Multiword Expressions in Spontaneous Speech:
Do we really speak like that?

H. Strik, D. Binnenpoorte, C. Cucchiarini

Centre for Language and Speech Technology (CLST)
Radboud University Nijmegen, the Netherlands
[h.strik,d.binnenpoorte,c.cucchiarini]@let.ru.nl

Abstract

In this study, we examined the pronunciation characteristics of multiword expressions (MWEs). We first drew up an inventory of frequently occurring N-grams extracted from orthographic transcriptions of spontaneous speech contained in a large corpus of spoken Dutch. For about 10% of these N-grams phonetic transcriptions were available, which were examined. Our results show that the pronunciation of these N-grams differed to a large extent from the canonical form. In order to determine whether this is a general characteristic of spontaneous speech or rather the effect of the specific status of these N-grams, we analyzed the pronunciations of the individual words composing the N-grams in two context conditions: 1) in the N-gram context and 2) in any other context. We found that words in N-grams do indeed have peculiar pronunciation patterns. This seems to suggest that these N-grams may be considered as MWEs that should therefore be treated as lexical entries with their own specific pronunciation variants in the pronunciation lexicons used for e.g. automatic speech recognition (ASR) and automatic phonetic transcription (APT).

1. Introduction

In [1] it is mentioned that multiword expressions (MWEs) are "Expressions consisting of multiple words for which at least one aspect (syntactic, semantic, pragmatic, translation, etc.) is not predictable from the individual words and their normal mode of combination. Therefore, such expressions and their unpredictable properties must be stored in a lexicon.". MWEs have been studied in theoretical linguistics [2, 3], and more recently also in the field of Natural Language Processing (NLP) [1, 4, 5, 6]. So far, most of the research on MWEs has concerned their extraction and handling in written rather than in spoken language. However, it seems that pronunciation might be one of those aspects of MWEs that are not predictable from the individual words. The aim of the present paper is to investigate whether this is indeed the case.

In this paper, we study pronunciation variation in MWEs in the Spoken Dutch Corpus (CGN: Corpus Gesproken Nederlands) [7]. We focus on spontaneous speech, because we think that the problem of pronunciation variation in MWEs is most acute in this style. Speech recognition performance for spontaneous speech is far below the performance for read speech [8], and there are indications that a large proportion of the performance gap is due to the inability to effectively model pronunciation variation in spontaneous speech [9].

One way to model pronunciation variation is to add pronunciation variants to the lexicon. However, it has been found that adding variants of individual words to the ASR lexicon becomes counter-productive as soon as the average number of variants per word exceeds a threshold of about 2.5 [10, 11]. At the same time, it appears that adding frequent N-grams to the lexicon, and treating these as words with their own specific pronunciation variants improves ASR performance (for an overview, see [9]). However, in the above mentioned studies, the notion of MWE is mainly deployed for the benefit of reducing word error rate in ASR. In the present paper, we investigate to what extent the pronunciation of the words in MWEs differs from the pronunciation of the same words in other contexts.

In our research, we first extracted frequent word sequences, which, for convenience, will hereafter be referred to as MWEs (see section 3). Then we proceeded to a more detailed analysis of MWEs in which we focused on reduction phenomena (see section 4).

2. Material

The MWEs were extracted from the Spoken Dutch Corpus (CGN) [7], a database containing about 9 million words of contemporary Dutch as spoken in the Netherlands and Flanders. In this study, we focused on 3,301,503, Northern-Dutch words recorded in lessons (LS), spontaneous dialogues (SD), and spontaneous telephone conversations (ST). These words and their orthographic transcriptions constitute *Corpus 1*. Manually generated phonetic transcriptions were available for 333,284 of these words. They make up *Corpus 2*. For more details see [12].

3. Experiment 1

3.1. Method

First, the orthographic transcriptions of *Corpus 1* were used to study how many and what sort of MWEs are present in spontaneous spoken Dutch in *Corpus 1*. We selected N-grams that meet the following criteria:

- *Length: $3 \leq N \leq 6$*. Given the size of the corpus, we did not expect to find frequent sequences longer than six words. For theoretical and practical reasons, we decided to omit bigrams. First, many frequent bigrams were part of frequent N-grams with $N > 2$; we can observe and analyze their pronunciation variation even if we do not include bigrams. Second, the number of frequent

bigrams was extremely large, and the sheer number complicates analysis considerably.

- *Minimum frequency*. For this we used the criterion proposed in Chapter 13 of [13], who stated that expressions containing three or four words should have a minimal frequency of 10 per million words, and expressions containing more than four words should have 5 or more occurrences per million words. In our case, with a source text of 3.3 M words, we require the frequency of a unique 3-gram or 4-gram to be at least 30, and the frequency of a unique 5-gram or 6-gram at least 15.

