uniphones	267704 (31)	8770 (23)	32957 (27)
biphones prec.	213445 (327)	6676 (82)	31180 (150)
biphones foll.	211486 (319)	7822 (58)	31763 (143)
triphones	112478 (829)	5242 (54)	25888 (351)

[mɛns] or [mɛs]. We selected as word pronunciation features, word pronunciations that occurred more than twenty times in the entire corpus and produced by at least two speakers. In addition, the words they represented had to have at least one other pronunciation variant that occurred at least twice in the entire corpus. This is to ensure that the features could indeed capture pronunciation variation and not just lexical information, which would certainly be the case if there is only one pronunciation variant. Note that if the other pronunciation variant does not meet the selection criteria, this other variant is not used as a feature. In total, 290 word pronunciation variants met these restrictions, and these variants represent 157 word types (52 word types had one pronunciation variant; 86 word types had two variants; eleven word types had three variants; seven word types had four variants; and one word type had five variants). These variants represented 128 unreduced forms, for example [mɛnsə] *mensen* 'people', and 162 reduced variants, for example [mɛs], a variant of *mensen* 'people'.

2.3 Classification algorithm

We used the Balanced Winnow classifier (Dagan et al., 1997; Littlestone, 1988) implemented in the Linguistic Classification System (LCS; Koster et al., 2003; Koster and Beney, 2009). This algorithm assigns two weights (w^+ and w^-) to each feature for every speaker, and the overall (Winnow) weight is their difference. Features of a certain speaker with a positive overall weight are used to classify a fragment as belonging to that speaker, whereas those with a negative overall weight are used for classification as not belonging to that speaker. The value of an overall weight indicates how useful the feature is to distinguish the speaker from all other speakers in the corpus. The output for each speaker from the LCS is a model, henceforth *speaker profile*, which consists of two lists, one with positive overall weights and one with negative

overall weights. We used them to identify the speaker of a new fragment and thus to test how well the classifier performed. In addition, we used the profiles to characterize the differences among speakers: in a certain speaker profile, features with a positive overall weight are assumed to be characteristic for that speaker, whereas those with a negative overall weight are assumed to be uncharacteristic for that speaker.

The classifier created speaker profiles in the training phase in which it receives all (training) fragments as input, labeled as a positive or negative example for a given speaker. A fragment produced by a certain speaker is a positive example for only that speaker and is a negative example for all other speakers. The classifier first constructs initial speaker profiles on the basis of all fragments. Each initial profile consists of a list of all features with their initial weights, calculated through the LTC algorithm (Salton and Buckley, 1988). Subsequently, these speaker profiles are adapted during multiple training iterations, in each of which all (training) fragments are again presented to the classifier. For each fragment, the classifier calculates the correspondences between that fragment and every speaker based on the fragment's features and the weights for these features in each of the speaker profiles. Balanced Winnow is a mistake-driven classifier, which means that weights in a speaker profile are only updated if a fragment is classified incorrectly during training. If a fragment belonging to a speaker scores for that speaker above threshold θ^+, the fragment is correctly classified as a positive example and the weights remain unchanged. In contrast, if it scores below this threshold, the classification is treated as a mistake and the weights in the speaker profile are updated by multiplying the positive weights of the active features, i.e. those that occur in both the fragment and the speaker profile, with parameter α and the negative weights of the active features with parameter β. In addition, if a fragment that does not belong to a given speaker has a score for that speaker below another threshold, θ^-, it is correctly classified as a negative example, whereas if this fragment scores above the threshold, the weights of the active features in that speaker's profile are updated by multiplying positive weights with β and negative weights with α. After testing values for the parameters around the default settings of LCS, we used those settings that resulted in the highest performance, namely $\alpha = 1.05$, $\beta = 0.98$, $\theta^+ = 0.8$, $\theta^- = -0.8$, and a maximum of 20 training iterations.

2.4 Classification tests

To test how well the classifier performed, we used ten-fold-cross-validation: The classifier was trained on 180 fragments (nine from each speaker) and tested with the remaining 20 fragments (one from each speaker). This procedure of training and testing was repeated ten times, so that each fragment was used as a test fragment exactly once. Each fragment used in this study belonged to only one speaker, and therefore our tests are mono-classifications.

The order in which the training module of the classifier processes the fragments cannot be controlled and is entirely random. As a consequence, running the classifier twice does not usually lead to exactly the same results. We therefore ran each ten-fold-cross-validation 100 times. From the 100 performances of each classification test, we determined the lowest, highest, and average performance. The difference between the lowest and highest performance was on average 7.24% (range: 5.5% to 9.5%). Average performances are reported in Section 3. To compare the performance of different classification tests, we performed several unpaired t-tests on obtained performance scores (see below). As we performed multiple t-tests, we applied Bonferroni correction and used an alpha level of 0.0045.

To obtain information about which features contribute to the identification of speakers and thus in what aspects of choice and pronunciation of words speakers differ from each other, we manually inspected speaker profiles. For this manual inspection, we trained the classifier with the best performing combination of features (see below). Furthermore, we used all 200 fragments for training and thus obtained speaker profiles that are based on as much data as possible. As these speaker profiles are created by running the classifier only once, running it again will probably result in slightly different speaker profiles. To use only those features that are robust, i.e., likely to be part of the speaker profiles if we run the classifier again, we examined only those features that have an overall Winnow weight that is larger than the median weight, which was calculated separately for positive and negative weights. The positive overall weights ranged from 0.00025 to 2.54 with a median of 0.08, and the negative overall weights varied between -0.005 and -6.43 with a median of -0.15. We focused on the features that are characteristic for only a few speakers and thus provide information about differences between speakers. We therefore determined which features have a positive overall weight larger than the positive median for only one to four speakers and a negative weight larger than the negative median for 15 or more speakers.

