Pauses as an Indicator of Psycholinguistically Valid Multi-Word Expressions (MWEs)?

Irina Dahlmann and Svenja Adolphs
School of English Studies
University of Nottingham
University Park, Nottingham, NG7 2RD, UK
{aexid,svenja.adolphs}@nottingham.ac.uk

Abstract

In this paper we investigate the role of the placement of pauses in automatically extracted multi-word expression (MWE) candidates from a learner corpus. The aim is to explore whether the analysis of pauses might be useful in the validation of these candidates as MWEs. The study is based on the assumption advanced in the area of psycholinguistics that MWEs are stored holistically in the mental lexicon and are therefore produced without pauses in naturally occurring discourse. Automatic MWE extraction methods are unable to capture the criterion of holistic storage and instead rely on statistics and raw frequency in the identification of MWE candidates. In this study we explore the possibility of a combination of the two approaches. We report on a study in which we analyse the placement of pauses in various instances of two very frequent automatically extracted MWE candidates from a learner corpus, i.e. the n-grams *I don't know* and *I think I*. Intuitively, they are judged differently in terms of holistic storage. Our study explores whether pause analysis can be used as an objective empirical criterion to support this intuition. A corpus of interview data of language learners of English forms the basis of this study.

1 Introduction

MWEs are ubiquitous in language (e.g. Erman and Warren, 2001; Wray, 2002; Pawley and Syder, 2000) but at the same time they present researchers, especially in the areas of NLP, descriptive linguistics and (second) language acquisition (see for example Sag et al., 2002; Wray, 2000, 2002) with a number of challenges. Two of the most serious challenges are the identification and definition of MWEs. These are interdependent and cause a circular problem: As long as we cannot identify and describe the properties of MWEs fully, a definition remains only partial and, in return, without a full definition the identification process is incomplete.

Nevertheless, methods of identification have been developed and used, based on broad criteria, e.g. human intuition, frequency information or semantic and grammatical properties (e.g. idioms, light-verb constructions, adjective noun collocations).

A considerable amount of research in NLP and in linguistics draws on two broad definitions by Sag et al. (2002) and Wray (2002), respectively. Sag et al. define MWEs 'very roughly' as

'idiosyncratic interpretations that cross word boundaries (or spaces)' (Sag et al. 2002:2).

They specify further that MWEs can be classified broadly into two categories according to their syntactic and semantic flexibility, i.e. lexical phrases and institutionalised phrases.

Wray (2002), coming from a psycholinguistic perspective, wants to be 'as inclusive as possible, covering any kind of linguistic unit that has been considered formulaic in any research field' (p.9). She defines the term 'formulaic sequence' as

'a sequence, continuous or discontinuous, of words or other elements, which is or appears to be prefabricated: that is, stored and retrieved whole from

Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, pages 49–56,
Prague, June 2007.

memory at the time of use, rather than being subject to generation or analysis by the language grammar.' (Wray 2002:9)

The main difference between the two definitions is the inclusion of holistic storage of MWEs in the mental lexicon by Wray, whereas Sag et al.'s definition, which has been used extensively in NLP research, focuses mainly on syntactic and semantic properties of the MWE.

One of the possible reasons why holistic storage has not found its way into NLP research may be related to the fact that this criterion is almost impossible to measure directly. However, it has been proposed that prosodic cues and pauses are indirect indicators of prefabricated language and holistic storage as MWEs in speech exhibit more phonological coherence (e.g. Hickey, 1993).

If we assume that MWEs are stored as holistic units in memory, we would firstly not expect to find pauses *within* MWEs. Pawley (1986) states that 'pauses within lexicalised phrase are less acceptable than pauses within free expressions, and after a hesitation the speaker is more likely to restart from the beginning of the expression' (p.107, quoted from Wray, 2002). This is in line with Raupach (1984) who studied spontaneous L2 speech production and stresses that 'a formal approach to identifying formula units in spontaneous speech must, as a first step, list the strings which are not interrupted by unfilled pauses' (p.116).

