Muddying The Multiword Expression Waters: How Cognitive Demand Affects Multiword Expression Production

Adam Goodkind
CUNY Graduate Center
agoodkind@gradcenter.cuny.edu

Andrew Rosenberg
CUNY Queens College
andrew@cs.qc.cuny.edu

Abstract

Multiword expressions (MWEs) are vexing for linguists, psycholinguists and computational linguists, as they are hard to define, detect and parse. However, previous studies have not taken into account the cognitive constraints under which MWEs are produced or comprehended. We present a new modality for studying MWEs, keystroke dynamics. We ask subjects to respond to a variety of questions, varying in the level of cognitive demand required to generate an answer. In each response, a subject's pause time preceding each word – within and outside an MWE – can illuminate distinct differences in required effort across tasks. By taking advantage of high-precision keystroke loggers, we show that MWEs produced under greater cognitive demands are produced more slowly, at a rate more similar to free expressions. We hypothesize that increasingly burdensome cognitive demands diminish the capacity of lexical retrieval, and cause MWE production to slow.

1 Introduction

Multi-word expressions (MWEs) are vexing for both theoretical linguists and those working in Natural Language Processing. For theoretical linguists, MWEs occupy a liminal space between the lexicon and syntax (Langacker, 2008). For NLP practitioners, MWEs are notoriously difficult to detect and parse (Sag et al., 2002).

This paper presents a new modality for studying MWE production, keystroke dynamics, which allows for large-scale, low-cost, high-precision metrics (cf. (Cohen Priva et al., 2010)). Keystroke dynamics looks at the speed at which a user's hands move across a keyboard (Bergadano et al., 2002). It has the distinct advantage of using written text, with clear word and sentence boundaries, while combining it with dynamic production features, allowing for greater insight into the language creation process.

This study explores the notion that many of the principles that guide intonation and speech prosody are also present during the typing production process. Principles related to prosody need not be limited to spoken language production. The *Implicit Prosody Hypothesis*, for example, posits that a "silent prosodic contour" is projected onto a stimulus, and may help a reader resolve syntactic ambiguity (Fodor, 2002). Previous studies applied this hypothesis to silent reading (Fodor, 2002). The present study, in turn, applies this same principle to (silent) typing: Language users take advantage of prosodic contours to help organize and make sense of language stimulus, whether in the form of words they are perceiving or words they are producing.

Moreover, in previous studies, the *type* of question a subject is asked, in order to elicit a response, has not been taken into consideration. We take advantage of the low cost and high precision of keystroke dynamics to uncover trends in MWE production, by eliciting responses from subjects using a variety of questions with very different cognitive demands. Our findings show that the cognitive demands of an elicitation task have a noticeable effect on how MWEs are produced during a response. These findings have important ramifications for linguists performing MWE-related experiments, and cognitive scientists studying how lexical items are stored and retrieved.

In order to run our analysis, we collected free response typing data from a large set of subjects. The subjects responded to a wide array of cognitively demanding prompts, from simple recall to more complex, creative analysis. From this data, we then perform two experiments. In a preliminary experiment, we analyze how linguistic attributes such as word length and predictability shape keystroke produc-

Proceedings of NAACL-HLT 2015, pages 87–95,
Denver, Colorado, May 31 – June 5, 2015. ©2015 Association for Computational Linguistics

tion. In our main experiment, we then use these findings to analyze how multiword expression production is affected by the cognitive demands imposed upon the subjects.

We hypothesize that the cognitive demands of a task will impede MWE production, as the overall demands will interfere with lexical retrieval, creating a cognitive bottleneck. Our study aims to shed light on three sets of questions:

- Are MWEs produced differently depending upon the type of task they are produced within? If so, how?
- Can patterns in MWE production provide insights regarding constraints on lexical retrieval?
- What are the benefits of keystroke dynamics for psycholingistics studies?