Table 2. Confusion matrix with the number of classifications based on lexical features, uniphone features, and triphone features. Underlined numbers are combinations of speakers that participated in the same conversation.

Actual	Classified Speaker																			
Speaker	**A**	**B**	**E**	**Q**	**F**	**G**	**H**	**R**	**I**	**S**	**J**	**T**	**K**	**L**	**M**	**N**	**O**	**P**	**U**	**V**
A	647	100	0	0	0	46	75	55	0	17	7	0	49	0	0	0	3	0	0	1
B	0	830	0	9	40	1	29	0	0	0	0	0	24	10	0	56	0	0	0	1
E	0	2	437	281	25	0	77	13	118	0	9	21	13	0	1	0	0	0	0	3
Q	0	0	123	709	81	0	0	0	0	0	0	0	0	2	0	7	0	0	0	78
F	5	14	1	0	834	0	0	0	0	0	31	0	1	0	0	15	0	0	0	99
G	4	0	0	0	1	980	0	0	0	2	12	0	0	0	0	0	0	1	0	0
H	99	1	0	0	0	0	877	21	0	0	0	0	0	0	0	0	0	0	2	0
R	0	0	51	0	0	1	33	888	0	0	24	0	0	0	0	0	0	0	3	0
I	0	13	7	10	1	0	9	0	564	134	0	34	46	47	0	0	30	1	1	103
S	38	39	27	0	1	7	2	0	0	761	0	86	0	0	0	0	10	8	21	0
J	0	0	0	0	0	0	25	0	0	0	969	0	0	0	6	0	0	0	0	0
T	0	0	0	4	0	0	1	0	0	0	0	984	0	0	0	0	0	11	0	0
K	0	0	0	0	9	48	2	10	7	0	44	36	669	167	6	0	2	0	0	0
L	0	25	0	0	0	0	0	5	16	1	14	0	12	924	1	2	0	0	0	0
M	0	71	0	0	0	0	0	0	0	0	0	0	100	0	775	2	0	0	52	0
N	8	14	1	0	0	0	0	0	0	0	15	0	1	0	0	959	0	2	0	0
O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1000	0	0	0
P	0	0	0	0	0	100	0	0	0	0	0	77	0	0	0	0	1	822	0	0
U	0	3	0	1	39	0	0	67	0	0	0	3	40	2	13	0	0	0	735	97
V	0	0	0	8	0	36	6	0	84	0	0	2	73	0	0	6	0	98	79	608
Total	801	1112	647	1022	1031	1219	1136	1059	789	915	1125	1243	1028	1152	802	1047	1046	943	893	990

3 Results and discussion

As it is only meaningful to investigate differences between speakers with a well-performing classifier, we first investigated how well our best classifier performs in general. We therefore examined how often it correctly classified a fragment. Moreover, we investigated whether speakers in the same conversation were often confused with each other, which may be due to the topic of the conversation or to interactive speech alignment (Pickering and Garrod, 2004). Thereafter, we investigated the relevance of the different types of features by comparing classifiers trained and tested with various combinations of feature sets. Furthermore, for each feature type, we examined which features are especially relevant for characterizing individual speakers. As a performance measure for these classifiers, we used the harmonic means of their precision and recall (F_1), presented in Table 3.

3.1 General performance

The best performing classifier made use of both lexical features and of the uniphone and triphone features. The average performance of this classifier is as high as 79.86%, which may be surprising given the homogeneity of the group of speakers. As shown in the confusion matrix (Table 2), the percentage

of correct classifications for individual speakers ranged from 43.7% to 100%. Some speakers (e.g., Speakers J, N, and O) were seldom confused with other speakers. In contrast, a substantial number of fragments were incorrectly attributed to Speakers G and T, as were fragments of Speakers E and I to other speakers. Apparently, some speakers were more difficult to classify than others. Note that the classifier was still able to correctly classify the more difficult speakers well above chance level (i.e., 5%), suggesting that they were not indistinctive in their word choice or pronunciations.

We examined whether speakers who participated in the same conversation were more often confused with each other than with other speakers. Speakers in the same conversation discuss the same topics, which inevitably results in the use of the same content words. Moreover, several studies have shown that speakers tend also to align their speech on other levels, including syntactic and phonological levels (interactive speech alignment, e.g., Pickering and Garrod, 2004). As a consequence, speakers in the same conversation may be more confused with each other than with other speakers. This was only the case for six out of the twenty speakers. As shown by the underlined numbers in Table 2, Speakers A, E, I, K, Q, and U were more often classified as the other speaker in the conversation than as speakers from other conversations. Interestingly, the partners for Speakers A, I, K, and U (i.e., Speakers B, S, L, and V respectively) were not more often classified as the other speaker in the conversation, which suggests that these speakers also display more idiosyncratic properties.