Secondly, we would expect that pauses, i.e. silent pauses and hesitation phenomena, may also serve in the delineation of MWE boundaries (Raupach, 1984:114).

The research outlined above is echoed in more recent studies of MWEs and pauses in the development of speech fluency. The placement, quantity and lengths of pauses are important markers of fluency (e.g. Riggenbach 1991) and the stretches between pauses may be fluent because pauses provide planning time to formulate the next utterance (Pawley and Syder, 2000) and the utterance may be (partly) a prefabricated string of words (MWE).

Previous research into MWEs and fluency is especially important from a methodological perspective, as it provides methodological frameworks for the study of pauses, for example, the integration of silent and filled pauses, which both provide planning time (Raupach, 1984; Pawley and Syder, 2000), or the significance of pause lengths (Pawley and Syder, 2000). These aspects are, for instance, not sufficiently reflected in existing pause annotation schemes in spoken corpora (see also section 3.1), which has hampered the study of pauses and MWEs on a large scale so far.

The aim of our study is therefore twofold. Firstly, in terms of methodology, we combine insights from fluency and MWEs research with a corpus approach and automatic extraction of MWEs.

Secondly, we analyse whether units which have been extracted automatically also comply with predicted pause behaviour (no pauses within MWEs, pauses as indicator of MWE boundaries) and therefore whether they are psycholinguistically valid.

This kind of study may help develop our understanding of MWEs in naturally occurring discourse. In addition, it allows us to explore further whether the study of pause phenomena might be a useful tool in the evaluation of automatic extraction methods.

2 Pauses and MWEs

As outlined above research on prosodic features and MWEs has found that MWEs tend to exhibit more phonological coherence (e.g. Hickey, 1993; Read and Nation 2004; Wray, 2002). Van Lancker et al. (1981), for instance, found phonological differences depending on whether a string carried literal or idiomatic meaning in a read aloud task (e.g. *skating on thin ice*). The differences in the literal and idiomatic contexts were partly mirrored in the number and placement of pauses. Idiomatic expressions are uttered at a faster speed which is to some extent related to the lack of pauses within the idiomatic expression (Van Lancker et al. 1981:331). Additional indicators are the pace at which key words were used (increased word duration of major lexical items in the literal version), the length of the whole utterance, pitch changes, and articulatory precision (Van Lancker et al., 1981). Phonological coherence and further prosodic features (stress and intonation) may therefore be regarded as physical indicators of the storage and retrieval of MWEs which in turn can help to identify MWEs in spoken language.

Problems with this kind of investigation are mainly related to the lack of consistent methodology for studying pauses as physical markers of holistic storage in an empirical manner, i.e. using naturally occurring corpus data. Key problems are the shortage of suitable spoken corpora and inconsistent pause annotation schemes.

3 Methodological challenges

3.1 Corpora and pause annotation

As the aim of this study is to explore holistic storage and retrieval of MWEs in naturally occurring speech, a corpus of spontaneous speech is required. Both, audio data and transcriptions are needed for the automatic extraction of MWEs and pause annotation respectively.

Unfortunately, not many available spoken corpora have been marked up for pauses as it is a very labour intensive process and currently has to be done largely manually. In cases where pause marking has been applied, it does not necessarily meet the specific requirements for phonological analysis (Read & Nation 2004:32). For example, pauses may not have been defined sufficiently for this purpose, as in the spoken part of the BNC where a pause is defined as a 'silence, within or between utterances, longer than was judged normal for the speaker or speakers' [1]. The definition of pause length – unlike in fluency research – can be too broad in existing corpus annotation, e.g. pauses have to be perceived as a pause (short, medium, long) or, when timing is included it is often very vague, e.g. a 'comma indicates a brief (1-2 second) mid-utterance pause with non-phrase final intonation contour' in the MICASE corpus.[2] In comparison, the minimum threshold for a pause lies at around 0.2-0.3 seconds in fluency research. Furthermore, not all corpora which contain silent pause annotation have also annotated filled pauses. In fact, a survey of 12 corpus pause coding schemes (native and learner language) shows that none complies with the requirements needed for the study of fluency and MWU related research.[3]

Due to the lack of corpora which combine spontaneous speech and appropriate pause annotation we have developed a learner corpus which we then selectively annotated for pauses. The corpus contains 290,000 transcribed words of spontaneous interview discourse produced by Chinese learners of English (with accompanying audio files). The proficiency level of the Chinese students in the whole corpus is based on IELTS scores and ranges from 5.0 – 6.5 (of max. 9). Scores from around 5.5 onwards (depending on the intended studies) are required for foreign students for admission at a British university. The two speakers investigated here have scores of 5.0 and 5.5 respectively.

