Lang Resources & Evaluation (2010) 44:41–58
DOI 10.1007/s10579-009-9095-y

Analyzing and identifying multiword expressions in spoken language

Helmer Strik · Micha Hulsbosch · Catia Cucchiarini

Published online: 4 August 2009

Abstract The present paper investigates multiword expressions (MWEs) in spoken language and possible ways of identifying MWEs automatically in speech corpora. Two MWEs that emerged from previous studies and that occur frequently in Dutch are analyzed to study their pronunciation characteristics and compare them to those of other utterances in a large speech corpus. The analyses reveal that these MWEs display extreme pronunciation variation and reduction, i.e., many phonemes and even syllables are deleted. Several measures of pronunciation reduction are calculated for these two MWEs and for all other utterances in the corpus. Five of these measures are more than twice as high for the MWEs, thus indicating considerable reduction. One overall measure of pronunciation deviation is then calculated and used to automatically identify MWEs in a large speech corpus. The results show that neither this overall measure, nor frequency of co-occurrence alone are suitable for identifying MWEs. The best results are obtained by using a metric that combines overall pronunciation reduction with weighted frequency. In this way, recurring "islands of pronunciation reduction" that contain (potential) MWEs can be identified in a large speech corpus.

H. Strik (✉) · M. Hulsbosch · C. Cucchiarini
Department of Linguistics (Section Language and Speech), Radboud University, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
e-mail: strik@let.ru.nl; h.strik@let.ru.nl
URL: http://lands.let.ru.nl/~strik/

M. Hulsbosch
e-mail: mhulsbosch@student.ru.nl

C. Cucchiarini
e-mail: c.cucchiarini@let.ru.nl

Present Address:
H. Strik
Erasmus Building, room 8.14, Erasmusplein 1, 6525 HT Nijmegen, The Netherlands

Keywords Multiword expressions · Spoken language · Transcription · Pronunciation reduction · Identification

Abbreviations

MWE	Multiword expression
CGN	Corpus Gesproken Nederlands (Spoken Dutch Corpus)
Sub	Substitutions
Del	Deletions
Ins	Insertions
Dif	Differences
%Dis	Percentage disagreement
PhDist	Phonetic distance
LCa	Length of the canonical transcription
LRe	Length of the realization
ALD	Absolute length difference
RLD	Relative length difference
Dur	Duration
Freq	Frequency

1 Introduction

Multiword expressions (MWEs) have been studied for many years by researchers working in various disciplines, i.e., psycholinguistics, phonetics, language acquisition and NLP, and are still a topical issue (Schmitt and Carter 2004; Rayson et al. 2006; Villada Moirón et al. 2006; Gregoire et al. 2007). The literature indicates that MWEs are pervasive in language use, but in spite of their apparent frequency of occurrence and the considerable attention they have received in numerous studies, MWEs are still a notion open to interpretation (Schmitt and Carter 2004). Many similar and/or overlapping terms have been used to indicate multiword sequences that are somehow "prefabricated" so as to exhibit a degree of cohesion that is generally not present in other utterances. Terms like MWEs, formulaic sequences, fixed expressions, stock phrases, sayings, clichés, speech formulae, lexical phrases, automatized chunks, prefabricated phrases and collocations have been used in the various disciplines to denote such sequences or specific subcategories of them (Wray and Perkins 2000; Van Lancker Sidtis and Rallon 2004).

In general the definitions used in NLP tend to be related to the behavior of these word sequences, i.e., MWEs are "expressions whose linguistic behavior is not predictable from the linguistic behavior of their component words" (Van de Cruys and Villada Moirón 2007: 25) or MWEs are "idiosyncratic interpretations that cross boundaries (or spaces)"(Sag et al. 2002:2). In psycholinguistics and language acquisition research it is more common to use descriptions of MWEs that refer to the function they fulfill (Nattinger and DeCarrico 1992) or the way in which they are processed (Wray and Perkins 2000; Sprenger et al. 2006; Wood 2004; Schmitt and Carter 2004; Conklin and Schmitt 2007). Examples of this type of definition can be

found in Wray and Perkins (2000: 1) who define formulaic sequences as being "prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar".

In addition, while psycholinguistic and language acquisition studies have addressed MWEs also in spoken language, the majority of MWE studies in NLP have concerned MWEs in written language. Excellent overviews with many references can be found in the PhD theses by Evert (2004) and Villada Moirón (2005). More recent work can be found in the proceedings of workshops on MWEs: e.g., the EACL workshop in April 2006 in Trento (Rayson et al. 2006), the COLING/ACL workshop in July 2006 in Sydney (Villada Moirón et al. 2006), and the ACL workshop of Prague in June 2007 (Gregoire et al. 2007). These recent PhD theses, workshops, and the present special issue make it clear that MWEs are still a topical issue.

One of the reasons why MWEs in spoken language have attracted the attention of researchers working in psycholinguistics and language acquisition is that MWEs appear to contribute to reducing cognitive load and promoting fluency (Towell et al. 1996; Chambers 1998; Wray and Perkins 2000; Wood 2004; Sprenger et al. 2006). In particular, MWEs appear to be less interrupted by pauses and to lead to increased speech rate (Underwood et al. 2004; Dahlmann and Adolphs 2007; Erman 2007). Since MWEs are stored in a holistic way, they can be retrieved more quickly than other word sequences and by providing a form of scaffolding, they promote speech fluency (Schmitt and Carter 2004). Research indicates that MWEs are abundant in speech that is typically produced under pressure (Kuiper 1996, 2004; Pluymaeckers 2003), such as sports commentaries and auctioneering.

In NLP, on the other hand, MWEs in spoken language have been studied in the field of automatic speech recognition (see, e.g., Beulen et al. 1998; Finke and Waibel 1997; Kessens et al. 1999; Sloboda and Waibel 1996), generally with the aim of establishing to what extent modeling such expressions can help reduce word error rate (Strik and Cucchiarini 1999). For instance, in Kessens et al. (1999) it appeared that handling frequent word sequences that showed substantial reduction, such as 'ik heb', 'dat is', and 'dat hoeft niet' (in English: 'I have', 'that is', and 'that isn't necessary') in the appropriate way indeed contributed to reducing word error rate. The main aim in these studies was to improve the performance of the speech recognizers, not to study the properties of MWEs in detail.