Only the speakers of the pair E-Q were often confused with each other. Twenty-one percent of these two speakers' top 50 positive features concern a content word, which is either *eens* 'once', *is* 'is', *ja* 'yes', *kunnen* 'can', *maken* 'make', *natuurlijk* 'of course', *nu* 'now', *vind* 'find', *weet* 'know', or *wil* 'want'. These content words are not very informative about what the speakers talked about. It is therefore unlikely that confusion between Speakers E and Q is the result of the topic of the conversation. It probably results from alignment at various linguistic levels or coincidental similarities between these speakers.

3.2 Lexical features

In order to investigate differences in speakers' choice of words, we first trained and tested the classifier on the basis of lexical features only, namely uni- and bigrams. These tests are presented in the first three rows of Table 3. T-tests showed that including both unigrams and bigrams resulted in a significantly better performance than using only unigrams ($t(198.0) = -104.4$, $p < 0.0001$) or bigrams ($t(197.5) = -69.6$, $p < 0.0001$). The performance with both uni-

Table 3. Averaged harmonic means (F_1) obtained in our classification tests with different sets of features. A plus sign indicates that the feature type was included in training and testing the classifier. For the lexical features, the columns *Uni.* and *Bi.* denote the unigram and bigram features, respectively. For the pronunciation features, the columns *Uni.*, *Bi.prec.*, *Bi.foll.*, and *Tri.* denote the uniphone features, target phone features with the preceding phone, target phone features with the following phone, and target phone features with both surrounding phones, respectively. The column *Word* denotes the word pronunciation features.

	Lexical		Pronunciation					F_1
	Uni.	Bi.	Uni.	Bi.prec.	Bi.foll.	Tri.	Word	
1	+							51.7%
2		+						59.6%
3	+	+						73.9%
4	+	+	+					75.5%
5	+	+		+				77.9%
6	+	+			+			76.2%
7	+	+				+		79.2%
8	+	+	+	+				79.0%
9	+	+	+		+			77.3%
10	+	+	+			+		79.9%
11	+	+	+	+	+			77.8%
12	+	+		+	+			77.4%
13	+	+		+	+	+		79.1%
14	+	+	+	+	+	+		78.9%
15	+	+					+	75.9%
16	+	+	+				+	76.3%
17	+	+		+			+	75.7%
18	+	+			+		+	77.7%
19	+	+				+	+	78.0%
20	+	+	+	+			+	76.2%
21	+	+	+		+		+	78.0%
22	+	+	+			+	+	78.4%
23	+	+	+	+	+		+	75.9%
24	+	+		+	+		+	75.8%
25	+	+		+	+	+	+	78.9%
26	+	+	+	+	+	+	+	78.9%

Table 4. The lexical features that are characteristic for only a few speakers. Within a bigram, ∅ indicates a silence. Numbers indicate the numbers of speakers for which the features were characteristic / uncharacteristic.

unigrams:
ben 'am' (1 / 15), *bij* 'by' (1 / 19), *goed* 'good' (2 / 17), *hè* 'isn't it?' (4 / 15), *heel* 'very' (3 / 16), *jij* 'you' (4 / 15), *mij* 'me' (3 / 15), *nee* 'no' (3 / 16), *om* 'to' (1 / 18), *toch* 'still' (2 / 16), *want* 'because' (2 / 16), *we* 'we' (1 / 18), *ze* 'they' (1 / 18), *zo* 'so' (1 / 19)
bigrams:
∅ *de* '∅ the' (4 / 16), ∅ *dus* '∅ so' (1 / 16), ∅ *een* '∅ an' (1 / 19), ∅ *nee* '∅ no' (2 / 16), *dat* ∅ 'that ∅' (2 / 18), *een* ∅ 'an ∅' (1 / 19), *en dan* 'and then' (4 / 15), *het* ∅ 'it ∅' (2 / 15), *in de* 'in the' (4 / 15), *maar* ∅ 'but ∅' (4 / 16), *niet* ∅ 'not ∅' (2 / 15), *nou* ∅ 'well ∅' (3 / 15)

grams and bigrams was approximately 74% (row 3), indicating that speakers greatly differ in their choice of words.

Table 4 shows features that are characteristic for only one to four speakers and thus provide information about differences between speakers. The majority of the characteristic bigrams contain a silence (e.g., ∅ *dus* '∅ so'), indicating that speakers differ especially in which words they produce directly before or after a pause. Furthermore, the majority of the features are function words (e.g., *want* 'because' and *we* 'we') and those that are content words are highly frequent and semantically relatively weak (e.g., *goed* 'good' and *nee* 'no').

To further investigate the contribution of content and function words to the classification performance, we also trained and tested the classifier on unigrams only including either of these word types. Based on function words only, we obtained a performance of 26.0%, whereas classification based on content words only resulted in a performance of 39.5%. Both results were significantly lower than the classification performance of 51.7% based on all unigrams (function words: $t(198.0) = 121.2$, $p < 0.0001$; content words: $t(196.8) = 59.6$, $p < 0.0001$), suggesting that neither of the word types can solely account for the performance based on all unigrams. Interestingly, classification based on content words performed significantly better than the classification based on function words ($t(197.2) = 66.0$, $p < 0.0001$). Importantly, the number of features cannot explain this difference in classification performance, since the number of content word features (27,381) was lower than the number of function word features (51,377). These results suggest that speakers differ from each other especially in their use of semantically weak content words, such as discourse markers.