Only two students have been chosen for this study in order to reduce the number of possible variables affecting the results, especially with regard to idiosyncratic usage.

The choice of learner data rather than native speaker data evolved not only from practical considerations, but also from the wider aim of our study which is related to fluency and language acquisition. In addition, when applying preliminary pause annotations to extracts of both native and non-native speech, we observed that learners seem to pause a lot more than native speakers. Native speakers seem to apply some other modes of 'pausing' – such as using fillers, repeating words or rephrasing – more extensively. Therefore, we might expect clearer results from the learner data initially. In fact, it will be interesting to see in comparison, whether pauses might even tell us more about learners than about native speakers with regard to the use of MWEs.

It nevertheless has to be acknowledged that there might be considerable differences in learner and native speech; however, both varieties are valid in their own right, especially with respect to holistic storage and usage.

Careful pause annotation was then carried out around a selected set of automatically extracted MWEs from the learner data (see 3.2 and 3.3) to explore the approach outlined above.

3.2 Automatic extraction – n-grams

Different MWE extraction methods abound but we decided to begin our study with an investigation of n-grams as a way into the proposed ap-

[1] http://www.natcorp.ox.ac.uk/docs/userManual/cdif.xml.ID=cdifsp (last accessed 25/03/2007)

[2] http://www.lsa.umich.edu/eli/micase/MICASE_MANUAL.pdf (last accessed 25/03/2007)

[3] This is especially unfortunate in the case of the London-Lund Corpus (LLC), which in theory lends itself to this kind of study for native English MWEs usage: The LLC contains not only pause annotation but also marking of other prosodic features such as tone unit boundaries, the nucleus, and varying degrees of stress. These can serve as additional indicators for MWEs in use. However, only silent pauses are marked and only in broad terms, i.e. '-' indicates a 'brief pause of one light syllable', '–' indicates a 'unit pause of one stress unit or 'foot'. This is one of the limitations of the only large-scale study in the field of pauses and MWEs (Erman, 2007), as it is based solely on the LLC and its annotation.

proach. The choice of n-grams, described as one of the most successful statistical models (Gil and Dias, 2003), was based on several reasons.

Firstly, the assumption behind n-grams is that continuous strings of words, which are used repeatedly and frequently in the same form in a speech community, are also likely to be stored holistically in memory.

Secondly, simple n-grams are continuous sequences. This aids the study of pauses at this early stage as discontinuous sequences or sequences with variable slots might exhibit different pause behaviour and/or prosodic features.[4]

In addition, the special case of learner language requires an extraction method which is based on the actual corpus data itself and not on preconceived ideas of whether or not a particular multi-word string is in fact a valid MWE, as is the case with symbolic or knowledge based extraction methods. Learners may have their own (sub-)set of MWEs (Wray 1999). These may be characterised by idiosyncratic MWEs, which nevertheless may be used frequently either by individuals or by a certain speech community, e.g. Chinese learners of English.

A further advantage of using n-grams is that the extraction is fully automated and therefore does not require human intervention. This extraction method does not take into account the additional factor of 'meaning' as the process of extraction itself is very mechanical and not dependant on meaning.