An attempt in this latter direction was made by Binnenpoorte et al. (2005). Since there was no generally accepted definition of MWE in spoken language, the Binnenpoorte et al. (2005) investigation was based on what was considered a reasonable operational definition of this concept: MWEs are contiguous sequences of words that are characterized by unpredictable pronunciation. The criterion of contiguity was considered to be necessary for defining MWEs in spoken language because pronunciation variation is expected to be caused by phenomena of cross-word assimilation and degemination, which will not work in sequences that are broken up by interspersed words. The aim of the study was to determine whether in spontaneous speech the words contained in frequent *N*-grams exhibit different pronunciation patterns in the *N*-gram context and in other contexts. For this purpose an inventory of frequently found *N*-grams was extracted from orthographic

transcriptions of spontaneous speech contained in a large corpus of spoken Dutch, the CGN ('Corpus Gesproken Nederlands'; Oostdijk 2002). These *N*-grams were filtered according to a number of criteria: they had to be contiguous, be between 2 and 6 words long, not straddle a deep syntactic boundary, and not contain disfluencies, hesitations and repetitions. For a small selection of these *N*-grams the phonetic transcriptions contained in the corpus were examined and were found to differ to a large extent from the canonical forms. To establish whether this was due to the specific status of these *N*-grams, the pronunciations of the individual words composing the *N*-grams were studied in two context conditions: (a) in the *N*-gram context and (b) in any other context. It appeared that words in the selected *N*-grams exhibited peculiar pronunciation patterns that were not found in other contexts and that these pronunciation patterns were specifically characterized by increased reduction when compared to the pronunciation patterns of the same words in other contexts. It was concluded that these frequent *N*-grams should be considered as MWEs, which should receive special attention, e.g., they should be treated as lexical entries in the pronunciation lexicons used in automatic speech recognition, with their own specific pronunciation variants.

Given the large amount of reduction observed in MWEs, an interesting question is how human listeners deal with reduced forms. Ernestus et al. (2002) report that although listeners in general cannot recognize highly reduced word forms in isolation, they manage to do so when these forms are presented in context. Furthermore, when listeners perceive reduced forms, they are generally not aware of the reduction present in these forms; in fact, listeners report that they have heard phonemes that were not present in the reduced forms (Kemps et al. 2004). For instance, if listeners hear 'vreesk', short for 'vreselijk' ('terrible'), many of them report that they have heard the sound /l/, which is present in the citation form but not in the reduced form they heard. These findings suggest that highly reduced word forms in MWEs need not be problematic in human communication because MWEs do provide a context by themselves with the consequence that listeners might even ignore the large amount of reduction.

Considering that MWEs thus appear to be characterized by a considerable amount of reduction (Binnenpoorte et al. 2005) it remains to be seen whether reduced pronunciation patterns are a prerogative of highly frequent stock phrases or whether they are also encountered in other contiguous sequences that are not readily recognized as being stock phrases. Research has shown that predictable words are more likely to be reduced (Bell et al. 2003; Gregory et al. 1999; Jurafsky et al. 2001). One can imagine that there may be word sequences that are not readily categorized as stock phrases, but that occur frequently enough as to exhibit high predictability and therefore considerable reduction in pronunciation. Fixed word sequences do occur frequently in spontaneous speech. In Binnenpoorte et al. (2005) it was found that 21% of the source corpus investigated consisted of fixed word sequences. As cognitive load increases, speakers are more likely to use prefabricated expressions (Kuiper 1996; Pluymaeckers 2003). In commentaries of sports games such expressions can cover up to 48% of the whole speech material (Pluymaeckers 2003).

Studying MWEs in spoken language can be relevant for different disciplines in various ways. In psycholinguistics it is important to investigate how MWEs are

perceived and stored in the lexicon and how they should be handled in psycholinguistic models. For language acquisition research it is relevant to know how MWEs are acquired and how they contribute to L2 fluency. For automatic speech recognition it is important to know how to identify and handle MWEs in order to improve recognition performance. Studying the pronunciation properties of MWEs is relevant also for phonetic research and automatic phonetic transcription. In order to study MWEs large speech corpora are needed; however given the size of the corpora it will not be possible to transcribe all material by hand, so automatic phonetic transcription could play a crucial role here. In speech synthesis proper handling of MWEs can also contribute to improving the quality and naturalness of the synthesized speech. And, finally, studying the pronunciation of MWEs is also important for automatic speech-to-speech translation, just as MWEs are important for machine translation of written texts: MWEs first have to be recognized correctly (automatic speech recognition for MWEs), have to be translated into the correct equivalent in the other language, and made audible in a correct way (speech synthesis of MWEs).

Having established that contiguous word sequences with unpredictable, usually reduced, pronunciation exist, the question that arises is whether and how these sequences can be detected automatically, because in the end this is the only way that data from large corpora can be handled to the benefit of research in speech science and speech technology. Several methods for identifying MWEs in written language, defined as "expressions whose linguistic behaviour is not predictable from the linguistic behaviour of their component words" (Van de Cruys and Villada Moirón 2007: 25), have already been proposed in the literature (see e.g., the overviews presented in Evert 2004, and Villada Moirón 2005). However, as far as we know, something similar for detecting "contiguous multiword expressions whose pronunciation is not predictable from the pronunciation behavior of their component words", i.e., MWEs in spoken language, has not been done.

The current study is a first step towards developing methods for identifying MWEs in spoken language. In other words, the question we address in this paper concerns the criteria that can be applied to spot MWEs in spoken language corpora. Since the definition of MWEs in spoken language refers to their pronunciation characteristics, and in particular to their reduced pronunciation, we will need to look for criteria and metrics that are able to capture pronunciation reduction in a meaningful way. In this connection it is important to underline that there is no gold standard that states which *N*-grams are MWEs and which not. This applies in particular to MWEs in spoken language. Since the definition of MWE is related to the degree of reduction in pronunciation, this is not something that we could ask human judges to evaluate. After all, human judges are often insensitive even to cases of extreme pronunciation reduction and apparently "restore" sounds that never appeared in the speech signal (Kemps et al. 2004).