The rest of the paper is organized as follows: Section 2 situates our study in context, illustrating how prosody is affected by MWEs, and keystroke dynamics relates to cognition. Section 3 outlines our experiments, with results reported in Section 4. Our results are discussed in Section 5, with a conclusion and look towards future work in Section 6.

2 Related Work

Our study brings together MWEs, cognition and keystroke dynamics in a novel manner. In order to situate our investigation in context, we explore relevant previous studies below, and explain how their findings contribute to the present work.

2.1 MWEs in speech production

Many studies have concluded that multiword expressions are stored and retrieved as single lexical units (Wray, 2005; Dahlmann and Adolphs, 2007, and references therein). As such, MWEs exhibit unique phonological and prosodic characteristics. For example, MWEs have been found to exhibit greater phonological consistency than free expressions (Hickey, 1993). Specifically, pauses have been found to be less acceptable in lexicalized phrases (Pawley, 1985). In addition, and most relevant to our study, Dahlmann and Adolph study how pausality differs in and around MWEs (Dahlmann and Adolphs, 2007). They conclude that "...where pauses occur they give valuable indications of possible [MWE] boundaries". (Dahlmann and Adolphs, 2007, p. 55)

In many ways, the present study can, and should, be viewed as an extension of Dahlmann and Adolphs' study. If we view keystroke dynamics as a reflection of many speech production principles in the typing process, then this is a reasonable extension. We augment the previous findings, though, by investigating how varying cognitive demands affect MWE production.

In studies of speech, Erman (2007) notes that a pause can be caused by the cognitive demands of lexical retrieval, and Pawley (1985) notes that pauses are much less acceptable within a lexicalized phrase than within a free expression. This led Dahlmann and Adolphs (2007) to study pausing within spoken MWEs. A central finding of Dahlmann & Adolphs is that MWEs are often surrounded by pauses, and that pausality is unique within and around MWEs.

In addition, Dahlmann & Adolphs note the difficulty of accurately measuring pauses in speech; keystroke dynamics does not face that obstacle.

2.2 Typing Behavior and Cognition

Typing is an interesting blend of cognitive and physical activity. On the cognitive side, a typist must undertake the cognitively demanding task of text production. Although literate people produce text on a nearly daily basis, researchers have gone so far as to call the writing process "one of the most complex and demanding activities that humans engage in" (Alves et al., 2008, p. 2). The act of typing involves juggling both the high-level text creation process, and low-level motor execution.

Beginning in the 1980s (Rumelhart and Norman, 1982), investigators used typing data to construct cognitive and motor models of language production. As expounded by Salthouse (1986), a typist must simultaneously employ multiple cognitive and motor schemata, often with a formidable amount of noise between signals. Translating from lexical retrieval into physical action is a non-trivial task, which involves multiple pipelines that can be occluded, and also result in mixed up signals.

The typing task is especially daunting for novice typists. Gentner, et al. (1988) investigated the linguistic characteristics of skilled versus unskilled typists, finding marked differences in the behavior (and thus cognitive model) of each population. A novice typist is so burdened by the physical execution cycle of typing that the quality of his or her writing is noticeably diminished.

However, Alves et al. (2008), in studying narrative construction in typing conclude that while differences do exist between the populations, this might not be as significant a differentiation as originally thought. They conclude, "Although motor execution is more demanding for slow typists, this higher demand neither prevented them from activating high-level processes concurrently with typing, nor changed the distribution of occurrences of the writing processes." (Alves et al., 2008, p. 10)

The importance of pauses during the typing process is borne out in a number of studies. Schilperoord (2002) concludes that writers pause for a number of reasons, such as cognitive overload, writing apprehension or fatigue. Alves et al. (2008) similarly concluded that pauses are usually a sign of cognitive competition. Many of the reasons given for pausing during typing are similar to the reasons given for pausing during speech production, thus providing further motivation to use the typing process to test phenomena observed during speech.