- *Uniform presence*. The sequence should have a high frequency in all the three sub-corpora (LS, SD, ST). This stipulation removes sequences such as 'een twee drie vier' ('one two three four'), which are frequent in the SD sub-corpus, due to the fact that the speakers were encouraged to play games to keep the conversation going.

- *Contiguous*. We expected more pronunciation variation in contiguous sequences due to assimilation and degemination processes. For similar reasons, we also excluded word sequences that straddle a deep syntactic boundary.

- *Fluent*. Finally, we also excluded all N-grams containing disfluencies, hesitations, filled pauses, repetitions, and (speaker) noise or unintelligible speech.

3.2. Results

Table 1 summarizes the results of the MWE extraction on *Corpus 1*. It can be seen that the number of types and tokens, as well as the token/type ratio decrease as the sequences grow longer. The number of types would have been much larger if we had not applied the 'uniform presence' criterion, which removed many sequences from the subcorpus of face-to-face dialogs that were directly related to playing card or board games. Removing setting-specific types resulted in a large increase in the average token/type ratio.

Table 1: Number of types, tokens and the token-type ration of the selected N-grams

	3-gram	4-gram	5-gram	6-gram
#types	3,015	247	48	1
#tokens	217,230	13,495	1,285	19
ratio	72.1	54.6	26.7	19

From Table 1 it can be deduced that the 3,311 N-gram types (3,015+247+48+1) cover about 21% of the source corpus. Apparently, spontaneous conversations consist of 'stock phrases' and/or true MWEs to a large extent. The 6-gram was excluded from further analysis, as generalizations cannot be made on the basis of one type.

We also studied what kinds of N-grams were selected. The majority of the N-grams are the beginning of (what is likely to become) a main clause. In Dutch, given information tends to appear at the beginning of a clause, whereas new information tends to occur at the end. The high proportion of conventional expressions at the beginning of a clause points to an almost automatic generation, which may well help speakers to overlap cognitive processing needed to express the new information in the clause. Listeners may also profit from such an alternation of predictable and new information. In any case, the high frequency of a small number of fixed clause-initial 'formulae' suggests that the variety of introductory clauses is not very broad in conversational Dutch.

4. Experiment 2

4.1. Method

In the second experiment, the manually verified phonetic transcriptions of *Corpus 2* were used to study the pronunciation of MWEs. According to the criterion proposed in [13], the minimum frequency for this 333,284 word corpus would be about 3 for 3-grams and 4-grams, and 1.5 for 5-grams and 6-grams. Since these numbers are too small to draw meaningful conclusions on pronunciation variation, we decided to select only the N-grams that occur at least 7 times. In *Corpus 2*, none of the 5-grams or 6-grams fulfilled this minimum frequency criterion. Consequently, this experiment is limited to an analysis of 3-grams and 4-grams.

A number of disagreement measures were calculated for the selected MWEs. We first used the program Align [14] to calculate the percentage disagreement between the actual and canonical pronunciations of words and MWEs:

$$\%disagreement = 100 * \frac{(\#Sub + \#Del + \#Ins)}{\#pc}$$

#Sub, #Del and #Ins are the number of substitutions, deletions, and insertions, respectively.
#pc is the number of phonemes in canonical transcription.

Next, we computed the average weighted percentage disagreement (AWD) for the words in the MWEs (AWD_MWE), and for the same words in all other contexts (AWD_other):

$$AWD = \frac{\sum_{n=1}^{N} \%avg\_dis_n * \#pc_n}{\sum_{n=1}^{N} \#pc_n}$$

%avg_dis$_n$ is the average percentage disagreement of word n.
#pc$_n$ is the number of phonemes in canonical transcription of word n.

Words in 'other context' are defined in the following way. Suppose the target word is *B* in the MWE context *ABC*. Then the pronunciation of all occurrences of word *B* in MWE context *ABC* is compared to the pronunciation of all occurrences of *B* in the context *XBZ* (X ≠ A & Z ≠ C). Words in 'other context' are defined similarly when the target word is the first or last word of a 3-gram, and for the four word positions in a 4-gram.

4.2. Results

Statistics for the 3-grams and 4-grams under investigation are presented in Table 2.

Table 2: Properties of the selected N-grams

	#types	token/type ratio	frequency range
3-gram	110	17.5	7-118
4-gram	21	13.8	7-50

For these selected N-grams, the percentage disagreement between the actual and the canonical pronunciation was calculated. The average results are shown in Table 3.

Table 3: Percentage disagreement for selected N-grams

	%sub	%del	%ins	%dis
3-gram	12.66	11.21	0.36	24.23
4-gram	13.44	14.20	0.78	28.42

On average, about 13% of the canonical phonemes were pronounced differently, 11-14% were deleted, and less than 1% inserted. Thus, in total about one quarter of the phonemes in the MWEs were pronounced in a non-canonical way.