Another point to be considered is that all definitions of MWEs mentioned so far do not contain frequency as a criterion for defining MWEs, while all studies indicate that an important characteristic of MWEs is their frequency. As a matter of fact, their being prefabricated, cliché, fixed and automatized is considered to be the result of their frequency of occurrence. So it seems that the element of frequency should somehow be used to define MWEs and to identify them in speech corpora.

We address the issue of MWE identification in spoken language on the basis of a study on the 'Spoken Dutch Corpus' (CGN), in which a number of possible indicators of reduced pronunciation are investigated to determine which of them are most promising for selecting potential MWEs. The current study is exploratory to a large extent. We start by studying two cases, two MWEs that emerged from the Binnenpoorte et al. (2005) study, which occur frequently in Dutch (see Sect. 3) and which can be categorized as either "sentence builders" (Granger 1998; Schmitt et al. 2004) or "discourse devices" (Nattinger and DeCarrico 1992). Many tokens of these two cases were extracted from the 'Spoken Dutch Corpus'; the properties of these tokens were studied and compared to the average properties of all other utterances in the corpus. The insights gained from these two case studies are subsequently used to develop methods for identifying MWEs in spoken language (see Sect. 4). We end with discussion and conclusions in Sect. 5.

2 Material

The database used for the current study is the 'Spoken Dutch Corpus' (CGN) a corpus containing about 9 million words of contemporary Dutch as spoken in the Netherlands and Flanders (Oostdijk 2002; CGN website 2004). All recordings are orthographically transcribed, lemmatized and enriched with part-of-speech (POS) information.

For about 10% of the corpus, more detailed annotations are available, such as manually checked broad phonetic transcriptions, word alignments, and syntactic and prosodic annotations. For the phonetic transcriptions a computer phonetic alphabet was used (CGN website 2004) that is a slightly modified version of SAMPA (for Dutch SAMPA, see Wells 1996). This sub-corpus of 900,000 words, called the core corpus, was composed in such a way that it faithfully reflects the design of the full corpus. In this paper we report results for all components of the core corpus, thus including many different speech styles and modalities, ranging from spontaneous to read, and from monologues to dialogues and even multilogues. As pointed out by Read and Nation (2004: 32), one of the difficulties in studying MWEs in spoken language is related to the limited availability of spoken corpora of adequate size with detailed annotations. Although this corpus might seem limited compared to those used for research in written language, it is quite large for a corpus of spoken language with phonological annotations.

3 Two case studies: 'in ieder geval' and 'op een gegeven moment'

We studied two MWEs that are frequently used in Dutch: 'in ieder geval' (IIG, 'in any case') and 'op een gegeven moment' (OEGM, 'at a given moment'/'at some point') (Binnenpoorte et al. 2005), which can be categorized as either "sentence builders" (Granger 1998; Schmitt et al. 2004) or "discourse devices" (Nattinger and DeCarrico 1992). 'op een gegeven moment' could also be classified as a specific case of discourse device, namely as a "temporal connector" (Nattinger and DeCarrico 1992).

Table 1 Realizations of the MWE "in ieder geval"

N	Realization	Sub	Del	%Dis
22	In id@ x@fAl	1	1	18.2
10	In id@ x@vAl	0	1	9.1
8	In i xfAl	1	4	45.5
7	In id@r x@vAl	0	0	0
7	In i vAl	0	5	45.5
6	n id@ x@vAl	0	2	18.2
6	In id@ xfAl	1	2	27.3
5	n i vAl	0	6	54.5
5	In id@ G@vAl	1	1	18.2
5	In i x@fAl	1	3	36.4
5	@n i vAl	1	5	54.5
4	n i fAl	1	6	63.6
4	In id@ vAl	0	3	27.3
4	In i x@vAl	0	3	27.3
4	In i vA	0	6	54.5
4	In i fAl	1	5	54.5
4	@n id@ x@fAl	2	1	27.3
…	…	…	…	…
204	Total Mean	0.8	3.0	34.6

For the phonetic transcriptions a computer phonetic alphabet was used (CGN website 2004) that is a slightly modified version of SAMPA (for Dutch SAMPA, see Wells 1996)

In all components of the core corpus of the CGN a total of 114 occurrences of OEGM, and 204 occurrences of IIG were found. In Table 1 the most frequent realizations of IIG are presented, all other pronunciations occurred less frequently, with the majority of them occurring only once. In total, 91 different realizations were observed for the 204 occurrences. The diversity for OEGM was even larger: 93 different realizations for 114 occurrences.

The differences between the actually observed pronunciations and the canonical transcriptions were determined by means of a dynamic programming algorithm (Cucchiarini 1996; Elffers et al. 2005). The canonical transcription of IIG that was used is **/In id@r x@vAl/** (11 phonemes and 5 syllables), and the canonical transcription used for "op een gegeven moment" is **/Op en G@gev@ momEnt/** (16 phonemes and 7 syllables). The canonical transcription represents the transcription that is most commonly encountered in Dutch. This explains why the 'n' is not contained in the canonical transcription of the word 'gegeven', 'n' after schwa is often deleted in Dutch spontaneous speech (Booij 1995). Thus, some reduction is already represented in the canonical transcriptions; if we had taken citation forms as the point of reference, the amount of pronunciation reduction would have been even larger.

The output of the dynamic programming algorithm contains the following information: number of substitutions (Sub), deletions (Del), and insertions (Ins), percentage disagreement (%Dis), and phonetic distance (PhDist). Some results for the MWE IIG are presented in Table 1. It can be observed in this Table that the canonical

transcription occurs only in 7 of the 204 cases (3.4%), and that percentage disagreement in some cases is higher than 50%. On the bottom row the mean values for all occurrences of the MWE IIG are given. The mean values for the MWE OEGM (of 1.2, 5.9, and 44.3%, respectively) are somewhat higher. In terms of the number of syllables, the smallest number of syllables in both cases is 2, which is a reduction by 3 syllables for IIG (60%), and a reduction by 5 syllables for OEGM (71%).