3.3 Phone features

We refrained from testing models with pronunciation features only, as these would mainly signal lexical variation instead of pronunciation variation. For instance, when using this type of classifier, if we found that the word pronunciation variant /tyk/ for *natuurlijk* 'of course' is characteristic for a few speakers, this would probably be because these speakers more often produced this word and not because they pronounced it differently from the other speakers. All our classifiers therefore included lexical features.

We first combined the lexical features with phone features. The best performance was obtained with the classification including the uniphone and triphone features (compare row 10 of Table 3 to rows 4 to 9 and 11 to 14). This set of features improved performance by approximately 6% compared to the classification that was based on lexical features only (row 3 vs. 10: $t(197.9) = -28.4$, $p < 0.0001$).

As classification including biphones resulted in lower performances than classification including triphones, information about either neighboring segment is apparently less helpful than information about both neighboring phones. The probable explanation is that the exact pronunciation of a phone depends on both the preceding and following phone (e.g., /ə/ may especially be absent after /x/ and before a sonorant); both neighboring phones are therefore necessary for variation to be meaningful.

Table 5 presents the phone features that are characteristic for only a few speakers. There are two reasons for why we cannot immediately conclude that these differences among speakers reflect individual differences in reduction. First, if for these phones only one variant is incorporated as a phone feature, speaker differences in the use of these phone features cannot represent individual differences in pronunciation; they rather show that speakers differ in how often they produce words with these phones. Second, if speakers differ in the words in which these phones occur, differences among these speakers may reflect differences in word choice rather than in phone realization. Research has shown that a segment is more likely to be reduced in one word than in another depending on the word's frequency of occurrence, the preceding or following segment, and whether the segment carries stress (e.g., Bell et al., 2009; Cho and McQueen, 2005; Mitterer and Ernestus, 2006).

Two-thirds of the features in Table 5 represent unreduced phones with their neighboring phones. For 51.1% of these unreduced characteristic features (which is 35.4% of all features in Table 5), counterparts with an absent or replaced phone do not occur frequently enough in the corpus to be included in the phone features. For instance, the unreduced feature /ty͡r/ was charac-

Table 5. The uniphone and triphone features that are characteristic for only a few speakers. The target phone is underlined and a silence is indicated by ∅. A replacement of phone A by phone B is denoted by A → B.

unreduced phones:
ø, xət, kə**∅**, pət, təf, təp, mal, **∅**ɑl, kɑn, tɑl, wɑn, nɑʊ**∅**, xen, mer, wet, pɛn, tɛn, tit, nɪk, tɪk, son, kɔm, xʊt, tyr, ɛɪxə, ɑlə, ɑls, elə, ant, ɑns, enp, ens, ɛns, ɪnə, ɛrk, ɔrt, nsə, ətØ, ətə, atə, its, xtə, ltə, ntØ əwo,
absent phones:
x, i, s, xəl, xəw, jət, təØ, Ø**ɪ**k, ələ, ɑls, ens
replaced phones:
a → ə, ɑ → ə, o → ə, a → ɑ, nət → nɛt, Øen → Øən, ren → rən, sen → sən, ten → tən

teristic for Speaker I, who produced it in the words /natyrlək/ *natuurlijk* 'of course', /prosədyrə(s)/ *procedure(s)* 'procedure(s)', /litərɑtyr/ *literatuur* 'literature', /dyr/ *duur* 'expensive', /dyrə/ *dure* 'expensive', /dyrt/ *duurt* 'lasts', and /dyrdə/ *duurde* 'lasted' (remember that voiced and voiceless obstruents have been collapsed in the pronunciation features). Its counterpart in which /y/ is absent occurs only twice in the entire corpus (both occurrences concern an inflection of /kʏltyrel/ *cultureel* 'cultural' produced by Speaker K) and the replacement of this segment occurs only once (in the pronunciation [natʏr] /natyr/ *natuur* 'nature' produced by Speaker H). Since this phone sequence is only represented with its full pronunciations in the feature sets, its relevance cannot be attributed to pronunciation variation but probably results from variation in word choice.

In contrast, the remaining 48.9% of the unreduced characteristic features appear to represent pronunciation variation. For example, an unreduced schwa in the sequence /xət/ is characteristic for Speakers P and T, who produced it mainly in /xədan/ *gedaan* 'done', /zɛxə/ *zeggen* 'say' followed by /t/ or /d/, /texə/ *tegen* 'against' followed by /t/ or /d/, and /xədeltə/ *gedeelte* 'part'. In contrast, the absence of this schwa is characteristic for Speakers K, L, P, R, and V. These speakers produced the sequence /xət/ mainly in the words /xədan/ *gedaan* 'done', /zɛxə/ *zeggen* 'say' followed by /t/ or /d/, and /xətɑl/ *getal* 'number'. As the sequence /xət/ occurs in approximately the same words for the two speaker groups, the difference in pronunciation between these two groups is not likely the result of differences in word choice but in genuine pronunciation variation.

A second example represents the unreduced pronunciation of the sequence /ɑlə/, which is characteristic for Speaker V only. He produced this sequence in the words /ɑlə/ *alle* 'all', /ɑləs/ *alles* 'everything', /ɑlərlɛɪ/ *allerlei* 'all kinds of', /ɑləmal/ *allemaal* 'all', /antɑlə/ *aantallen* 'numbers', and /xəvɑl/

geval 'case' followed by a schwa. The counterpart in which /l/ was absent was characteristic for six other speakers, namely Speakers E, H, M, P, R, and T. These speakers produced this counterpart in the word /ɑləmal/ *allemaal* 'all' only. Speaker V therefore differs from these other speakers in how he pronounces the sequence /ɑl̲ə/ in the semantically weak word *allemaal*.

