6-2016

Utilizing Linguistic Context To Improve Individual and Cohort Identification in Typed Text

Adam Goodkind
Graduate Center, City University of New York

CUNY GRADUATE CENTER

MASTER'S THESIS

Utilizing Linguistic Context To Improve Individual and Cohort Identification in Typed Text

Author:

Adam GOODKIND

Advisor:

Dr. Andrew ROSENBERG

and

Dr. Martin CHODOROW

A thesis submitted in partial fulfillment of the requirements

for the degree of Master of Arts

in the

Concentration in Computational Linguistics

Department of Linguistics

2016

Declaration of Authorship

This manuscript has been read and accepted by the Graduate Faculty in Linguistics in satisfaction of the dissertation requirement for the degree of Master of Arts.

UTILIZING LINGUISTIC CONTEXT TO IMPROVE INDIVIDUAL AND COHORT IDENTIFICATION IN TYPED TEXT

ADAM GOODKIND

Professor Martin Chodorow

Date	Chair of Examining Committee

Professor Gita Martohardjono

Date	Executive Officer

Professor Andrew Rosenberg

Professor Martin Chodorow

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

Utilizing Linguistic Context To Improve Individual and Cohort Identification in Typed Text

BY

Adam GOODKIND

The process of producing written text is complex and constrained by pressures that range from physical to psychological. In a series of three sets of experiments, this thesis demonstrates the effects of linguistic context on the timing patterns of the production of keystrokes. We elucidate the effect of linguistic context at three different levels of granularity: The first set of experiments illustrate how the nontraditional syntax of a single linguistic construct, the multi-word expression, can create significant changes in keystroke production patterns. This set of experiments is followed by a set of experiments that test the hypothesis on the entire linguistic output of an individual. By taking into account linguistic context, we are able to create more informative feature-sets, and utilize these to improve the accuracy of keystroke dynamic-based user authentication. Finally, we extend our findings to entire populations, or demographic cohorts. We show that typing patterns can be used to predict a group's gender, native language and dominant hand. In addition, keystroke patterns can shed light on the cognitive complexity of a task that a typist is engaged in. The findings of these experiments have far-reaching implications for linguists, cognitive scientists, computer security researchers and social scientists.

Co-Advisors: Dr. Andrew Rosenberg and Dr. Martin Chodorow

Acknowledgements

This thesis would not have been possible without the insightful guidance of my advisor, Andrew Rosenberg. Much of the work was also done in close collaboration with David-Guy Brizan, who acted as a sounding board and wellspring of ideas. I would also like to thank Martin Chodorow for his insightful feedback and comments on the final draft of this these, as well as Janet Fodor for her constant guidance. The individual chapters of this thesis are based on papers submitted to NAACL-HLT 2015, BTAS 2015 and the International Journal of Human-Computer Studies. Thank you to the anonymous reviewers, as well as audiences at the conferences for their questions and suggestions.

Finally I would like to thank my family and friends for their unwavering encouragement throughout the entire Masters degree process. They have been my best critics and best supporters.

This work was supported in part by DARPA Active Authentication Phase I grant FA8750-12-2-0201 and FA8750-13-2-0274.

Large portions of this thesis have appeared or will appear in other publications. The table below indicates the chapter and the publication it is based upon.

Chapter	**Appears in Other Publication**
3	Goodkind, A., & Rosenberg, A. (2015, June). Muddying The Multiword Expression Waters: How Cognitive Demand Affects Multiword Expression Production. In *Proceedings of NAACL-HLT* (pp. 87-95).
4	Goodkind, A., & Brizan, D.G., & Rosenberg, A. (In revision). Utilizing Overt and Latent Linguistic Structure to Improve Keystroke-Based Authentication. *Image and Vision Computing: Best of Biometrics Special Issue*.
5	Brizan, D. G., Goodkind, A., Koch, P., Balagani, K., Phoha, V. V., & Rosenberg, A. (2015). Utilizing linguistically enhanced keystroke dynamics to predict typist cognition and demographics. *International Journal of Human-Computer Studies, 82*, 57-68.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

James Gleick's magnum opus *The Information* (Gleick, 2012) traces the history of information transmission from drummers in the African rain forest, to telegraph operators in World War II, all the way through modern cryptography and fiber optics. A single beat of a drum can convey a single bit of information. And in order for these bits to accumulate meaning, the drum beats must repeat, creating patterns from randomness.

Most telling, however, is that these streams of information – drum beats, telegraph signals, and computer keystrokes – have always been intertwined with their composers, never truly a sterile language wiped free of traces of personal identification. During World War II, counterintelligence experts learned to identify unique telegraph patterns, called the "fist of the sender," which could be used to identify individual signal operators. This knowledge, in turn, was used to track various troop movement.

The temporal patterns associated with keystrokes on a computer are similarly known to be unique to individuals. As such, the timing and choice of a sequence of keystrokes can be thought of a unique personal identifier, or biometric, similar to a fingerprint or DNA code. The present study combines the unique attributes of keystroke production with another unique identifier, language production, in order to improve the accuracy of biometric systems. Just as it is well-established that individuals have unique speech patterns and that these speech patterns are affected by linguistic context, this thesis advances the hypothesis that linguistic context affects

keystroke production, and that combining linguistic and temporal signals can improve the informativeness of keystroke pattern identification.