Next, we selected the ten 3-grams and 4-grams that showed the highest degree of discrepancy between the actual pronunciation and the canonical reference. For these N-grams, the pronunciation of words in MWE context was compared to the pronunciation of the same words in other contexts. AWD was calculated for the words in MWE context and in all other contexts, together with the difference between these two measures (see Tables 4 and 5).

Table 4: Disagreement measures for 3-grams

3-gram	AWD		diff
	MWE	other	
zoiets van ja	57.27	15.75	41.52
in ieder geval	37.17	12.26	24.91
af en toe	34.76	15.15	19.61
Op die manier	31.94	12.99	18.95
't is natuurlijk	45.59	31.11	14.48
weet ik niet	29.22	21.52	7.70
Dat is natuurlijk	34.62	28.76	5.86
Hoe heet dat	30.43	24.95	5.48
Ook helemaal niet	27.78	24.40	3.38
Als 't ware	23.15	35.88	-12.73

For both the selected 3-grams and 4-grams in Table 4 and Table 5, t-tests revealed that the differences in percentage disagreement between the two context conditions are significant (for 3-grams: p=0.010 and for 4-grams p=0.030): the average weighted percentage disagreement (AWD) in MWE context is significantly larger than in other contexts. In general, words in MWE context are (much) more reduced. In some cases the amount of reduction is extremely large: more than half of the canonical phonemes are deleted, and the number of syllables is reduced by a factor two or three (due to resyllabification and deletion).

Table 5: Disagreement measures for 4-grams

4-gram	AWD		diff
	MWE	other	
dat vind 'k ook	48.89	29.00	19.89
op een gegeven moment	47.13	27.91	19.22
dat maakt niet uit	42.42	26.49	15.93
dat is niet zo	40.00	28.47	11.53
of wat dan ook	31.54	22.10	9.44
'k weet niet precies	28.57	22.73	5.84
dat weet ik niet	29.03	25.96	3.07
weet ik veel wat	26.45	25.08	1.37
dat weet ik nog	24.55	26.15	-1.60
als 't goed is	18.57	32.41	-13.84

5. Discussion

The analysis of frequent N-grams showed that a very large proportion (21%) of the words in the spontaneous speech in the CGN corpus are part of word sequences that occur frequently. This highly repetitive and predictable nature of extemporaneous speech deserves more attention in the future than it has received in the past. Furthermore, while compiling the set of frequent N-grams, we also found that there are quite a number of N-grams that occur frequently in very specific communicative settings, and not at all in other settings. Whether this finding is coincidental or systematic can only be decided by comparing and analyzing more and larger spoken corpora than just the CGN.

The results presented in section 4.2 clearly indicate that for all the words in the N-grams investigated, the actual pronunciation is reduced as compared with its canonical representation. The amount of reduction in pronunciation is mainly caused by vowel reduction (substitutions), and deletion of many segments.

In our study, we wanted to determine whether this amount of reduction is characteristic of spontaneous speech across the board, or whether it is related to specific contexts, in particular those of frequent N-grams. To answer this question we examined the pronunciation variants of the same words in the context of the MWE and in all remaining (other) contexts. We found that, for almost all words investigated, it holds that the degree of reduction is higher when these words appear in the context of frequent MWEs as opposed to when they appear in any other context.

Pronunciation thus appears to be one of the aspects of MWEs that cannot be predicted from the individual words. This finding seems to suggest that, at least for the purpose of pronunciation modeling, it is necessary to add a number of frequent MWEs with their characteristic pronunciation variants to the (pronunciation) lexicon. This may be a better solution than indiscriminate addition of all the pronunciation variants observed to the individual words in the lexicon, which can be counter-productive [15].

The most important reason to start the research reported in this paper was to determine whether these MWEs, and their pronunciation variants, require special handling in ASR and automatic phonetic transcription (APT). Previous research has shown that modeling pronunciation variation can be beneficial for both APT and ASR: for APT, because the quality of the resulting transcriptions can be improved [16, 17], and for ASR, because the word error rates can be reduced

[9]. In ASR research, it has also been shown that if too many variants are added, word error rates will start increasing. Specific modeling of pronunciation variation in MWEs has been studied in the field of ASR but, as far as we know, not in the field of APT. In ASR, MWEs are referred to as phrases, word tuples, multiword units, or multiwords. Different criteria are used to select (usually a small number of) MWEs. Adding these MWEs and their pronunciation variants to the lexicon usually reduces word error rate, which is generally the main goal of ASR studies and, consequently, no detailed study of pronunciation variation of MWEs has been carried out. In our study, we examined the type of pronunciation variation that characterizes a selected number of frequent MWEs and found that these MWEs exhibit uncommon pronunciation patterns that are not found in other contexts. We therefore suggest these MWEs to be included as lexical entries in the pronunciation lexicons employed in ASR and APT. In both cases, this is likely to increase the performance of the system.

