The study of second language speech fluency, pauses and MWUs – Background and Methodology[1]

Irina Dahlmann
(November 2007)

In this section some light will be shed on terms and relationships which are important for the project and the development of methodological aspects.

What are MWUs?

Multi-word unit (MWU) is an umbrella term for sequences of interrelated words. Several other terms abound, (Wray, 2002) lists some 50 different ones, including formulaic language, formulaic expressions, lexicalised phrases, formulas, chunks *etc.* which are used to describe (aspects of) formulaicity. MWUs are omnipresent in the English language (e.g., Pawley and Syder 2000; Erman and Warren 2001; Wray 2002).

The description of MWUs is commonly carried out according to different characteristics, for example, their varying degrees of syntagmatic fixedness, pragmatic functions or their frequency of co-occurrence in discourse. In terms of (second) language acquisition the appropriate use of MWUs is seen as a key skill for the development of fluent speech (e.g., Pawley and Syder 1983, 2000; Wray 2002). This is because MWUs are believed to be stored holistically in the mental lexicon and thus ease processing efforts of the speaker (see below).

However, definitions and reliable criteria for the identification of MWUs pose two of the major problems in the field (Nattinger and DeCaricco 1992; Schmitt and Carter 2004; Weinert 1995; Wray 2002). The main reason is that the phenomenon of MWUs in itself is extremely diverse in form and function.

MWUs and fluency

Holistic storage and retrieval from memory of MWUs is one of the characteristics which seems very promising in enhancing our understanding of MWUs and of fluency. This is based on the interrelation presented in Figure 1.

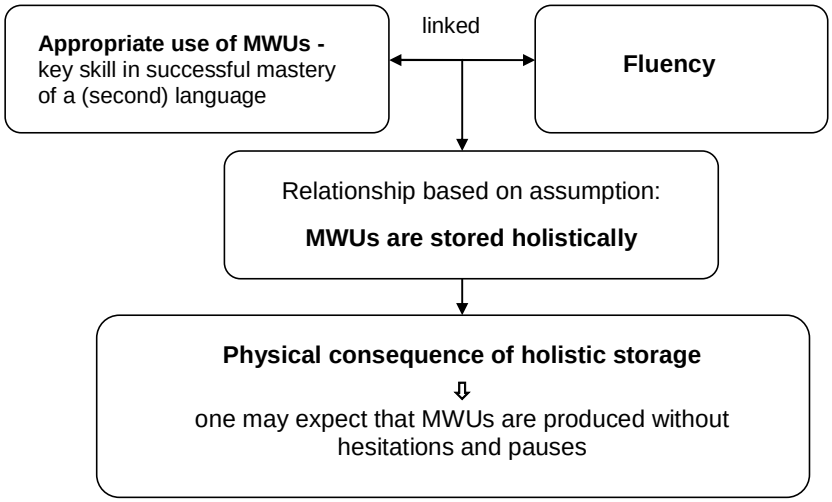

Figure 1: Overview of the relationship between MWUs, fluency and pauses

According to Pawley and Syder (1983, 2000), for example, the appropriate use of MWUs seems to enhance fluency. This is based on the assumption that MWUs are stored holistically as single units and therefore free processing time which allows the production of longer, thus more fluent, utterances. Miller's (1956) observations in his paper *The magical number seven* supports this argument. He states 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). MWUs, such as '*as soon as possible'* can be seen as one large holistically stored chunk, one unit, although containing several words.

Where pauses come in

The notion of holistic storage of MWUs as single chunks also permits for the supposition that MWUs are uttered in a fluent manner, i.e. without the exhibition of pauses. Pawley (1986) declares that 'pauses within lexicalised phrases are less acceptable than pauses

[1] Published on the project website 'Second language speech fluency (SLSF) and the role of pauses in automatically extracted multi-word units (MWUs)' https://www.nottingham.ac.uk/english/research/cral/doku.php?id=projects:slsf

within free expressions and after a hesitation the speaker is more likely to restart from the beginning of the expression' (p. 107, in Wray, 2002).

Although pauses are not conditional on MWUs and their boundaries (Dahlmann and Adolphs 2007) it has been suggested that pauses and, conversely, phonological coherence might be an additional factor in the identification process, if they do occur. Hickey, for example, proposes amongst other criteria 'phonological coherence' as a necessary condition for formula identification in first language acquisition (Hickey, 1993: 32; see also Peters, 1983) and Weinert (1995) lists phonological coherence as one of the criteria commonly used in the identification of formulaic language in relation to language acquisition (p.182).

What has been studied so far?