The other third of all phone features in Table 5 represent absent and replaced phones in approximately the same numbers. All but one (i.e., sə̲n → sə̲n) of these features have unreduced counterparts in the feature set and are therefore likely to represent genuine pronunciation variation. The absence of schwa in /jə̲t/ is an example of a feature that clearly represents pronunciation variation. It is characteristic for Speaker H and uncharacteristic for sixteen other speakers. Sixteen of this feature's occurrences (84.2%) represent the word /jə/ *je* 'you' followed by a /t/ or /d/, while the remaining occurrences represent the word types /spœytjə/ *spuitje* 'little syringe', /festjə/ *feestje* 'little party', and /betjə/ *beetje* 'a little' followed by /t/ or /d/. The unreduced pronunciation of /jə̲t/ was characteristic for five different speakers (i.e., Speakers E, F, K, L, and Q) and uncharacteristic for ten speakers (i.e., B, H, I, M, N, O, R, S, T, and U). These five speakers produced this sequence also mainly in the word /jə/ *je* 'you' followed by a /t/ or /d/ (83.9%). This indicates that the absence of schwa in /jə̲t/ represents pronunciation variation, primarily in the semantically weak word *je*, and is not likely to result from lexical choice. Our data thus show that our socially homogeneous group of speakers are not at all homogeneous in their pronunciation of the phone sequence /jə̲t/.

Since some phone features appear to represent occurrence of phone sequences rather than pronunciation variation, the question arises whether genuine pronunciation variation really contributes to speaker identification. In order to answer this question, we ran an additional classifier based only on lexical features and the full pronunciations of the phone features. For instance, the uniphone feature *absence of* /x/ was converted to *unreduced* /x/ and *replacement of* /a/ *by* /ə/ was replaced by *unreduced* /a/. All unreduced features remained unchanged. This classifier thus contained no features representing pronunciation variation. Consequently, if the best performing classifier so far uses information about pronunciation variation, the classifier without pronunciation variation should perform worse. The performance of the classifier without pronunciation variation was 70.0% (not incorporated in Table 3), which is significantly less than the classification based on lexical features and uniphone and triphone features (row 10 of Table 3; $t(198.0) = -48.1$, $p < 0.0001$). Interestingly, this performance is also worse than the performance of the classifier based on lexical features only (row 3; $t(197.9) = -19.0$, $p < 0.0001$). We con-

clude that pronunciation variation contributes to classification, and speakers thus differ in how they pronounce phones and phone sequences.

Interestingly, the classifier achieved better performance if it was not only based on lexical and triphone features but also on uniphone features (Table 3 row 7 vs. 10: $t(197.9) = -3.4$, $p < 0.001$). This suggests that variation in the single phones, regardless of the context, also reflects speaker-specific behavior. Most of the relevant uniphones represent reductions (see Table 5). For instance, the absence of /x/ is characteristic for Speaker V. This speaker did not produce /x/ 33 times out of 422 times, and this was mainly in the semantically weak words /nɔx/ *nog* 'yet' and /tɔx/ *toch* 'still' (also in /xəxan/ *gegaan* 'went', /ɛɪxələk/ *eigenlijk* 'actually', /ɔnxəver/ *ongeveer* 'approximately'). As only one of the characteristic uniphones is an unreduced pronunciation, we conclude that speakers hardly differ in how often they produce most single phones. In contrast, they differ in how they reduce single phones.

Comparison of the relevant features that regard vowels and consonants (Table 5) showed that for both unreduced and absent uniphones and triphones, the numbers are approximately the same (i.e., unreduced: 24 vowels and 21 consonants; absent: 6 vowels and 5 consonants). This is unsurprising for the uniphones, given the similarity in the numbers of vowels and consonants among the unreduced and absent uniphones provided to the classifier (i.e., unreduced: 16 vowels and 15 consonants; absent: 14 vowels and 13 consonants). In contrast, the numbers of vowels and consonants differ among triphones in the input (i.e., unreduced: 450 vowels and 379 consonants; absent: 202 vowels and 149 consonants). The fact that approximately the same number of triphones with vowels and consonants characterize speakers therefore suggests that, for both the unreduced and absent segments, a larger proportion of the consonants than vowels is speakers specific.

Furthermore, the replaced phones that distinguish speakers from each other are all vowels. This is in line with the number of replaced vowel and consonant features in the input of the classifier: 22 uniphone and 47 triphone replacements concerned a vowel, but only 1 uniphone and 7 triphone replacements concerned a consonant which all concerned the pronunciation of [m] instead of /b/ or /p/. A likely explanation is that vowels are easily reduced to schwa, while consonants are not often replaced by another consonant (aside from the consonant that has the same characteristics except for voicing; e.g., /v/ is often replaced by /f/). As explained in the Method, variation in voicing is not specific for casual speech and was therefore ignored.

In conclusion, the classification including phone features showed that speakers differ in their pronunciations of single phones and of sequences of three phones. Interestingly, we found that several triphone features mainly

originate from semantically weak words, which shows that speakers differ in the pronunciations of these words.