N	3-grams	Freq.	%
1	A LOT OF	352	0.17
2	**I DON'T KNOW**	**327**	**0.16**
3	**I THINK I**	**300**	**0.15**
4	I THINK IT'S	252	0.12
5	SO I THINK	220	0.11
6	I WANT TO	211	0.1
7	I THINK THE	188	0.09
8	BUT I THINK	185	0.09
9	I DON'T THINK	146	0.07
10	I THINK ER	143	0.07

Table 1. 10 most frequent 3-grams extracted from 290,000 words of learner interview data

[4] Discontinuous MWEs and n-grams are nevertheless important, which is reflected in the development of more refined extraction methods (e.g. positional n-grams (Gil and Dias, 2003) and ConcGrams (Chen et al. 2006)). However, they are only of secondary interest for us at this stage.

This is one of the disadvantages at the same time. Frequent examples in our spoken learner corpus are n-grams such as *I think er*, *I I I* or *and er I* which at first glance do not appear to be holistically stored MWEs.

Drawing on n-grams as an approach also allows us to study MWE candidates, which – on the basis of intuition – do not appear to be stored holistically, but nevertheless occur very frequently in the corpus.

For our analysis we have chosen two very frequent 3-grams (see Table 1) which contrast in terms of their internal consistency. *I don't know* seems to be an example of a self contained MWE candidate whereas *I think I* is an example of a MWE candidate which intuitively does not seem to be psycholinguistically valid, i.e. stored as a holistic item.[5]

3.3 Pause annotation and research questions

The analysis has been carried out for two different speakers and the following number of n-grams (see Table 2).

MWE candidate	Speaker MS001	Speaker MS003
I don't know	21	26
I think I	16	28

Table 2. MWE candidates per speaker

Pauses have been measured manually with audio-visual clues, i.e. the combination of audio recording and waveforms, both displayed by Adobe Audition. Within this software the pause length (in seconds, correct to the third decimal) is calculated by marking up a stretch of the wave form, which has been identified as a pause.

[5] The analysis of other contrastive pairs, e.g. on the basis of syntactic properties such as *I don't know* vs. *I don't see* (keeping the syntactic structure but changing the lexical verb - as suggested by one of the reviewers) also seems sensible. However, the choice of the substituting items has to be well informed by factors such as frequency of the single lexical verbs, compared to frequency of the whole string, as for example done by Tremblay et al. (2007). However, this does not necessarily lead to an unproblematic comparison: *I don't see*, for instance, only occurs two times in our data set of spontaneous speech, which is not frequent enough to find pause patterns or to compare it to the pause patterns of *I don't know*. Such an approach thus seems to lend itself more readily to experimental studies (such as the self-paced reading experiments by Tremblay et al. 2007) with carefully designed stimuli, and not to the study of natural occurring speech.

Pause measurement in fluency research commonly suggests thresholds between 0.2-0.3 seconds as a minimum for a silence to be regarded and perceived as a pause (e.g. Goldman Eisler, 1968, Towell et al., 1996). To account for this, pauses between 0.2 and 0.3 seconds length were measured correct to two digits in order to allow for a later adjustment of minimal pause length, pauses above 0.3 were measured to one digit. Filled pauses were measured if they seemed exceptionally long. Both, silent and filled pauses are marked here for the purpose of placement indication with '< >'.

The main focus of our analysis is on pause distribution and the following five cases of placements of pauses have been identified as pertinent to our study: ('____'indicates text which can theoretically be of any length, < > indicates pause)

a. M W < > E (pause within the MWE candidate)
b. < > MWE < >
c. < > MWE ____< >
d. < > _____MWE < >
e. < > _____MWE _____ < >

In the annotation of pause patterns around the two different MWE candidates the following questions are explored:

(1) Do the two candidates seem to be stored holistically, i.e. do they contain pauses within the extracted form or not? (Referring to pause placement pattern a.)

(2) Do pauses assist in the determination of MWE boundaries, i.e. are there any regular pause patterns which indicate boundaries? Do pauses seem to align MWEs in the form in which they were extracted? (Referring to b.-e.)

(3) Do the results comply with intuition, i.e. does *I don't know* fit the predicted behaviour better than *I think I*?

4 Results and discussion

4.1 'I don't know'

Forty seven *I don't know*'s, used by two different speakers within approximately 71,000 words of interview data have been studied for pause phenomena. The distribution is summarised in Table 3.