3 Methodology

3.1 Procedure

Our typing data was collected from 189 Louisiana Tech students (hereinafter referred to as "subjects"). The subjects reported themselves to be 41.3% female, 56.4% male and 88.3% right-handed and 9.1% left-handed. (Note that these do not sum to 100%; on each question some percentage of subjects chose not to respond to one or more of the demographic questions.)

We limited our study to only native English speakers. This was to avoid the additional confound of language familiarity, though this is certainly an important area for study. Specifically, Riggenbach (1991) found that in speech, placement and length of pausing around MWEs is seen as a sign of fluency.

Further, we limited our study to only "touch typists", or those subjects who only look at the screen when typing. This is in comparison to "visual typists" who look at their fingers when typing. As proposed by Johanssen et al. (2010), touch-typists and visual typists employ distinct cognitive models, as visual typists also need to dedicate cognitive effort to figuring out where the next key is. For touch typists, this is a less conscious process.

Subjects were seated at a desktop computer with a QWERTY keyboard, and freely responded to prompts of varying complexity. A keylogger with 15.625 millisecond clock resolution was used to record text and keystroke event timestamps. There was no time limit, although subjects had to type at least 300 characters before proceeding to the next prompt. Each subject responded to $10-12$ prompts, with the average response comprising 448 characters and 87 words.

Prompts were designed to test all aspects of Bloom's Taxonomy of learning (Krathwohl, 2002), from simple to more complex tasks. Bloom's Taxonomy includes six different types of tasks: *remember, understand, apply, analyze, evaluate* and *create*. The Bloom Taxonomy is ordered by complexity, in that mastery of one learning objective is necessary in order to progress to the next. It is a useful way for educators to structure a curriculum, in order to ensure that learners possess the necessary cognitive abilities before progressing to more complex tasks. The taxonomy has been refined and expanded in recent years; as such, we treat each type of task as a discrete type of task, rather than having a continuous relationship.

The order that the prompts were presented in was randomized, with an equal distribution from each type of task. Examples of prompts include (1) and (2):

(1) *List the recent movies you've seen or books you've read. When did you see or read them? What were they about?* [Remember]

(2) *How would you design a class if you were the teacher? What subject would you teach? How would you structure your course?* [Create]

The full data set is part of a long-term longitudinal study relating to subject biometrics. Although the current data is not publicly available, we hope to release future data sets.

3.2 Materials

All texts were tokenized using OpenNLP (Baldridge, 2005). We then automatically extracted all multiword expressions using jMWE (Finlayson and Kulkarni, 2011). For the present studies we only looked at contiguous MWEs. jMWE has reported an F$_1$ measure of 83.4 in detecting continuous, unbroken MWEs in the Semcor (Mihalcea, 1998) Brown Concordance (Finlayson and Kulkarni, 2011).

Contiguous MWEs should show more signs of being a cohesive lexical unit, although non-contiguous MWEs should still exhibit some degree of the same. As a result of this exclusion, MWEs such as *ran up* in (3) would be included in our study, while the same non-contiguous MWE in (4) would not.

(3) *Jack <u>ran up</u> the bill.*

(4) *Jill <u>ran</u> the bill <u>up</u>.*

While keystroke dynamics is concerned with a number of timing metrics, such as key holds (*h* in Figure 1) and pauses between every keystroke (*p* in Figure 1), the current study looked only at the pause preceding a word (the second *p* in Figure 1). This interval consists of the time between the spacebar being released and the first key of the word being pressed.

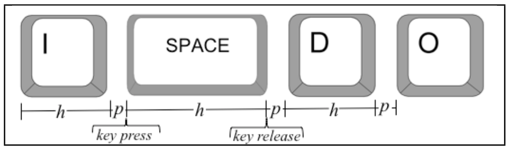

Figure 1: Timing Intervals in Keystroke Dynamics
p = pause *h* = hold

We also did not remove any outliers, although this is common in keystroke dynamics (Epp et al., 2011; Zhong et al., 2012). We feel it is difficult-to-impossible to discriminate between a "true" pause that is indicative of a subject's increased cognitive effort and any other type of pause, such as those caused by distraction or physical fatigue. As such we include any idiosyncrasies, such as long pauses, in our analyses rather than dismiss them as noise.