3.4 Word pronunciation features

Finally, we ran a classifier that used all types of features, that is, lexical features, phone features, and word pronunciation features. Several words were represented by both lexical and word pronunciation features; we decided not to include both to avoid that a single token was represented several times (as a lexical and as a word pronunciation feature) and to only include its word pronunciation feature. For instance, in the classification tests without word pronunciation features, the word *mensen* 'people' was represented as unigram, in five bigrams (i.e., *mensen die* 'people who', *die mensen* 'those people', *de mensen* 'the people', *mensen* ∅ 'people ∅', and ∅ *mensen* '∅ people'), and in the phone features. As [mɛnsə], [mɛsə], [mɛns], and [mɛs] were part of the word pronunciation features, when including these, we replaced each occurrence of *mensen* that was produced as one of these variants by its actual pronunciation. If an occurrence of *mensen* was pronounced differently from the pronunciations that are part of the word pronunciation features (e.g., as [mɛnsən] or [mɛsn]), the occurrence was not replaced. For bigram features, this means that none of the words, both of the words, or either of the words could be replaced. This procedure gave a preference to the word pronunciation features over the lexical features and thus gave the word pronunciation features every chance to distinguish between speakers.

The best performance including word pronunciation features was achieved with the combination of all lexical features, phone features, and word pronunciation features (78.9%; final row of Table 3). Importantly, this performance is significantly worse than the best performance with lexical and phone features only (row 10; $t(196.9) = 4.9$, $p < 0.001$). Hence, the addition of word pronunciation features does not result in an improvement. One explanation may be that speakers do not differ in their pronunciations for complete words. This is however unlikely as classification based on lexical and word pronunciation features only (i.e., without any phone features) resulted in an improved performance compared to classification with only lexical features (row 15 vs. row 3 of Table 3; $t(186.5) = -10.3$, $p < 0.0001$). An alternative and more probable explanation is that the variation in the pronunciation of entire words is already captured by the phone features. This is supported by the triphone features discussed in the previous section which upon closer inspection appeared to represent a small number of words (e.g., reduced /jə̲t/ mainly represented

the word *je* 'you' followed by /t/ or /d/ and unreduced /ɑl̲ə/ represented mainly the word *allemaal* 'all'). Moreover, all words are also represented by uniphone and triphone sequences. For example, the pronunciation [mɛs] for /mɛnsə/ *mensen* 'people' is presented as a word pronunciation variant, but is also represented in the uniphones /m/, /ɛ/, absence of /n/, /s/, and absence of /ə/, and in the triphones /?m̲ɛ/ (in which ? indicates any possible preceding segment), /mɛ̲n/, absence of /n/ in /ɛn̲s/, /ns̲ə/, and absence of /ə/ in /sə̲?/ if these features met the restrictions described in the Method. The phone features outperformed the word pronunciation features probably because they additionally capture variation that spans word boundaries (e.g., resulting from cross-word assimilation), and they therefore contain more information about the speaker's pronunciation habits than word pronunciation features. The addition of features that contribute little to discrimination, because they largely covariate with features already in the model, often lowers performance of several feature-based classifiers.

4 General discussion

This paper investigated whether individual speakers sampled from a socially homogeneous group differ in their choice and pronunciation of words when engaged in casual conversations, and if so how. We studied the homogeneous group of twenty male speakers of the ECSD and tested whether a classification algorithm was able to distinguish between these speakers. In order to focus on the speakers' choice and pronunciation of words rather than characteristics of the speech signal (including voice characteristics), we trained and tested the classifier on the basis of features extracted from hand-made orthographic and from automatically generated broad phonetic transcriptions. We hypothesized that if the classifier was able to distinguish between these speakers, they have to differ in their choice and pronunciation of words. To study how speakers differ from each other, we inspected which features in the speaker profiles created by the classifier were relevant for classification.

Our classification tests based on only lexical features resulted in a high performance (73.9%), indicating that the speakers differed in the words they used. Two types of words appeared to be relevant in distinguishing between speakers. The first are function words (e.g., *want* 'because' and *dat* 'that'), which is as expected as they are often regarded as useful features in research on authorship attribution (e.g., Koppel et al., 2009; Stamatatos, 2009). The second type are highly frequent content words that are semantically relatively

weak (e.g., *goed* 'good' and *no* 'nee'). In authorship attribution, content words are often argued to be topic dependent and thus less suitable for distinguishing between writers. Our results suggest that including semantically weak content words may provide more information than including function words only and may thus be beneficial.

Classification tests including features that represent pronunciation variation typical for casual speech performed better than tests based on lexical features alone. However, inspection of the relevant pronunciation features showed that some of them probably represent lexical information rather than pronunciation variation. Others represent pronunciation variation and, importantly, they are the ones responsible for the increase in classifier performance.

These pronunciation features show that speakers differ in how often they reduce certain phones and phone sequences. Closer inspection of the speaker specific triphone features showed that some of them mainly originate from a few semantically weak words, including *allemaal* 'all' and the pronoun *je* 'you'. This suggests that speakers differ in how they realize phones not only given the immediate phone context but also given the carrier word. Moreover, it shows that semantically weak words contribute to speaker classification at two levels: speakers differ in their use and pronunciation of these words.

Interestingly, the performance of classification with uniphones and triphones was higher than that of classifications including biphones. Apparently, the pronunciation of single phones by themselves or with two neighboring segments is more informative than the pronunciation of these phones given only one neighboring segment. A likely explanation is that generally a phone's pronunciation does not only depend on one of the neighboring phones but on both neighbors. Moreover, as mentioned above, speakers differ from each other in how much they reduce certain semantically weak words and a given word is better identified by a triphone than by a biphone.