In general there are many deletions, some substitutions (usually indicating vowel reduction), and almost no insertions. Sometimes more than half of the phonemes are not pronounced in the canonical way, and the number of syllables is reduced substantially.

As explained above, the question to be addressed in this paper is whether it is possible to identify MWEs automatically by resorting to some measure of pronunciation reduction that can be calculated automatically. Several measures were calculated for all occurrences of the tokens of (possible) MWEs, and—for comparison—the same measures were also calculated for the complete corpus consisting of 900,000 words. The measures obtained for the whole corpus thus function as a kind of baseline, and measures obtained for the MWEs are compared to the measures for all utterances. Table 2 shows the values of seven measures of reduction which are calculated for the two MWEs OEGM and IIG, and for all other utterances in the corpus (i.e., the mean values and standard deviations for about 900,000 words). Indeed, the results presented in Table 2 make clear that for some measures the values obtained for the tokens of the MWEs are much larger than those obtained for 'all utterances' (i.e., the whole corpus consisting of 900,000 words).

Many measures depend on the length of the units for which they are calculated. Since MWEs often differ in length, both in terms of the number of phonemes and the duration, direct comparison of absolute measures is not very informative. This is even more so when measures for MWEs are compared to corresponding measures for all other utterances, because in that case the differences in length are even more substantial. To obviate this problem we therefore calculated relative measures to make it possible to compare between units of different lengths. The results for seven relative measures are presented in Table 2: the first four measures are divided by the length of the canonical transcription (LCa), and then multiplied by 100% to express the results in percentage points; the last three measures are divided by the duration (Dur).

Since we observed some substitutions and many deletions, we calculated the relative number of substitutions and deletions: Sub/LCa and Del/LCa. Our findings

Table 2 Mean (and standard deviation) values for all utterances and the two MWEs (OEGM & IIG)

Measures	All utt.	OEGM	IIG
1. Sub/LCa	7.8 (4.7)	7.3 (5.7)	7.2 (6.2)
2. Del/LCa	6.5 (6.5)	37.0 (12.7)	27.4 (17.7)
3. Dif/LCa	15.2 (8.6)	44.3 (13.6)	34.6 (17.6)
4. ALD/LCa	5.6 (6.8)	36.9 (12.6)	27.4 (17.8)
5. PhDist/Dur	5.0 (4.3)	38.7 (19.8)	25.8 (21.4)
6. LCa/Dur	14.3 (3.5)	30.6 (9.2)	27.0 (11.1)
7. LRe/Dur	13.4 (3.0)	18.7 (5.0)	19.0 (8.5)

suggest that other relative measures that express differences between the realization and the canonical transcription could also be indicators of pronunciation reduction. Therefore, we took the output of the dynamic programming algorithm to calculate the following two measures:

- Dif = Sub + Del + Ins; total number of differences
- ALD = LCa − LRe; absolute length difference

ALD is the difference between the length (number of phonemes) of the canonical transcription (LCa) and the length of the realization transcription (LRe). ALD/LCa is the absolute length difference relative to the length of the canonical transcription. Note that 100%*Dif/LCa is percentage disagreement, which for the sake of consistency and clarity here will be denoted as Dif/LCa. Furthermore, we calculated three measures relative to duration. The unit of LRe/Dur and LCa/Dur is the number of phonemes per second. LRe/Dur is the articulation rate, and PhDist/Dur is the phonetic distance (between realization and canonical transcription) per unit of time.

In Table 2 it can be observed that the mean number of substitutions in MWEs does not differ much from the mean value for all utterances; in fact, it is even somewhat smaller. Insertions are rare in these two MWEs and in all other utterances (the mean value (of 100%*Ins/LCa) for all utterances is 0.9%) so results for insertions are not presented here. Thus, if we compare the mean values for these two cases to the values for all other utterances, we see that the differences are small for number of substitutions and insertions, but very large for number of deletions: 6.5% for all utterances, and 37.0% (factor 37.0/6.5 = 5.7) and 27.4% (factor 27.4/6.5 = 4.2) for OEGM and IIG, respectively. Except for the substitutions in row 3, there are large differences in the mean values observed for MWEs and all utterances. These differences are all highly significant (t-test, $p < 0.01$). All values are (much) higher for the MWEs, indicating (much) more reduction in the case of the MWEs.

For the last six measures, we can see that the differences are somewhat smaller for LRe/Dur: for the other five measures the values for MWEs are more than twice as high (sometimes up to a factor 5), while for LRe/Dur the values are about 40% higher.

The articulation rate (LRe/Dur) indicates how fast speech sounds are articulated. In general, this measure is quite constant, even when we compare read to spontaneous speech, and native speech to non-native speech, as was done in Cucchiarini et al. (2002). Therefore, it is all the more remarkable that articulation rate turned out to be 40% higher for MWEs. Apparently MWEs constitute special cases in which we do manage to speed up articulation rate to a considerable extent. On the other hand, it is plausible that the articulation rate is not a factor 2–7 as high in MWEs (as is the case for measures 2–6), as there are physical-physiological limits to the increase in articulation rate.

4 Identification of MWEs

In the previous section, results for two case studies were presented. For seven relative measures, the mean values for these two cases were compared to the mean

values for all other utterances: for five of the seven measures (numbers 2–6) the values are more than twice as high for the MWEs, which suggests that these five measures might be potential indicators of MWEs. We therefore went on to investigate to what extent these five measures are suitable for identifying MWEs. The results are presented in the current section.