Pause distribution	**MS001**	**MS003**	**Σ**
MW< > E	--	--	**--**
< > MWE < >	9	1	**10**
< > MWE __ < >	5	14	**19**
< > __ MWE < >	2	3	**5**
< > __ MWE__ < >	5	8	**13**

Table 3. Pause distribution around 47 instances of *I don't know*

As expected, in the speech examples at hand, *I don't know* is never interrupted by pauses, which is a good indicator for holistic storage of this particular string of words by the two learners.

In terms of boundary alignments it can be observed that almost two thirds of the examples contain pauses immediately preceding *I don't know* (29:18), which in turn can be interpreted as a sign of a MWE boundary. It has to be taken into account that MWEs can occur within other MWEs or within a stretch of creative speech. Therefore, pauses do not need to be present on all occasions even if it seems to be a boundary. The fact, that pauses nevertheless do occur very often and that these pauses are proper pauses - on average far longer than the suggested 0.2 seconds (on average 0.57 seconds) reinforces the case for an actual boundary.

The case is different for the final boundary. If pauses occur right at the end of *I don't know* they are shorter overall (0.39 seconds on average). The main point is, however, that in over two thirds of the instances (32:15) no pause occurs in this place.

A further observation is that the 'ideal' form (in terms of boundary recognition and validation) < > MWE < > with pauses at either side of the extracted MWE candidate, occurs infrequently. It seems rather idealistic to expect language to be organized neatly according to stored chunks. Instead speakers are generally capable of placing several chunks and/or creative language together in one stretch of speech. Pawley and Syder (2000) suggest that 'the average number of words per fluent unit is about six' (p. 195) for fluent (native) speakers. The actual average number of words might differ slightly for learners, however the point is that either way the numbers are averages and in single instances stretches might be considerably longer. It is therefore not surprising that 3-word n-grams might be embedded within longer stretches of speech and are not surrounded by pauses. Furthermore, Miller (1956) states in his paper *The magical number* *seven*, that ‘the memory span is a fixed number of chunks, we can increase the number of bits of information that it contains simply by building larger and larger chunks, each chunk containing more information than before.’ (p.93). In other words, if *I don’t know* is stored as one chunk or item (instead of three single words) it is more likely that it may be embedded in a larger portion of language as the memory is able to handle more language items.

Moreover, the form < > MWE < > is mainly used by one speaker (MS001; 9:1). This points towards the importance of the consideration of idiosyncratic usage, especially when dealing with learner language (but it also plays a role in native usage): learners may use MWEs in a much more restricted way, i.e. the way they have learned a particular phrase instead of using it appropriate to the context. For instance, learner MS003 evidently also has a preferred way of using *I don’t know*, namely < > MWE __ < > (14:5).

It also has to be taken into consideration that *I don’t know* can be used as a discourse marker/filler or in the more literal sense of ‘I don’t have the knowledge’. This distinction might be of significance for clearer descriptions of the MWE generally.

In summary, one may want to argue that *I don’t know* may function as a core MWE. It seems to be stored holistically as it does not exhibit pauses within the core, but it allows for variation and elongation at the end, preferably introduced by a question word (e.g. why, what, where, how). For example, four out of five instances of speaker MS001, using the form < > I don’t know __ < >, are followed by *why*. Speaker MS003 also prefers why (in 6 out of 14 instances). That raises the question as to whether *I dont know why* may even be regarded as a separate MWE. In fact, considering all results and the distribution of pauses, one could also argue that there may be several different MWEs:

- I don’t know
- I don’t know wh=
- I don’t know why
- I don’t know why but
- I don’t know if
- I don’t know [the (NP)]
- but I don’t know

Biber et al. (1999:1002), studying *lexical bundles*[6] also found plenty of such structures. For example, they find that the structure *personal pronoun + lexical verb phrase (+ complement–clause fragment)* - which fits most of the above examples – is very common in conversation. They also record many of the examples listed above in their category of four-word bundle expressions with *I + know.* (ibid.). However, whereas their analysis is based on frequency information alone, the very rare use of pauses between *I don’t know* and the subsequent word(s) gives more confidence in that these strings are actually valid units from *two* perspective, that of frequency and holistic storage.