Another type of pronunciation feature that did not increase classification performance is formed by the word pronunciations. In the classification including pronunciation features at the word level, we replaced lexical features that were also represented as word pronunciation features by their actual pronunciation. By doing so, we favored the word pronunciation features. Nevertheless, all classifications including word pronunciation features performed worse than the classification with lexical features and uniphone and triphone features. Probably, the triphone features contained all information present in the word pronunciation features, in addition to pronunciation variation spanning word boundaries.

Previous research has shown that speakers participating in the same conversation tend to align their speech at for instance lexical, syntactic, and

phonological levels (e.g., Pickering and Garrod, 2004). We expected that this alignment would demonstrate itself in that speakers within a conversation would be more often confused with each other than with other speakers in general. This expectation was borne out for only 30% of the speakers. This low number may indicate that the properties of the speech produced by a given speaker at a given moment is colored more by idiosyncratic speech habits than by processes of speech alignment with the conversation partner.

Van Bael and Van Halteren (2007) investigated classification of groups of speakers on the basis of phonetic transcriptions and reported that classification based on pronunciation features performed poorly. The authors discussed several possible explanations for this finding. Our results indicate which of their explanations is most likely. First, the authors noted that their pronunciation feature set may have been too small as it contained only 94 pronunciation features. This is a likely explanation since we used many more pronunciation features and our classification improved when we added any type of pronunciation features to the lexical features (see Table 3). A second explanation provided by Van Bael and Van Halteren (2007) concerns the heterogeneity within their classes. They classified speakers in terms of social groups defined by, for instance, regional background and age. As a consequence, one class contained multiple speakers who may have had different pronunciation habits. Our findings suggest that this is also a highly probable explanation, as our classifier was able to distinguish between speakers within the same social group.

Van Bael and Van Halteren (2007) also suggested that their phonetic transcriptions may not have been sufficiently detailed to capture pronunciation differences among speakers. Since the transcriptions that we used were also broad phonetic transcriptions without any fine-phonetic detail, our results suggest that this is an unlikely explanation. However, we agree with these authors that individual differences may be larger if also fine-phonetic detail is taken into account. Whereas our study only investigated whether speakers differ in which segments they realize or substitute by other segments (i.e., categorical reduction), previous research has shown that reduction may also be gradient in nature (e.g., Browman and Goldstein, 1990; Davidson, 2006; Hanique et al., 2013; Torreira and Ernestus, 2010): only a part of a segment may be reduced or segments may not be reduced sufficiently to be identified as different phones. Future studies focusing on both categorical and gradient reduction may report larger individual differences.

Currently, psycholinguistic models offer explanations of how the average speaker produces his speech. Our finding that individual speakers differ in their choice and pronunciation of words should to be incorporated in these models.

For instance, in a production model like WEAVER++ (Levelt et al., 1999), speakers should differ in the resting activation levels of words and perhaps even of pronunciation variants. Exemplar-based production models, such as the one described by Goldinger (1998), should assume that speakers differ in their number of exemplars for some words and pronunciation variants. Furthermore, speech comprehension models should assume that a listener adapts to the specific pronunciations of the speaker he is listening to.

In conclusion, the speakers that we investigated belonged to a homogeneous group, and may therefore be expected to show similar speech habits. Nevertheless, our classification tests showed that these speakers differ in the words they use, as well as in how they pronounce the words in casual conversations. Individual differences between speakers' pronunciations can be observed, even if these speakers have the same social background.

Acknowledgement: The authors would like to thank Eva d'Hondt for her help with the classifier and the anonymous reviewer for useful comments. This work was partly funded by a European Young Investigator Award and an ERC starting grant (284108) to Mirjam Ernestus.

References

Barlow, Michael. 2010. Individual usage: a corpus-based study of idiolects. In *Proceedings of LAUD Conference*, Landau, Germany.

Bell, Alan, Jason M. Brenier, Michelle Gregory, Cynthia Girand & Dan Jurafsky. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60. 92–111.

Booij, Geert. 1995. *The Phonology of Dutch*. Oxford: Clarendon Press.

Browman, Catherine P. and Louis Goldstein. 1990. Tiers in articulatory phonology with some implications for casual speech. In John Kingston & Mary E. Beckman (eds.), *Between the grammar and physics of speech [papers in Laboratory Phonology 1]*, 341–376. Cambridge: Cambridge University Press.

Cho, Taehong & James M. McQueen. 2005. Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress. *Journal of Phonetics* 33. 121–157.

Dagan, Ido, Yael Karov & Dan Roth. 1997. Mistake-driven learning in text categorization. In *Proceedings of 2nd Conference on Empirical Methods in Natural Language Processing*, 55–63. Providence, RI.

Davidson, Lisa. 2006. Schwa elision in fast speech: Segmental deletion or gestural overlap? *Phonetica* 63. 79–112.

Ebrahimpour, Maryam, Talis J. Putnin, Matthew J. Berryman, Andrew Allison, Brian W.-H. Ng & Derek Abbott. 2013. Automated authorship attribution using advanced signal

classification techniques. *PLOS ONE* 8. 1–12.

Ernestus, Mirjam. 2000. *Voice assimilation and segment reduction in casual Dutch. A corpus-based study of the phonology-phonetics interface*. PhD thesis. Utrecht: LOT.