6. Conclusions and future research

In this paper we presented an explorative study of MWEs in spontaneous speech, focusing on the pronunciation of MWEs. We showed that the words composing the MWEs investigated do indeed have different pronunciation patterns when they appear in the MWE context as opposed to when they appear in other contexts. This provides evidence for the fact that these MWEs require special handling in ASR and APT.

At the moment, we are continuing our research on MWEs. Apart from the spontaneous utterances from CGN, we are studying other speech styles, to investigate if there are differences between speech styles. Furthermore, apart from the selection criteria mentioned in this paper, we are also studying criteria such as mutual information and other probabilistic measures.

Future research could profit from the application of shallow syntactic parsing to the classification of N-grams; in this study, this was performed on the basis of the orthography alone. More detailed information about the type and the degree of completeness of the syntactic constituent formed by frequent N-grams should help in selecting the word sequences that are candidates for inclusion in a MWE lexicon.

Adding information about prosody, if only in the form of the strength of the juncture between adjacent words, is an obvious extension of the work reported in this paper. It seems evident that the presence of clear phonetic boundaries between adjacent words prevents the deletion of large phoneme clusters across the boundary. However, large corpora with accurate transcriptions would be needed to support the research.

7. Acknowledgements

8. References

[1] Odijk, J., "Reusable Lexical Representations for Idioms", *Proceedings LREC 2004, Lisbon 2004*, p. 903-906.

[2] Nunberg, G., Sag, I.A., and Wasow, T. "Idioms", *Language, 70, 1994*, p. 491-538.

[3] Wong-Fillmore, L. "Individual Differences in Second Language Acquisition", *C. Fillmore, D. Kempler and W. Wang (Eds.) Individual Differences in Language Ability and Language Behaviour.* Academic Press, New York; 1979. p. 203-228.

[4] Koster, C.H.A., "Transducing Text to Multiword Units", *Proceedings MEMURA 2004 workshop, Lisbon, 2004*, p. 31-38.

[5] Nivre, J., and Nilsson, J., "Multiword Units in Syntactic Parsing", *Proceedings MEMURA 2004 workshop, Lisbon, 2004*. p. 39-46.

[6] Sag, I.A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D., "Multiword expressions: A pain in the neck for NLP", *LinGO Working Paper (2001-03, 2001)* http://lingo.stanford.edu/pubs/WP2001-03.ps.gz.

[7] Oostdijk, N.H.J., "The design of the Spoken Dutch Corpus", *P. Peters, P. Collins and A. Smith (Eds.): New Frontiers of Corpus Research*, Amsterdam: Rodopi; 2002. p. 105-112.

[8] Pallett, D.S., "A look at NIST's benchmark ASR tests: past, present, and future", *Proceedings Workshop Automatic Speech Recognition and Understanding, 2003*, p. 483-488.

[9] Strik, H., and Cucchiarini, C., "Modeling pronunciation variation for ASR: A survey of the literature", *Speech Communication, Vol. 29 (2-4), 1999*, p. 225-246.

[10] Kessens, J.M., Cucchiarini, C., and Strik, H., "A data-driven method for modeling pronunciation variation", *Speech Communication, 40 (4), 2003*, p. 517-534.

[11] Yang, Q., and Martens, J-P., "On the importance of exception and cross-word rules for the data-driven creation of Lexica for ASR", *Proceedings 11th ProRisc Workshop, Veldhoven, The Netherlands, 2000*, p. 589-593.

[12] Binnenpoorte, D., Cucchiarini, C., Boves L. and Strik H., "Multiword Expressions in Spoken Language: an exploratory study on pronunciation variation", accepted for publication in *Computer Speech and Language, 2005*.

[13] Biber, D., Johannson, S., Leech, G., Conrad, S., and Finegan, E., *The Longman Grammar of Spoken and Written English*, Longman, Harlow, Essex, 1999. p. 987-1036.

[14] Cucchiarini, C., "Assessing transcription agreement: methodological aspects", *Clinical Linguistics & Phonetics, Vol. 10, No. 2, 1996*, pp. 131-155.

[15] Kessens, J.M., Wester, M., and Strik, H., "Improving the performance of a Dutch CSR by modelling within-word and cross-word pronunciation variation", *Speech Communication, 29 (2-4), 1999*, p. 193-207.

[16] Binnenpoorte, D., Cucchiarini, C., Strik, H., and Boves, L., "Improving Automatic Phonetic Transcription of Spontaneous Speech through Variant- Based Pronunciation Variation Modelling", *Proceedings of LREC 2004, Lisbon, 2004*, p. 681-684.

[17] Schiel, F., "Automatic Phonetic Transcription of Non-Prompted Speech", *Proceedings of the ICPhS 1999, San Francisco, 1999*, p. 607-610.