We call these five relative pronunciation measures: RP_i ($i = 1, 5$). Mean and standard deviation of RP_i, $M(RP_i)$ and $SD(RP_i)$ respectively, were first calculated for all utterances. Next, sequences of N words (N-grams) were extracted from the corpus. The values of RP_i were derived for all these N-grams ($Ng(RP_i)$), and then combined to obtain one overall pronunciation measure. This was done in the following way:

$$\text{If} |\text{Ng}(\text{RP}_i) - \text{M}(\text{RP}_i)| > \text{SD}(\text{RP}_i), \text{ then } \text{DP}_i = 1, \text{ else } \text{DP}_i = 0$$

$$\text{Overall Pronunciation Deviation : } \text{OPD} = \sum \text{DP}_i$$

$DP_i = 1$ indicates that the measure deviates more than one standard deviation from the average, i.e., it is an indication of a deviant pronunciation (DP). Overall Pronunciation Deviation (OPD) is the number of measures for which the deviation from the mean is more than 1 SD. OPD can thus vary between 0 and 5, an OPD value of 5 means that all 5 measures are outside the range, i.e., a strong indication of a deviant pronunciation. For each N-gram, values of OPD were first calculated for all occurrences (tokens) and then averaged to obtain a mean value for that N-gram (type).

Note that with this procedure we can identify cases of extreme hypoarticulation and hyperarticulation. If for certain N-grams there were many insertions compared to the canonical transcription, these N-grams would also be identified. However, in our data such N-grams with extreme hyperarticulation were not observed.

The analyses were carried out for N-grams with N larger than one. Obviously, the larger N, the smaller the observed frequency will be. For N larger than six, no N-grams were found that could qualify as MWEs. Below the final results are presented for $N = 2$–6. To make the data more transparent to readers that cannot read Dutch, in the tables we also provide literal, word-by-word translations of the various N-grams.

First, we made lists of the N-grams with the highest mean OPD values, i.e., the largest amount of reduction. The lists were ordered according to mean OPD, then frequency, and finally order of occurrence in the corpus. The results show that for all N ($N = 2$–6) the top-100 consists of N-grams for which the mean OPD is 5 (and the SD = 0); the frequencies are low for all these N-grams. Some of these N-grams contain (part of) the MWEs under study. For instance, if we look at the top-5 list of 6-grams in Table 3, we see that number 2 contains ‘in ieder geval’ and numbers 4 and 5 contain ‘op een gegeven moment’. These are cases of extremely reduced MWEs combined with other words. As was to be expected, using only the criterion of extreme reduction for identifying potential MWEs yields N-grams that are indeed extremely reduced, but not very frequent. So, going back to the question we posed in the introduction, namely whether reduced pronunciation patterns are a

Table 3 Top 5 6-grams ranked by mean OPD

N-gram	Mean OPD * $\sqrt{\text{Freq}}$	Mean OPD	Standard deviation	Freq.
dan heb je tenminste nog gips (then have you at least still plaster)	7.07	5	0	2
't is in ieder geval een (it is in any case a)	7.07	5	0	2
maar dat heb ik dat heb (but that have I that have)	7.07	5	0	2
je op een gegeven moment ook (you at a given moment also)	7.07	5	0	2
op een gegeven moment ook een (at a given moment also a)	7.07	5	0	2

prerogative of highly frequent stock phrases or whether they are also encountered in other contiguous sequences that are not readily recognized as being stock phrases, we have to conclude that the latter is the case: there are indeed word sequences that exhibit extreme reduction, but that are not highly frequent and are not readily recognized as being stock phrases.

Since frequency is considered to be another characteristic of MWEs, we went on to extract the *N*-grams with the highest frequency. These results show that for these *N*-grams the frequencies are much higher, that there is a large variation in mean OPD, and that for many *N*-grams the amount of reduction (the mean OPD) is quite small. In the top-5 list for the 6-grams, presented in Table 4, it can be observed that the mean OPD is smaller than 1 for numbers 2–5, and for the repetitions of 'ja' and 'nee' mean OPD is even smaller than 0,2. Sequences of 'ja', 'nee', and numbers, with little reduction, are also present in the other lists of *N*-grams (for $N = 2$–5). In addition, there are many other examples of frequent sequences with little reduction, e.g., 'ja dat is' ('yes that is', frequency = 224), and 'aan de andere kant' ('on the other side', frequency = 41); both have a mean OPD of less than 1. Apparently,

Table 4 Top 5 6-grams ranked by frequency

N-gram	Mean OPD * $\sqrt{\text{Freq}}$	Mean OPD	Standard deviation	Freq.
op de één of andere manier (in the one or other way)	11.30	2.41	2.19	22
ja ja ja ja ja ja (yes yes yes yes yes yes)	0.73	0.18	0.51	17
nee nee nee nee nee nee (no no no no no no)	0.24	0.06	0.24	17
één twee drie vier vijf zes (one two three four five six)	2.77	0.77	1.25	13
twee drie vier vijf zes zeven (two three four five six seven)	2.41	0.73	0.86	11

Table 5 Top 10 6-grams ranked by mean OPD * $\sqrt{\text{Freq}}$

N-gram	Mean OPD * $\sqrt{\text{Freq}}$	Mean OPD	Standard deviation	Freq
op de één of andere manier (in the one or other way)	11.30	2.41	2.19	22
speelt de bal even terug naar (plays the ball just back to)	9.00	4.50	0.87	4
'k weet niet of je dat (I know not whether you that)	7.51	4.33	0.47	3
de rechterkant van 't veld naar (the right side of the field to)	7.50	3.75	0.83	4
't is in ieder geval een (it is in any case a)	7.07	5.00	0.00	2
als je de advertentie in de (if you the advertisement in the)	7.07	5.00	0.00	2
daar hebben we 't vorige keer (there have we it last time)	7.07	5.00	0.00	2
't is wel zo dat er (it is surely so that there)	7.07	5.00	0.00	2
't op een gegeven moment toch (it at a given moment still)	7.07	5.00	0.00	2
wel 'ns een keer naar huis (surely once a time to home)	7.07	5.00	0.00	2

there are many frequent *N*-grams with little reduction, so a sequence of words that occurs often does not necessarily exhibit extreme reduction.

These results suggest that neither of the two measures in isolation, mean OPD and frequency, is able to capture MWEs. We therefore experimented with different combinations of these two measures. Since the range of values for frequency is much larger than that for OPD, we looked at ways of reducing its relative contribution to the final ranking. By taking the square root of frequency (Freq) the relative weight of frequency can be reduced. Shown in Tables 5, 6, 7, 8, 9 are the top-10 lists ranked according to mean OPD * $\sqrt{\text{Freq}}$. Again if we look at the top-5 of the 6-grams, we see that numbers 2 and 4 are sequences that figure in sports commentaries, and number 5 is a combination of two MWEs, i.e., 'het is' and 'in ieder geval', together with the word 'een'. Number 1 clearly stands out, in terms of the Mean OPD * $\sqrt{\text{Freq}}$. value in combination with a high frequency. This is also a common MWE that can be categorized as sentence builder and that also emerged in the Binnenpoorte et al. (2005) study.