Ernestus, Mirjam & R. Harald Baayen. 2011. Corpora and exemplars in phonology. In John A. Goldsmith, Jason Riggle, & Alan C.L. Yu (eds.), *The Handbook of Phonological Theory (2^{nd} edn.)*, 374–400. Chichester, West Sussex: Wiley-Blackwell.

Ernestus, Mirjam, Mybeth Lahey, Femke Verhees & R. Harald Baayen. 2006. Lexical frequency and voice assimilation. *Journal of the Acoustical Society of America* 120. 1040–1051.

Ernestus, Mirjam & Natasha Warner. 2011. An introduction to reduced pronunciation variants. *Journal of Phonetics* 39. 253–260.

Goldinger, Stephen D. 1998. Echoes of echoes? An episodic theory of lexical access. *Psychogical Review* 105. 251–279.

Guy, Gregory R. 1980. Variation in the group and the individual: The case of final stop deletion. In William Labov (ed.), *Locating language in time and space*, 1–36. New York: Academic Press.

Hämäläinen, Annika, Michele Gubian, Louis ten Bosch & Lou Boves. 2009. Analysis of acoustic reduction using spectral similarity measures. *Journal of the Acoustical Society of America* 126, 3227–3235.

Hanique, Iris, Mirjam Ernestus & Barbara Schuppler. 2013. Informal speech processes can be categorical in nature, even if they affect many different words. *Journal of the Acoustical Society of America* 133. 1644–1655.

Johnson, Keith. 2004. Massive reduction in conversational American English. In *Proceedings of the workshop on Spontaneous Speech: Data and Analysis*, 29–54. Tokyo, Japan.

Keune, Karen, Mirjam Ernestus, Roeland van Hout & R. Harald Baayen. 2005. Social, geographical, and register variation in Dutch: From written 'mogelijk' to spoken 'mok'. *Corpus Linguistics and Linguistic Theory* 1. 183–223.

Kipp, Andreas, Maria-Barbara Wesenick & Florian Schiel. 1997. Pronunciation modeling applied to automatic segmentation of spontaneous speech. In *Proceedings of Eurospeech-1997*, 1023–1026, Rhodes, Greece.

Koppel, Moshe, Jonathan Schler & Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology* 60. 9–26.

Koster, Cornelis H.A. & Beney, Jean G. (2009). Phrase-based document categorization revisited. In *Proceedings of the PAIR workshop at CIKM-2009*, 49–55. Hong Kong, China.

Koster, Cornelis H.A., Marc Seutter & Jean G. Beney. 2003. Multi-classification of patent applications with winnow. In Manfred Broy & Alexandre V. Zamulin (eds.), *Lecture Notes in Computer Science 2890*, 545–555. Berlin / Heidelberg: Springer.

Levelt, Willem J.M., Ardi Roelofs & Antje S. Meyer. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22. 1–38.

Littlestone, Nick. 1988. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning* 2. 285–318.

Mitterer, Holger & Mirjam Ernestus. 2006. Listeners recover /t/s that speakers reduce: Evidence from /t/-lenition in Dutch. *Journal of Phonetics* 34. 73–103.

Oostdijk, Nelleke. 2002. The design of the Spoken Dutch Corpus. In Pam Peters, Peter Collins & Adam Smith (eds.), *New Frontiers of Corpus Research*, 105–112. Amsterdam:

Rodopi.
Phillips, Betty S. 1994. Southern English glide deletion revisited. *American Speech* 69, 15–127.
Pickering, Martin J. & Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences* 27. 169–226.
Salton, Gerard & Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing Management* 24. 513–523.
Schuppler, Barbara, Mirjam Ernestus, Odette Scharenborg & Lou Boves. 2011. Acoustic reduction in conversational Dutch: A quantitative analysis based on automatically generated segmental transcriptions. *Journal of Phonetics* 39. 96–109.
Stamatatos, Efstathios. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* 60. 538–556.
Strik, Helmer, Joost van Doremalen & Catia Cucciarini. 2008. Pronunciation reduction: How it relates to speech style, gender, and age. In *Proceedings of Interspeech 2008*, 1477–1480. Brisbane.
Torreira, Francisco & Mirjam Ernestus. 2010. Phrase-medial vowel devoicing in spontaneous French. In *Proceedings of Interspeech 2010*, 2006–2009. Makuhari, Japan.
Van Bael, Christophe & Hans van Halteren. 2007. Speaker classification by means of orthographic and broad phonetic transcriptions of speech. In Christian Müller (ed.), *Speaker Classification II*, Lecture Notes in Computer Science, 293–307. Berlin / Heidelberg: Springer.
Van der Sijs, Nicoline. 2002. *Chronologisch woordenboek. De ouderdom en herkomst van onze woorden en betekenissen (Chronological dictionary. The age and origin of our words and meanings)*, 341–518. Amsterdam / Antwerpen: Veen.
Van Son, Rob J.J.H., Diana Binnenpoorte, Henk van den Heuvel & Louis C.W. Pols. 2001. The IFA corpus: A phonemically segmented Dutch 'open source' speech database. In *Proceedings of Eurospeech 2001*, 2051–2054. Aalborg, Denmark.
Young, Steve, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying (Andrew) Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev & Phil Woodland. 2002. *The HTK Book 3.2*. Cambridge: Entropic.

Received ...; accepted ....