4 Experiments

4.1 Experiment 1: Creating A Baseline

In the main experiment, we measure the pause preceding each word. However, we wanted to remove as many confounds as possible that were not related to whether the word was part of an MWE.

Our first line of investigation aimed to understand the distribution of pauses overall. As seen in Figure 2, pauses are not distributed normally around a mean (non-Gaussian). Rather, there is a strong log-linear relationship between length of pause and frequency. As such, results reported below use the logarithm of the pause time. We felt that reporting the raw pause time would obfuscate important patterns within pausality.

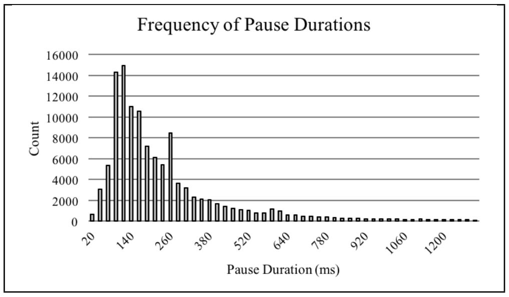

Figure 2: Distribution of All Pauses

As noted by Nottbusch & Weingarten (2007), the length of a written word affects pre-word pausing. We quantified this by mapping each pre-word pause to the length of the word, and found a strong logarithmic relationship, where pause length increased as a function of the log of the word length (see Figure 3). Since we expect cognitive demand to affect typing, we measured this affect on each task, and created different α and β parameters for our "Expected Pause" algorithm, as described in (5).

(5) $Pause_{expected}(w) = \alpha \cdot \ln(length(w)) + \beta$

The regression model illustrated in (5) provided a very reliable fit for all tasks. Between tasks α ranged from $0.107 - 0.112$ while β ranged from $2.20 - 2.24$. In the various versions of the Expected Pause algorithm, R^2 ranged from $0.93 - 0.98$ yet the differences were never significant, with $0.22 < p < 0.58$.

In our main experiment, all pauses were quantified as a deviation from the expected pause, based on word length and cognitive demand.

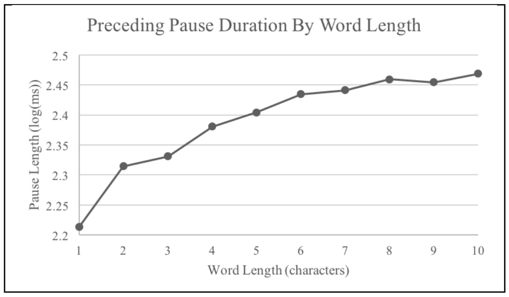

Figure 3: Duration of Pre-Word Pause By Word Length

A final confound to be investigated was sequence likelihood. The effects of predictability are well documented, in that more likely sequences are produced and comprehended at a faster rate (Goldman-Eisler, 1958; Hale, 2006; Nottbusch et al., 2007; Levy, 2008; Smith and Levy, 2013, and references therein). Since MWEs are frequently made up of collocations, i.e. words that are often seen together, they are inherently highly predictable.

For the present study, we wanted to ensure that we were not simply detecting faster rates of highly predictable sequences, but rather that we were detecting a signal idiosyncratic to MWEs. To test this, we grouped all word tokens according to the bigram predictability of the sequence they occurred within. Bigram predictability was calculated using a development set of users to create a language model. Smoothing was done using the Laplace technique with the inverse vocabulary size, as described in (6), where V is the total number of possible bigrams, i.e., the vocabulary size for a bigram model, and C is the total count of occurrences.

(6) $$P(w_n|w_{n-1}) = \frac{C(w_{n\text{-}1}w_n)+1}{C(w_{n\text{-}1})+V}$$

The grouping was done by rounding the log probability of the bigram sequence. We looked at the most highly predictable groups, to see if MWEs were still produced differently from free expressions, when compared to sequences of similar likelihood.