Apart from one recent exception (Erman, 2007) the few studies on pauses and MWUs so far have been carried out only on limited data sets (e.g., Peters, 1983; Raupach, 1984), limited referring here to the number of speakers, the quantity of speech and therefore consequently to the number of MWUs. But especially pause patterns around many instances of one MWU - ideally produced spontaneously and by different speakers – in order to gain a better picture of certain MWUs *in use* have not been studied in an empirical manner. This becomes possible with a corpus approach.

For example, pauses within and around many instances of a MWU candidate such as the very frequent *I don't know* could indicate whether this sequence is predominantly used with phonological coherence and without pauses and thus, may be stored holistically. With this quantitative approach *patterns* of usage may be observable. In a more qualitative approach these would go unnoticed, as there are normally not enough instances of the same MWU to find patterns, especially in spontaneously produced speech.

The approach used here enables us for example to reconsider MWU boundaries. They might be classified differently with the additional pause annotation, as in the example of *I don't know* (Dahlmann and Adolphs, 2007). MWU boundaries are not always as clear cut as automatic extraction methods suggest. In this study almost 50 instances of the second most frequent 3-word cluster *'I don't know'* (out of 327 instances) of our learner corpus NICLE(s) were considered. *I don't know* appears to be a self contained unit in the top 10 list of 3-word cluster. But it turned out that 'I don't know' may be a core

unit which can be elongated (e.g. *I don't know why, I don't know how* etc.) without exhibiting pause interruptions and thus these instances may even be regarded as different units (ibid.).

Moreover, in the same study none of the examples exhibits a pause within the phrase *I don't know*, which is seen as indication that this phrase may be stored holistically in the mental lexicon, especially when compared to the equally frequent 3-word cluster *I think I'* which contained internal pauses in 8 instances. With this approach we may gain a more rounded picture of the actual use of particular MWUs.

These results will are presently being compared to native English speech in order to study differences between native speakers and learners. This will inform the study of formulaic language regarding its definition and identification and form the basis for the study of potential differences and similarities in usage of formulaic language in learners and native speakers.

To sum up, the overall aim of the project is to investigate how best to merge the characteristic of holistic storage of multi-word units with the automatic extraction of such units from corpora of native speaker and learner English so as to enhance our understanding of

- Second language speech fluency
- The difference between native speaker and non-native speaker use of multi-word units
- The juncture profiles of automatically extracted multi-word units and thus the overlap between the psycholinguistic conceptualisation of multi-word units and their statistically-based extraction.

Methodological challenges, approach and research questions

The reasons for the negligence of this kind of study lie mainly on the methodological side.

- **Lack or suitable corpora and pause coding**

A corpus approach may be used to address the overall lack of progress in this area as it lends itself to this type of analysis, but there is a lack of spoken corpora suitable for this kind of investigation as spoken corpora have to fulfil certain criteria. They need to consist of spontaneous speech, as 'the essential nature of language (…) is most clearly revealed

in the unselfconscious activity of speaking' (Halliday 2004:25) and that is most likely achieved in a dialogue situation (ibid.). Furthermore, corpora need suitable coding of pauses which complies with practices used in fluency research. Coding schemes should for instance include silent and filled pauses as well as appropriate measurements of pause lengths.

Of the relatively small number of spoken (British) English corpora and Learner English Corpora (all in all, 34 were surveyed here), only very few contain real spontaneous speech (e.g. CANCODE, LLC). Sometimes the type of speech is mixed or it is not obvious from the information available what type of language has been collected. Occasionally a corpus claims to contain spontaneous speech but it may be only valid for parts of the corpus. For example, it has been claimed for the BNC that its spoken part consists of natural, spontaneous conversations, but in fact the lectures and chat shows which are included may be (partly) scripted.

Regarding pause coding, very few corpora contain pause coding. A survey of 12 corpus pause coding schemes of spoken corpora shows that none complies with the requirements needed for the study of fluency and MWU related research (Dahlmann and Adolphs 2007:51).

- **Manual implementation of pause coding**

Due to the lack of corpora which combine spontaneous speech and appropriate pause annotation we have decided to use two corpora of spontaneous speech available at Nottingham and carry out the pause annotation from scratch, that includes the developing a suitable pause coding scheme as well as the (manual) implementation into the two corpora. This decision has far-reaching consequences for the further development of our demonstrator project. For instance, the pause annotation will only be carried out in the immediate surroundings of the chosen MWU candidates instead of annotating both corpora in full. This restricts the number of MWU candidates which can actually be studied and compared.

The two corpora used for the present project are ENSIC and NICLE(s). Both consist of interviews between two speakers, the interviewer and interviewee. We are only interested in the interviewees' speech as it is spontaneous speech, whereas the interviewer's speech may be at least partly scripted.

- **Limitations**

This decision has far-reaching consequences for the further development of our demonstrator project. For example, due to practical constraints of manual coding the pause annotation will only be carried out in the immediate surroundings of the chosen MWU candidates which this restricts the number of MWU candidates which can actually be studied and compared.