4.2 ‘I think I’

Forty four instances of *I think I* have been annotated. The pause distribution within these examples is as follows:

Pause distribution	**MS001**	**MS003**	**Σ**
MW< > E	5	3	**8**
< > MWE < >	1	3	**4**
< > MWE __ < >	5	7	**12**
< > __ MWE < >	--	3	**3**
< > __ MWE__ < >	5	12	**17**

Table 4. Pause distribution around 44 instances of *I think I*

I think I had been chosen for analysis because – intuitively – it does not seem to be a holistically stored MWE. Especially in comparison with no single pause occurring within 47 *I don’t know’s* the results seem to (at least partly) confirm this. Eight out of 44 examples do exhibit pause phenomena in *I think I* which is a first indicator that probably not all instances of *I think I* are stored holistically. A closer assessment of the eight MW< >E instances reveals that all but one exhibit the pause after *I think.* This is not surprising as *I think* is the most frequent occurring bi-gram in the data (almost 3000 instances in the 290,000 word learner corpus and 3 times more frequent as the second most frequent bi-gram *you know*). In fact, *I think I* could be regarded as a sub-unit of *I think*, similar to the relationship between *I don’t know* and *I don’t know*

[6] The definition of lexical bundles is essentially based on frequency - they are ‘sequences of words that most commonly co-occur in a register.’ Furthermore, Biber et al. observed that ‘most lexical bundles are not structurally complete at all’ (Biber et al. 1999:989).

why. Thus, the eight instances with pause breaks may be actually instances of the MWE candidate *I know* where *I* happens to mark the beginning of the next clause.

Interestingly, all 44 instances are followed by a full clause, which has the second *I* of *I think I* as the subject at the beginning of the new clause. In addition, *I think* seems to be used rather in the function of filler, possibly in order to provide *thinking* time for the next utterance. This happens extensively in the eight *I think* < > *I*____ cases where *I think* is followed by a pause. However, and as discussed earlier, the absence of a pause does not necessarily mean the absence of a MWE boundary. Therefore the 17 < > __ I think I __ < > cases and the 12 < > I think I __ < > cases may follow the same pattern with using *I think* as a filler. In these instances no further pause is necessary. However, this does not explain the 7 instances where pauses do occur at the end of *I think I*. Idiosyncratic usage might be one explanation as it is mainly a feature used by MS003 (6 times) and the only instance of MS001 coincides with a false start. Further investigations using a larger data-set might be able to confirm whether this pattern is due to idiosyncratic usage.

4.3 Summary and limitations

The analysis of pauses in our data would suggest that *I don't know* might be stored holistically while it is questionable that this is the case for *I think I* which is interrupted by pauses in some of the instances that were investigated.

In terms of the delineation of boundaries, it can be said that pauses are only helpful to a limited extent as boundaries are not conditional on them. The absence of a pause does not exclude the possibility that it might in fact be a boundary. However, where pauses occur they give valuable indications of possible boundaries. The results can give useful information on actual MWE usage to fields such as lexicography, (second/computational) language acquisition and teaching.

These initial findings are encouraging, but they are nevertheless based on limited data in terms of the number and forms of MWEs investigated, and also the number of speakers considered.

Future research should thus draw on more instances by different speakers in order to determine idiosyncratic usage and to arrive at more stable patterns. A comparison with native speaker usage seems crucial and promising for a more comprehensive description of MWEs.

In addition, studying intonation and stress patterns of these instances may indicate boundaries more clearly.

Finally, MWEs may be used in more than one sense, as in the case of *I don't know* which has to be considered for each different MWE candidate individually.

5 Conclusion: Value for NLP and future work

In this paper we have reported on a study which combines approaches within NLP for the identification of MWE candidates with pause analysis. The aim was to explore an approach which might lead to a frequency-based and psycholinguistically motivated description of MWEs.

The results of our study seem to suggest that the placement of pauses might be valuable as an additional criterion for the identification of holistically stored MWEs, however, larger data-sets and further pause annotation is necessary to confirm our initial findings.