Our results are illustrated in Figure 4. Using a two-tailed t-test, and assuming equal variance, the differences for the two most highly predictable groups (where rounded log probability was -1 and 0) is significant at the 0.00001 level, while it is not significant for left-most grouping (rounded log probability of -2). The overall difference for all levels of predictability is significant at the 0.000001 level.

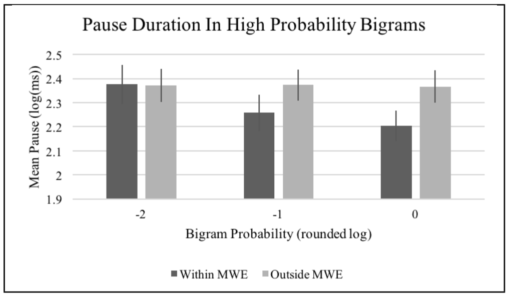

Figure 4: MWE Production in High Predictability Sequences.

4.2 Experiment 2: MWEs in Varying Cognitive Tasks

MWEs were produced at a fairly consistently rate across all tasks, comprising approximately $12-13\%$ of all word tokens, as reported in Table 1. It should be noted that this figure is markedly lower than often cited figures such as Erman & Warren (2000), who point out that half of spoken and written language comes from multiword constructions. In the present case, however, we are dealing with a small subset of MWEs, namely those that were produced contiguously (cf. examples (3) and (4) above). A total of $1,982$ different MWEs were produced, across the entire spectrum of "MWE types," from verb-particle constructions to idioms.

Task	Within-MWE Tokens	Outside MWE Tokens	Total Tokens	MWE Rate (%)
Remember	3,285	23,631	26,916	12.2%
Understand	3,986	25,008	28,994	13.7%
Apply	1,807	12,674	14,481	12.5%
Analyze	3,375	21,300	24,675	13.7%
Evaluate	4,957	35,290	40,247	12.3%
Create	3,629	24,042	27,671	13.1%
Total	**21,039**	**141,945**	**162,984**	**12.9%**

Table 1: MWE Production Rates and Counts By Task

Pauses that took place before the first word and directly after the last word of an MWE were not considered to be 'within' the MWE. An example of the pauses we *did* measure is seen in Figure 5. In this figure, the underscores represent measured pauses, while a whitespace gap represents a pause that was not taken into consideration for the present study. Pauses that occur on the edges of MWEs may represent distinct "barrier" pauses (Dahlmann and Adolphs, 2007), and therefore merit a further, but distinct study.

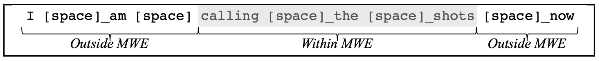
I [space]_am [space] calling [space]_the [space]_shots [space]_now
Outside MWE
Within MWE
Outside MWE

Figure 5: An example sentence. Measured pauses are represented with an underscore.

In each task, words within MWEs were consistently produced with a shorter preceding pause than were words in free expressions. As seen in Figure 6, pauses are shorter within MWEs across all tasks.

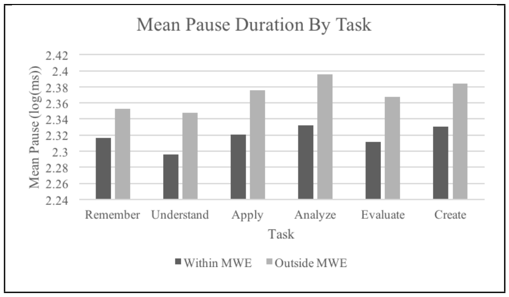

Figure 6: Pause Duration By Task, Within and Outside MWEs

However, the distributions of the means as reported in Figure 7 is not uniform[1].