We now see that well-known MWEs figure high on these lists: OEGM and IIG on place 1 in the lists of 4- and 3-grams, respectively; furthermore, many of the 'multiwords' mentioned in Kessens, Wester, Strik (1999) do also appear on the list of 2-grams (rank order numbers + word sequence): 2. 't is (it is), 3. da's (that is), 6. ik heb (I have), 10. dat is (that is), 20. heb ik (have I), 24. is 't (is it), and 49. 'k heb (I have). Apparently, these well-known MWEs are present in the top-100 lists ranked according to mean OPD * $\sqrt{\text{Freq}}$, while they were not present in the top-100 lists ranked according to OPD. In Tables 5, 6, 7, 8, and 9 we can observe that the values of mean OPD * $\sqrt{\text{Freq}}$ usually vary gradually, and that several of these

Table 6 Top 10 5-grams ranked by mean OPD * $\sqrt{\text{Freq}}$

N-gram	Mean OPD * $\sqrt{\text{Freq}}$	Mean OPD	Standard deviation	Freq
'k weet niet of je (I know not whether you)	12.97	4.10	1.22	10
en op een gegeven moment (and at a given moment)	12.85	4.86	0.35	7
op één of andere manier (in one or other way)	12.20	3.38	1.60	13
't is in ieder geval (it is in any case)	11.84	4.83	0.37	6
de één of andere manier (the one or other way)	11.73	2.50	2.08	22
op een gegeven moment ook (at a given moment also)	11.34	4.29	1.75	7
je op een gegeven moment (you at a given moment)	11.18	5.00	0.00	5
de bal even terug naar (the ball just back to)	10.73	4.80	0.40	5
op een gegeven moment uh (at a given moment erm)	10.61	3.75	0.43	8
op de één of andere (in the one or other way)	10.43	2.17	1.95	23

Table 7 Top 10 4-grams ranked by mean OPD * $\sqrt{\text{Freq}}$

N-gram	Mean OPD * $\sqrt{\text{Freq}}$	Mean OPD	Standard deviation	Freq
op een gegeven moment (at a given moment)	45.05	4.55	1.13	98
'k weet niet of (I know not whether)	18.14	3.78	1.53	23
en dan moet je (and then must you)	17.26	2.84	2.06	37
en dan kun je (and then can you)	17.21	4.06	1.18	18
ik weet niet of (I know not whether)	16.25	2.42	1.97	45
één of andere manier (one or other way)	16.23	2.74	2.07	35
ja ik weet niet (yes I know not)	15.67	2.61	1.64	36
maar dan moet je (but then must you)	15.64	3.26	1.94	23
weet niet of je (know not whether you)	15.50	4.14	1.55	14
dat vind 'k wel (that find I surely)	15.23	4.07	1.44	14

Table 8 Top 10 3-grams ranked by mean OPD * $\sqrt{\text{Freq}}$

N-gram	Mean OPD * $\sqrt{\text{Freq}}$	Mean OPD	Standard deviation	Freq
in ieder geval (in any case)	54.64	3.79	1.67	208
op een gegeven (at a given)	46.65	4.55	1.00	105
een gegeven moment (a given moment)	45.86	4.63	1.01	98
dan moet je (than must you)	41.15	2.81	1.93	214
dan kun je (than can you)	37.75	3.91	1.45	93
dat vind ik (that find I)	33.53	3.34	1.89	101
't is een (it is a)	33.45	2.80	1.98	143
dat vind 'k (that find I)	33.27	3.82	1.39	76
ik weet niet (I know not)	32.61	2.62	2.02	155
ja da 's (yes that 's)	31.71	2.21	0.63	205

Table 9 Top 10 2-grams ranked by mean OPD * $\sqrt{\text{Freq}}$

N-gram	Mean OPD * $\sqrt{\text{Freq}}$	Mean OPD	Standard deviation	Freq
als je (if you)	125.39	4.14	1.14	918
't is (it is)	99.82	2.79	1.92	1,282
da 's (that 's)	78.79	2.32	0.69	1,157
en dan (and then)	78.30	2.12	2.12	1,366
kun je (can you)	75.05	3.96	1.17	359
Ik heb (I have)	71.70	2.24	1.88	1,022
dan moet (then must)	71.45	3.78	1.44	357
van de (of the)	71.08	1.56	1.35	2,078
volgens mij (according to me)	69.42	3.91	1.56	316
dat is (that is)	65.28	1.44	1.67	2,041

N-grams are formed by well-known MWEs or parts of them, combined with other words.

5 Discussion and conclusions

In this study we have proposed and investigated a number of measures that could be used to (semi)automatically identify MWEs in spoken language and which refer to a definition of MWEs as contiguous multiword expressions with reduced pronunciation. Since this definition appeals to properties of speech that human judges are not very good at evaluating—research indicates that human listeners are not sensitive to pronunciation reduction—it is not possible to ask human judges to draw up a list of such expressions, which could be used to cross-validate our identification measures. Therefore, there is no gold standard. To establish whether the metrics proposed here are suitable for automatic identification of MWEs, we studied whether these measures managed to identify well-known MWEs that had emerged from previous studies.

Closer inspection of the results shows that in many cases the *N*-grams identified on the basis of these measures contain an MWE or part of one, sometimes combined with other (often reduced) words. For instance, the first and last three words of the MWE 'op een gegeven moment' appear on place two and three in the top 10 list of 3-grams; and the MWE 'op een gegeven moment' in combination with another word appears 5 times in the top 10 list of 5-grams. Furthermore, (parts of) the MWEs 'in ieder geval' and 'op de één of andere manier' also appear several times in the presented lists. These *N*-grams are related to MWEs, because they either contain part of an MWE, or a frequent combination of another (reduced) word with part of or a whole MWE.

