Detecting Multi-word Expressions Through Typing Patterns

Multi-word Expressions (MWEs) have been cited as one of the great remaining problems for Natural Language Processing (Sag, et al. 2002). Unlike literal expressions, in which meaning can be composed syntactically and semantically, an MWE shows markedly peculiar linguistic behavior in terms of lexicalization, since the meanings of the individual lexical items do not compose to the meaning of the expression (Kunchukuttan 2007). Examples of MWEs include *kick the bucket* and *shoot the breeze*. When processing a text, a parser must decide if *kick the bucket* is being used in the context of physically moving a metal container, or as a holistic phras. We have devised a novel method to detect MWEs, when analyzing a dynamically typed text, in which timing data is made available.

To study the phenomena of dynamically-typed text, we extracted temporal and linguistic data from 1,013 typing sessions. The exact timing of each key press was recorded, from when a key was depressed, to when it was released. We then extracted the pausality, or latency, before each word was typed, and categorized each word into one of four categories, relative to MWEs: Outside an MWE, Starting an MWE, Within an MWE, and Ending an MWE. We then "normalized" each preceding pause time by measuring the predictability of the word sequence it occurred within. In other words, given the preceding word, we calculated the likelihood that this word will follow. This is in line with creating n-gram language models.

The results of our analysis appear in *Figure 1* on the following page. We divided subjects further by whether they were L1 speakers of English, or L2, and then by whether they were considered "fast" or "slow" typists. (Units of measurement are difficult to define in *Figure 1*, as numbers were calculated as pause time divided by sequence predictability. We propose the term *predictabilized pause*.) Perhaps the most salient trend in *Figure 1* is the markedly shorter pause duration occurring within an MWE, relative to pauses starting, ending or outside of an MWE. *Figure 2*, below, is a visual representation of pausality relative to an MWE, where a longer underscore signifies a longer pause time.

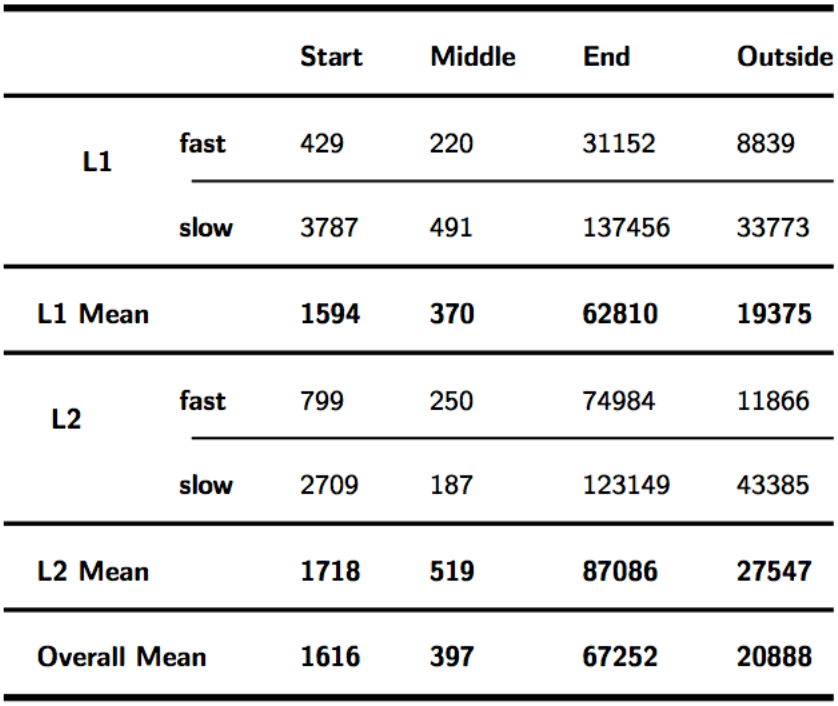
She____looked_up____the____word_record

Figure 2

Our findings closely parallel similar findings in speech production. Dahlmann and Adolphs (2007) also found that speakers pause for greater lengths of time outside of an MWE, relative to within an MWE.

Our current findings have broad applications, from Information Retrieval to Machine Translation. As a (somewhat contrived) example, suppose a search engine is presented the query *How to handle a pain in the neck*. A search engine must decide whether to return results for local chiropractors, or stress relief tactics. By measuring the temporal data surrounding each term, a search engine could retrieve more relevant results. Additionally, our methods allow for the possibility of detecting MWEs organically, rather than from a pre-compiled dictionary of MWEs. This would be especially useful in translating low-resource languages, where precompiled dictionaries are not readily available. We believe our findings represent an exciting development in NLP.

Preceding Pause Duration, Relative to MWE Location



		Start	Middle	End	Outside
L1	fast	429	220	31152	8839
	slow	3787	491	137456	33773
L1 Mean		**1594**	**370**	**62810**	**19375**
L2	fast	799	250	74984	11866
	slow	2709	187	123149	43385
L2 Mean		**1718**	**519**	**87086**	**27547**
Overall Mean		**1616**	**397**	**67252**	**20888**

Figure 2

References

Dahlmann, I., & Adolphs, S. (2007, June). *Pauses as an indicator of psycholinguistically valid multi-word expressions (MWEs)?*. In Proceedings of the Workshop on a Broader Perspective on Multiword Expressions (pp. 49-56). Association for Computational Linguistics.

Kunchukuttan, A. (2007). *Multiword Expression Recognition* (Doctoral dissertation, Indian Institute of Technology, Bombay).

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). *Multiword expressions: A pain in the neck for NLP*. In Computational Linguistics and Intelligent Text Processing (pp. 1-15). Springer Berlin Heidelberg.