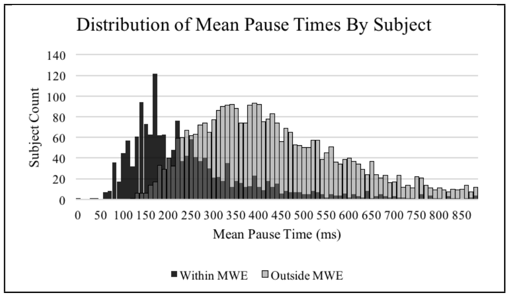

Figure 7: Distribution of Mean Pauses Within and Outside MWEs

Within-MWE pauses are not only shorter in duration, but we see evidence that the distribution is somewhat more concentrated around the mean. Although the standard deviations of each distribution are similar ($s_{within-mwe} = 197.5$, $s_{outside-mwe} = 209.8$), the interquartile ranges were more distinct ($IQR_{within-mwe} = 160$, $IQR_{outside-mwe} = 240$).

However, our investigation aimed to look at how pausing within MWEs varies between cognitive loads, rather than an overall distribution. These results are illustrated in Figure 8. A one-way between category ANOVA was conducted on the pause times, to compare the effects of cognitive demands on pausality. There was a significant effect of cognitive demand at the $p < 0.001$ level, $[F(5, 11796) = 4.19, p = 0.000815]$.

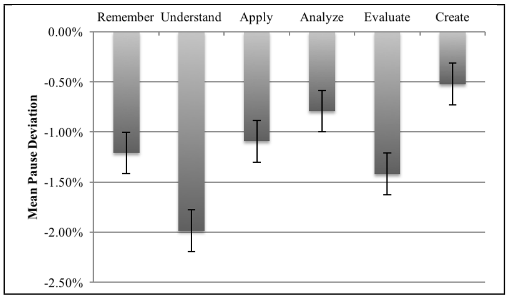

Figure 8: Within-MWE Pause Duration Deviation By Cognitive Task
(Tasks are arranged from (generally) simplest to most complex)

5 Discussion

As demonstrated above, the overall cognitive demands of a task have a significant effect on pauses within an MWE. While the trend is generally upward, in that MWEs produced under greater cognitive demand behave more similar to free expressions, i.e. they exhibit longer pauses, we note that this is not perfectly consistent. This is to be expected, as there are many dimensions to each of Bloom's tasks, and each dimension could have greater or lesser effects on pauses within typing. This could also be an artifact of the difficulty of assigning labels using Bloom's Taxonomy, as has been demonstrated even among a group of subject-matter experts (van Hoeij et al., 2004)

These results seem to demonstrate competing cognitive demands, operating in parallel. The canonical theory of MWE production holds that MWEs are retrieved as a single unit. Our results, however, imply that a more nuanced view may be

[1] Figure 7 took the mean pause per subject, rather than mean pause per word token, which is why it uses a linear scale, rather than a logarithmic scale.

justified. If an MWE is retrieved as a single unit, then somewhere between retrieval and execution the overall cognitive demands can interfere. Specifically, we theorize that the overall cognitive demands serve to narrow the bandwidth of lexical retrieval, occluding large units from being holistically moved into the executive buffer, as illustrated in Figure 9. To clarify this idea, though, subsequent investigations will investigate pauses at the boundaries of MWEs.

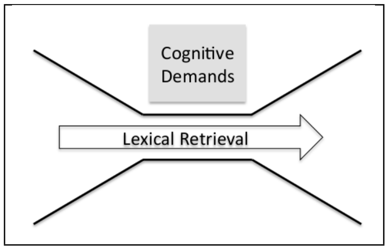

Figure 9: Model of Cognitive Bottleneck

The notion of various schemata interacting is supported by Kellogg (1996), who proposes that “resources from the central executive of Baddeleys model of the working-memory, e.g., Baddeley (1974), are needed to perform both lower-level writing processes such as spelling, grammar and motor movements and higher-level writing processes such as planning and revising.” (qtd. in Johansson, 2010).