The metrics proposed here combine indicators of pronunciation reduction and frequency measures and thus seem plausible potential markers of MWEs in spoken language. In general, the values of mean OPD * $\sqrt{\text{Freq}}$ vary gradually. Possible exceptions might be the first 6-gram (see Table 5), the first 4-gram (Table 7), the

first 3-gram (Table 8), and the first 2 2-grams (Table 9). However, it does not seem to be possible to simply draw a line somewhere, and classify everything above that line as MWEs.

Our study has indicated that OPD and frequency alone are not suitable metrics for identifying MWEs, and that a combination of—very broadly speaking—a pronunciation deviation score and a collocation measure is to be preferred. The metric employed in the present paper is OPD * $\sqrt{\text{Freq}}$, which gave more satisfactory results than other metrics we tried. But of course alternative combinations are also possible. For instance, we also tried Z-scores:

$$\text{DP}_{2,i} = (\text{N}(\text{RP}_\text{i}) - \text{M}(\text{RP}_\text{i}))/\text{SD}(\text{RP}_\text{i}), \text{ and } \text{OPD}_2 = \sum \text{DP}_{2,\text{i}}.$$

An advantage of this metric is that $\text{DP}_{2,i}$ is not just 0 or 1 (as is the case with DP_i), but can take on many other values. $\text{DP}_{2,i}$ denotes the difference between the measure and its mean in terms of the number of standard deviations. Apart from small differences regarding the exact position in the ranking, the differences in the top lists were small. Besides co-occurrence frequency, we also tried several other collocation measures which were calculated with the *N*-gram Statistics Package (NSP; Pedersen 2006). However, none of these collocation measures yielded better results than co-occurrence frequency. That frequency can perform as well as many collocation measures was also observed by other authors for other tasks, see e.g., Evert and Krenn (2001).

At this point it is important to notice that the resulting MWEs can differ between tasks. In the present study MWEs were extracted from the large general purpose corpus CGN, while in Kessens et al. (1999) they were extracted from a corpus (called OVIS) containing recordings of an interactive train timetable information system. MWEs like, e.g., 'ik heb' and 'dat is' were found in both cases, many more were found in the CGN, and some were found in OVIS and not in CGN, e.g., 'dat hoeft niet'. The latter was frequent and often extremely reduced in OVIS, because it was the answer to a question at the end of the conversation: "Do you want some other information?". Furthermore, each corpus has its own peculiarities, which can make comparisons between corpora difficult and can give rise to practical problems which depend to a large extent on small details of the corpus used. Although most of these are outside the scope of the present paper, some should be mentioned, since they are visible in the results presented here. They concern the notation conventions used in CGN, where reduction was already annotated in the orthographic form. For instance, in the expression 'ik weet het niet' the word 'het' is usually not (clearly) pronounced and then in CGN the orthographic transcription can be 'ik weet niet'. Furthermore, in the orthographic transcriptions of CGN the following notations occur: ''k', ''t' and ''s' (short for 'ik', 'het', and 'is', resp.). In the top 10 list of 2-grams we see 'da 's' and 'dat is', which could be combined into one entry. There are some other examples in our results. If we had carried out these adjustments (i.e., by combining some of the entries), the ranking would have been somewhat different for these entries, but the general results would have most probably been the same.

To summarize, it seems that the measures proposed and tested here are capable of detecting *N*-grams that may qualify as MWEs or are in some way related to MWEs,

where MWEs in spoken language are characterized as contiguous word sequences that frequently reoccur, and whose component words exhibit properties (in this case regarding pronunciation) that are different from the properties in other contexts. Similar characterizations (or definitions) can also be found in research on MWEs in written language (see, e.g., Piao et al. 2005).

However, it is clear that the measures investigated are not sufficiently refined to be able to also define the exact boundaries of MWEs. Of course, this does not imply that these measures are completely useless. In any case, these different top lists do give suggestions about what possible MWEs might be. Specifically, we can say that the measures proposed here are in any case helpful to identify recurring "islands of pronunciation reduction" that might contain potential MWEs or parts of these. So, even though this is not accurate enough for completely automatic detection, it can still be useful for semi-automatic detection with a two-step procedure in which these measures are first applied to distill potential MWE candidates from large speech corpora, which are subsequently scrutinized by human judges for further processing.

From the research presented in this paper the following conclusions can be drawn. First, MWEs exhibit a large amount of pronunciation variation, covering the whole gamut from complete citation form to severely reduced forms. Second, in MWEs more pronunciation reduction is observed than in other utterances. Third, metrics to detect MWEs in spoken language corpora can be developed by combining measures of pronunciation reduction and measures of frequency. In this way we managed to identify contiguous word sequences that are MWEs, or are in some way related to MWEs, in a large speech corpus.

Acknowledgments We would like to thank two anonymous reviewers for their useful comments.

References

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M. L., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America, 113*, 1001–1024.

Beulen, K., Ortmanns, S., Eiden, A., Martin, S., Welling, L., & Overmann, J. (1998). *Pronunciation modeling in the RWTH large vocabulary speech recognizer*. (Paper presented at the ESCA Workshop "Modeling pronunciation variation for automatic speech recognition", Kerkrade).

Binnenpoorte, D., Cucchiarini, C., Boves, L., & Strik, H. (2005). Multiword expressions in spoken language: An exploratory study on pronunciation variation. *Computer Speech & Language, 19*(4), 433–449.

Booij, G. (1995). *The phonology of Dutch*. Oxford: Clarendon Press.

CGN website (2004). http://lands.let.ru.nl/cgn/ehome.htm. Accessed November 1, 2007.

Chambers, F. (1998). What do we mean by fluency? *System, 25*(4), 535–544.

Conklin, K., & Schmitt, N. (2007). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics, 29*(1), 72–89.