Further investigations of other functions of pauses and other prosodic features within a given stretch of discourse need to be carried out in order to fully assess the role of pauses in relation to holistic storage. A discourse functional analysis would be necessary to identify functional motivation of pauses and to delineate these from n-grams where the placement of pauses is related to holistic storage.

However, our study has illustrated the potential of a multi-method and interdisciplinary approach to the identification and description of MWEs which may eventually be necessary to overcome some of the problems within NLP in terms of developing extraction methods, and some of the problems in descriptive linguistics and discourse analysis in terms of gathering evidence for different MWEs in use.

Acknowledgement

The research described in this paper is supported by the Engineering and Physical Science Research Council (EPSRC, grant EP/C548191/1). We would also like to thank the three anonymous reviewers for their comments on an earlier draft of this paper.

References

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of spoken and written English*. Harlow: Longman

Winnie Chen, Chris Greaves and Martin Warren. 2006. From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics* 11(4): 411-433.

Britt Erman. 2007. Cognitive processes as evidence of the idiom principle. *International Journal of Corpus Linguistics* 12(1): 25-53.

Britt Erman and Beatrice Warren. 2000. The idiom principle and the open choice principle. *Text* 20(1): 29-62.

Alexandre Gil and Gaël Dias. 2003. Using Masks, Suffix Array-based Data Structures and Multidimensional Arrays to Compute Positional Ngram Statistics from Corpora. In: Proceedings of the ACL 2003 'Workshop on Multiword Expressions: Analysis, Acquisition and Treatment', Sapporo, Japan 12th July 2003, 25-32.

Frieda Goldman-Eisler. 1968. *Psycholinguistics: experiments in spontaneous speech*. London, New York: Academic Press.

Tina Hickey. 1993. Identifying formulas in first language acquisition. *Journal of Child Language* 20:27-41.

George A Miller. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. *The Psychological Review* 63(2):81-97.

Andrew Pawley. 1986. *Lexicalization*. In: Deborah Tannen and James E. Alatis (eds.). Language & Linguistics: The interdependence of theory, data & application. Georgetown University Round Table on Languages & Linguistics 1985, 98-120.

Andrew Pawley and Frances Syder. 2000. *The One-Clause-at-a-Time Hypothesis*. In: Heidi Riggenbach (ed.). Perspectives on fluency. Ann Arbor: University of Michigan Press, 163-199.

Manfred Raupach. 1984. *Formulae in Second Language Speech Production.* In: Hans W. Dechert, Dorothea Möhle and Manfred Raupach (eds.). Second Language Productions. Tübingen: Narr, 114-137.

John Read and Paul Nation. 2004. *Measurement of formulaic sequences.* In: Norbert Schmitt (ed.). Formulaic Sequences. Amsterdam: John Benjamins, 23-35.

Heidi Riggenbach. 1991. Towards an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes,* 14: 423-441.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger. 2002. *Multiword expressions: A Pain in the Neck for NLP*. In: Proceedings of the 3rd International Conferences on Intelligent Text Processing and Computational Linguistics (CICLing-2002), Mexico City, Mexico, 1-15.

Antoine Tremblay, Bruce Derwing, Gary Libben and Chris Westbury. 2007. *Are Lexical Bundles Stored and Processed as Single Units?* Paper presented at the 25th UWM Linguistics Symposium on Formulaic Language. Milwaukee, Wisconsin, April 18-21, 2007

Richard Towell, Roger Hawkins and Nives Bazergui 1996. The development of fluency in advanced learners of French. *Applied Linguistics,* 17(1):84-119.

Diana Van Lancker, Gerald J. Canter and Dale Terbeek. 1981. Disambiguation of Ditropic Sentences: Acoustic and Phonetic Cues. *Journal of Speech and Hearing Research*, 24:330-335.

Alison Wray. 2002. *Formulaic Language and the Lexicon*. Cambridge, CUP.

Alison Wray. 1999. Formulaic language in learners and native speakers. *Language Teaching*, 32:213-231.