By comparing the production rates of different types of lexical unit retrieved from working memory – MWES versus free expressions – along with varying the overarching cognitive task, we believe our experiment lends quantifiable support to this notion.

Our findings also bear relevance to investigators performing psycholinguistic experiments. Although most experiments are prepared with careful attention to the linguistic structure of stimulus, such as an elicitation prompt, there exists little attention to the overall cognitive demands a stimulus response requires. Our results, however, demonstrate that overarching cognitive demands can have a significant effect on results.

Finally, we hope our results serve as an illustration of the utility of keystroke dynamics within the linguistic and cognitive science domains. Many studies cite the difficulty of accurately transcribing speech data, delineating word boundaries and quantifying pause duration. Keystroke dynamics is not impeded by any of these factors. Additionally, although the data of this study was collected in a laboratory study, similar studies could be conducted using much less overhead, e.g. Amazon Mechanical Turk (Cohen Priva et al., 2010), where subjects can participate remotely without compromising experiment quality (Snow et al., 2008). This allows for low-cost, high-precision experimentation, with a wider selection of experiment participants.

6 Conclusion and Further Work

In this paper, we found that pauses within an MWE can vary significantly, depending upon the cognitive demands of the task within which they were produced. We first controlled for linguistic factors that affect typing rate, such as word length and predictability, and formed an Expected Pause metric. This metric measures the length of time we expected a subject to pause before a word, based on linguistic attributes. We then measured the divergence of pauses within MWEs, and found they varied significantly depending on the overarching cognitive task.

We believe our study represents a significant finding within MWE and lexical retrieval research. We have been able to directly quantify the effects of overall cognitive demand as it interacts with lexical retrieval. These results should be kept in mind when performing MWE research, as they clearly demonstrate that MWE production can be significantly affected by the cognitive complexity of a task, even if the method of elicitation is kept consistent.

A potentially important factor in MWE production is “MWE type,” such as verb-particle construction or idiom. Vincze et al. (2011) found useful differences between types, as they relate to MWE identification. Similarly, Schneider et al. (2014) classified MWEs using “strong” and “weak” dimensions, depending on “the strength of association between words...ranging from fully transparent collocations to completely opaque idioms (Hermann et al., 2012)” (Schneider et al., 2014, p. 456). Future studies will investigate the effects of these dimensions on the dynamics of MWE production.

Subsequent studies will also look into other elements of MWE production, such as errors (typos) produced within and outside of MWEs. In the cognitive science tradition, errors are a telltale window into the mind's inner workings.

Finally, we will expand our investigation to all intervals surrounding and within an MWE. Similar to Dahlmann & Adolphs (2007), we will investigate pauses at the beginning and end of a multi-word expression. In addition, we will investigate non-contiguous MWEs, to determine how their production differs from contiguous MWEs.

Acknowledgements

Thank you to Janet Fodor, Uriel Cohen Priva and David Guy Brizan for many discussions, feedback and edits. Thank you also to three anonymous reviewers for your feedback and suggestions. This work was supported in part by DARPA Active Authentication Phase I grant FA8750-12-2-0201 and FA8750-13-2-0274.

References

Rui Alexandre Alves, Sao Luis Castro, and Thierry Olive. 2008. Execution and pauses in writing narratives: Processing time, cognitive effort and typing skill. *International journal of psychology*, 43(6):969–979.

Alan D Baddeley and Graham Hitch. 1974. Working memory. *Psychology of learning and motivation*, 8:47–89.

J. Baldridge. 2005. The opennlp project. `www.opennlp.sourceforge.net`.

F. Bergadano, D. Gunetti, and C. Picardi. 2002. User authentication through keystroke dynamics. *ACM Transactions on Information and System Security (TISSEC)*, 5(4):367–397.

U Cohen Priva, S Ohlsson, and R Catrambone. 2010. Constructing typing-time corpora: A new way to answer old questions. In *Proceedings of the 32nd annual conference of the cognitive science society*, pages 43–48.