However, this approach also allows for a test phase of the pause coding scheme. The ideal version of the coding scheme comprises a range of pause phenomena and related information such as overlap and utterance coding (see Adolphs, Dahlmann, Rodden 2007 and the website for more information on the pause coding scheme) in order to represent the speech situation as precise as possible. The annotation currently carried out strives to be very detailed, that includes the measurement of pauses, the marking-up of repetitions, repairs and false starts and the marking of overlapping speech. That means in practical terms that each stretch of which is being marked-up according to our scheme undergoes several rounds of annotation and checking in this process. That is very timely and will be evaluated and refined in terms of practicability and appropriateness with regard to large scale application.

- **The role of frequency**

As mentioned above, pauses alone cannot be used for the identification of MWUs; however, several researchers have suggested or reported lists of criteria for MWU identification and advocate a combination of several identification criteria for MWUs (Hickey, 1993; Peters, 1983; Wray and Namba, 2004; Weinert, 1995). The aspect of *frequency* or *community-wide use* of a formula is recurrent in these lists: 'Frequent and unchanging use is generally cited as a characteristic of the prototypical formula, and it was included in Peter's and Wong Fillmore's sets of criteria' (Hickey 1993:33).

The idea of recurrent and frequent word clusters and collocations is key in a corpus approach. 'If two words occur together a lot, then that is evidence that they have a special function that is not simply explained as the function that results from their combination' (Manning and Schütze 1999:153). This has been taken up for the development of statistical approaches and automatic extraction methods for the detection and extraction of collocations and MWUs. The use of a frequency approach as a first criterion for a MWU

candidate combined with the further pause analysis seems appropriate for an exploratory study of the role of pauses within MWU research.

Summary of the methodological approach & research questions

The methodological approach for our project can thus be summarised as follows.
In the exploration of pause patterns within and around MWUs, MWU candidates are being extracted automatically with two different automatic extraction methods from the two target corpora. The first tool for the extraction is *Wordsmith Tools* (extracting *clusters; Scott, 2004*) and the second tool is *WMatrix2 (*extracting *multi-word expressions (MWEs*); Rayson, 2001-7). Wordsmith Tools works with a corpus-driven approach, whereas Rayson's MWE extractions are based on a semantic lexicon which contains more than 18,000 MWE templates (Piao et al. 2003, 2005a, 2005b).

Only the top 10 most frequent candidates will be considered for the further study of actual pause placements and pause patterns around them. This is not only due to feasibility reasons but also due to the fact that only a large amount of instances per MWU candidate allows for a reasonable generalisation of pause patterns of these candidates. This is especially appropriate as this is the first study of this kind.

This approach allows us to address the following **research questions**:

1. Are statistical procedures for the extraction of MWUs an apt mirror of the mental processes associated with the storage and retrieval of multi-word units?

It can be expected that MWUs are not interrupted by pauses (silent or filled), as it is assumed that MWUs are stored and retrieved holistically from memory. The question is whether this also holds for MWU candidates which have been extracted automatically with techniques which are not designed in the first place to extract word strings which are stored holistically, but based on other characteristics of MWUs (e.g. frequency).

2. What type of pause phenomena occur within and at the boundaries of automatically extracted MWUs and how do they relate to mental processes?

The following five cases of placements of pauses are possible:
('____'indicates text which can theoretically be of any length, < > indicates pause phenomena)

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

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

- Pause placement pattern a) is regarded as being informative with respect to holistic storage (mental processing). If a MWU candidate is repeatedly interrupted by a pauses, it is questionable whether it is stored holistically, however if a candidate is never interrupted by pauses it is a good indicator of holistic storage.

- Cases b – e are informative in terms of the delineation of boundaries. One question is whether pauses align MWEs in the form in which they were extracted. 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, if pauses occur they may give valuable indications of possible boundaries.

Furthermore, and especially for cases b – e the interest is on a descriptive analysis on any regular pause patterns emerging from the data, for example whether there is a link between the placement of pauses and the type of pauses, such as 'the majority of pauses directly preceding the MWU candidate are pauses of a certain kind (e.g. silent/filled/lengthy pauses, pause clusters)'. A further question which has to be touched on is whether such patterns – if they exist – are universal or tight to particular MWUs.

3. Are the pause phenomena different in the performance of native speakers to those of non-native speakers in and around the same set of MWUs? And if so how can such a result be explained in terms of language acquisition and how does it affect SLSF?

Question 2 is being explored for both, native speaker and learner use of certain MWU candidates. This follow-up question is being concerned with differences and similarities of these results.

Possible differences may be explained drawing on second language acquisition and speech fluency theory.