Cucchiarini, C. (1996). Assessing transcription agreement: Methodological aspects. *Clinical Linguistics & Phonetics, 10*(2), 131–155.
Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America, 111*(6), 2862–2873.
Dahlmann, I., & Adolphs, S. (2007). Pauses as an indicator of psycholinguistically valid multi-word expressions (MWEs)? *Proceedings of the ACL-2007 workshop on 'A broader perspective on multiword expressions'*, *Prague*, 49–56.
Elffers, A., Van Bael, C., & Strik, H. (2005). *Adapt: Algorithm for dynamic alignment of phonetic transcriptions*. (CLST internal report).
Erman, B. (2007). Cognitive processes as evidence of the idiom principle. *International Journal of Corpus Linguistics, 12*(1), 25–53.
Ernestus, M., Baayen, H., & Schreuder, R. (2002). The recognition of reduced word forms. *Brain and Language, 81*, 162–173.
Evert, S. (2004). The statistics of word cooccurences—Word pairs and collocations. Dissertation, Universität Stuttgart.
Evert, S., & Krenn, B. (2001). *Methods for the qualitative evaluation of lexical association measures*. (Paper presented at the 39th annual meeting of the association for computational linguistics, Toulouse).
Finke, M., & Waibel, A. (1997). *Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition*. (Paper presented at EuroSpeech-97, Rhodes).
Granger, S. (1998). *Prefabricated patterns in advanced EFL writing: Collocations and formulae. Phraseology: theory, analysis, and applications* (pp. 145–160). Oxford: Clarendon Press.
Gregoire, N., Evert, S., & Kim, S. N. (Eds.). (2007). *Proceedings of the ACL-2007 Workshop on 'A Broader Perspective on Multiword Expressions', Prague*. http://www.aclweb.org/anthology-new/W/W07/W07-11.pdf.
Gregory, M. L., Raymond, W. D., Bell, A., Fosler-Lussier, E., & Jurafsky, D. (1999). The effects of collocational strength and contextual predictability in lexical production. *Chicago Linguistics Society, 35*, 151–166.
Jurafsky, D., Bell, A., Gregory, M. L., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 229–254). Amsterdam: John Benjamins.
Kemps, R., Ernestus, M., Schreuder, R., & Baayen, R. H. (2004). Processing reduced word forms: The suffix restoration effect. *Brain and Language, 90*, 117–127.
Kessens, J. M., Wester, M., & Strik, H. (1999). Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation. *Speech Communication, 29*, 193–207.
Kuiper, K. (1996). *Smooth talkers. The linguistic performance of auctioneers and sportscasters*. Mawah, NJ: Lawrence Erlabum Associates.
Kuiper, K. (2004). Formulaic performance in conventionalised varieties of speech. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 37–54). Amsterdam: John Benjamins.
Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.
Oostdijk, N. H. J. (2002). The design of the Spoken Dutch Corpus. In P. Peters, P. Collins, & A. Smith (Eds.), *New Frontiers of Corpus Research.* (pp. 105–112). Amsterdam: Rodopi.
Pedersen, T. (2006). *Ngram Statistics Package (NSP)*. Retrieved November 1, 2007, from http://www.d.umn.edu/~tpederse/nsp.html.
Piao, S., Rayson, P., Archer, D., & McEnery, T. (2005). Comparing and combining a semantic tagger and a statistical tool for MWE extraction. *Computer Speech & Language, 19*, 378–397.
Pluymaeckers, M. (2003) *Prefabs in sports commentary*. Master's thesis, Tilburg University.
Rayson, P., Sharoff, S., & Adolphs, S. (Eds.). (2006). *Proceedings of the EACL-2006 workshop on 'multi-word-expressions in a multilingual context'*, Trento, Italy.
Read and Nation. (2004). Measurement of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 23–35). Amsterdam: John Benjamins.
Sag, I., Baldwin, T., Bond, F., Copestake, A. & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd international conferences on intelligent text processing and computational linguistics*, 1–15.
Schmitt, N. (2004). *Formulaic sequences: Acquisition, processing and use*. Amsterdam: John Benjamins.

Schmitt, N., & Carter, N. (2004). Formulaic sequences in action: An introduction. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 1–22). Amsterdam: John Benjamins.

Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are Corpus-derived Recurrent Clusters Psycholinguistically Valid? In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 127–152). Amsterdam: John Benjamins.

Sloboda, T., & Waibel, A. (1996). *Dictionary learning for spontaneous speech recognition*. (Paper presented at 4th international conference on spoken language processing, Philadelphia).

Sprenger, S. A., Levelt, W. J. M., & Kempen, G. (2006). Lexical access during the production of idiomatic phrases. *Journal of Memory and Language, 54*, 161–184.

Strik, H., & Cucchiarini, C. (1999). Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication, 29*(2–4), 225–246.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in davanced learners of French. *Applied Linguistics, 17*(1), 84–119.

Underwood, G., Schmitt, N., & Galpin, A. (2004). The eyes have it: An eye-movement study into the processing of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 153–172). Amsterdam: John Benjamins.

Van de Cruys, T. & Villada Moirón, B. (2007). Semantics-based multiword expression extraction. In *Proceedings of the ACL workshop 'A broader perspective on multiword expressions'*, 25–32.

Van Lancker Sidtis, D., & Rallon, G. (2004). Tracking the incidence of formulaic expressions in everyday speech: Methods for classification and verification. *Language & Communication, 24*, 207–240.

Villada Moirón, B. (2005). *Data-driven Identification of fixed expressions and their modifiability*. Dissertation, University of Groningen, The Netherlands.

Villada Moirón, B., Villavicencio, A., McCarthy, D., Evert, S., & Stevenson, S. (Eds.). (2006). *Proceedings of the COLING/ACL 2006 workshop on 'Multiword expressions: Identifying and exploiting underlying properties'*, Sydney. http://acl.ldc.upenn.edu/W/W06/W06-1200.pdf.

Wells, J. C. (1996). SAMPA for Dutch. http://www.phon.ucl.ac.uk/home/sampa/dutch.htm Accessed November 1, 2007.

Wood, D. (2004). An empirical investigation into the facilitating role of automatized lexical phrases in second language fluency development. *Journal of language and learning, 2*(1), 27–50.

Wray, A., & Perkins, M. R. (2000). The functions of formulaic language: An integrated model. *Language & Communication, 20*, 1–28.