Irina Dahlmann and Svenja Adolphs. 2007. Pauses as an indicator of psycholinguistically valid multi-word expressions (mwes)? In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, pages 49–56. Association for Computational Linguistics.

Clayton Epp, Michael Lippold, and Regan L Mandryk. 2011. Identifying emotional states using keystroke dynamics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 715–724. ACM.

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

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

Mark Alan Finlayson and Nidhi Kulkarni. 2011. Detecting multi-word expressions improves word sense disambiguation. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 20–24. Association for Computational Linguistics.

Janet Dean Fodor. 2002. Prosodic disambiguation in silent reading. In *PROCEEDINGS-NELS*, volume 1, pages 113–132.

Donald R Gentner, Serge Larochelle, and Jonathan Grudin. 1988. Lexical, sublexical, and peripheral effects in skilled typewriting. *Cognitive Psychology*, 20(4):524–548.

Frieda Goldman-Eisler. 1958. Speech production and the predictability of words in context. *Quarterly Journal of Experimental Psychology*, 10(2):96–106.

John Hale. 2006. Uncertainty about the rest of the sentence. *Cognitive Science*, 30(4):643–672.

Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. 2012. An unsupervised ranking model for noun-noun compositionality. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 132–141. Association for Computational Linguistics.

Tina Hickey. 1993. Identifying formulas in first language acquisition. *Journal of Child Language*, 20(01):27–41.

Roger Johansson, Åsa Wengelin, Victoria Johansson, and Kenneth Holmqvist. 2010. Looking at the keyboard or the monitor: relationship with text production processes. *Reading and writing*, 23(7):835–851.

Ronald T Kellogg. 1996. A model of working memory in writing. In C. Michael Levy and Sarah Ransdell, editors, *The science of writing: Theories, methods, individual differences, and applications*, pages 57–71. Lawrence Erlbaum Associates, Inc.

D. Krathwohl. 2002. A revision of bloom's taxonomy: An overview. *Theory Into Practice*, 41(4):212–218.

Ronald W Langacker. 2008. *Cognitive grammar: A basic introduction*. Oxford University Press.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Rada Mihalcea. 1998. Semcor semantically tagged corpus. *Unpublished manuscript*.

Guido Nottbusch, Rüdiger Weingarten, and Said Sahel. 2007. From written word to written sentence production. *STUDIES IN WRITING*, 20:31.

A. Pawley. 1985. *Lexicalization*. the interdependence of theory, data, and application. Georgetown University Round Table on Languages and Linguistics, 98-120, Languages and Linguistics.

Heidi Riggenbach. 1991. Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse processes*, 14(4):423–441.

David E Rumelhart and Donald A Norman. 1982. Simulating a skilled typist: A study of skilled cognitive-motor performance. *Cognitive Science*, 6(1):1–36.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15. Springer.

T. A. Salthouse. 1986. Perceptual, cognitive, and motoric aspects of transcription typing. *Psychological bulletin*, 99(3):303.

Joost Schilperoord. 2002. On the cognitive status of pauses in discourse production. In *Contemporary tools and techniques for studying writing*, pages 61–87. Springer.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T Mordowanec, Henrietta Conrad, and Noah A Smith. 2014. Comprehensive annotation of multiword expressions in a social web corpus. *Proc. of LREC. Reykjavík, Iceland*.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics.

Maggy JW van Hoeij, JCM Haarhuls, Ronny FA Wierstra, and Peter van Beukelen. 2004. Developing a classification tool based on bloom's taxonomy to assess the cognitive level of short essay questions. *Journal of veterinary medical education*, 31:261–267.

Veronika Vincze, István Nagy, and Gábor Berend. 2011. Multiword expressions and named entities in the wiki50 corpus.

Alison Wray. 2005. *Formulaic language and the lexicon*. Cambridge University Press.

Yu Zhong, Yunbin Deng, and Anil K Jain. 2012. Keystroke dynamics for user authentication. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 117–123. IEEE.