As an introductory example, traditional keystroke timing has measured the timing parameters of keystrokes or sequences of keystrokes *prima facie* A typical feature of a typing pattern might be the mean pause time between the keys `E` and `D`. However, the letter combination *e+d* has markedly different properties in the word *canned* versus in the word *red*. Within the former, it has distinct morphological properties and is pronounced differently. Thus, it seems reasonable to suspect that the processes underlying the production of the `E` and `D` keys in the word *canned* versus in the word *red* are also different. A feature such as "the mean pause time between e and d" might actually be too coarse-grained and in actuality be a combination of multiple distinct features.

This thesis explores the notion above from multiple angles and multiple levels of focus and granularity. Although the experiments conducted in this thesis provide novel insights, there exists a large body of works in keystroke dynamics and language production, which is highly relevant. These works are reviewed in 2. To introduce the notion of linguistic context affecting keystroke dynamics, this thesis first concentrates on a single linguistic construct, the multi-word expression. Chapter 3 illustrates how similar keystroke patterns can be produced with marked temporal differences depending on whether or not a word is within or outside of a multi-word expression. Chapter 4 then widens the scope of inquiry by demonstrating that adding linguistic context to an individual subject's keystroke feature-set can improve the informativity of this feature-set. Finally, Chapter 5 expands these findings to entire populations, by combining keystroke features with linguistic features to better identify demographic cohorts, as well as the cognitive complexity of a task being completed.

Chapter 2

Literature Review

The act of producing text is exceedingly complex. This complexity is highlighted by Alves, Castro, and Olive (2008, p. 2), when they declare that the writing process is "one of the most complex and demanding activities that humans engage in". The psychological investigation of typing goes back to at least the 1920s (Coover, 1923). These early studies, and subsequent studies in the later half of the 20th century, recognize that typing is a learned motor skill, and illustrates ingrained behavioral characteristics (Shaffer, 1978; Rumelhart and Norman, 1982; Salthouse, 1986, among others). As a result, typing patterns can be utilized as a biometric for both individual identification (Monrose and Rubin, 1997; Epp, Lippold, and Mandryk, 2011a; Banerjee and Woodard, 2012, among others) as well as cohort identification, or "soft biometrics" (Bartlow and Cukic, 2006; Villani et al., 2006, among others). Importantly, though, an individual's typing patterns can still fluctuate greatly from typing session to typing session (Bartmann, Bakdi, and Achatz, 2007).

The reasons for fluctuations in typing patterns can be both physical (motor) and cognitive. As noted by Schilperoord (2002), writers pause for a number of reasons, such as cognitive overload, writing apprehension or fatigue. Early investigators actually used typing to create holistic models of the interaction between language production and motor control, in general (Rumelhart and Norman, 1982). It was found, however, that skilled typists and untrained typists exhibit markedly different behaviors and employ different cognitive models (Gentner, Larochelle, and

Grudin, 1988). As noted by Alves et al. (2007)[p. 10], "Although motor execution is more demanding for slow typists, this higher demand [did not] prevent[] them from activating high-level processes concurrently with typing".

An attribute of an individual – whether physical or cognitive – is called a biometric. The practice of identifying an individual based on biometrics is called both *authentication* and *verification*. Within the present work, these terms are used interchangeablyy. The practice of using keystroke dynamics for authentication dates back to at least the early 1990s (Joyce and Gupta, 1990). Today, with typing becoming a ubiquitous process employed on a multitude of devices, keystroked-based authentication has expanded beyond QWERTY keyboards on desktop computers to smartphones and tablets (Saevanee, Clarke, and Furnell, 2012; Villani et al., 2006).

Keystroke-based biometrics can be based on short texts such as a password, single phrase, sentence or numeric PIN (Monrose, Reiter, and Wetzel, 2002), or a longer text such as a multi-paragraph essay (Curtin et al., 2006; Montalvao, Almeida, and Freire, 2006). For a recent survey of the use of keystroke dynamics for authentication, see Banerjee and Woodard (2012).

Longer texts can produce unique linguistic patterns, which are studied using stylometry. Unlike keystroke dynamics, which studies text production, stylometry investigates patterns in the final, static text. Stewart et al. (2011) investigated the respective benefits of keystroke dynamics versus stylometry for the purposes of authentication, and found advantages to both. Similarly, Darabseh and Namin (2014) used surface-level linguistic features to authenticate the authors of texts. Sim and Janakiraman (2007) found that in free text, word-specific digraphs and trigraphs are useful for authentication.

Other investigations of the relationship between typing and linguistics have looked at the overall quality and content of a text. Nottbusch, Weingarten, and Sahel (2007) found that pause duration is correlated with word frequency, word length and task type. Alves et al. (2007) found that more proficient typists produce longer bursts of keystrokes, or keystrokes between pauses. Alves et al. (2007) also found that proficient typists produce longer texts overall with more complex linguistic structure.

Because of the utility of keystroke dynamics as a soft biometric, typing patterns have also been used to classify other attributes of a typist. For example, Vizer, Zhou, and Sears (2009) used keystrokes as a means of stress detection. Typing patterns can also be used for more general emotion detection (Epp, Lippold, and Mandryk, 2011a) as well as deception detection (Choi, 2014). Similar to the demographic prediction studies in Chapter 5, keystroke dynamics has also been employed for gender identification (Fairhurst and Costa-Abreu, 2011; Giot and Rosenberger, 2012) and handedness identification (Monrose and Rubin, 1997; Idrus et al., 2014).

As noted above, longer texts also exhibit linguistic patterns which can be measured using stylometry. Stylometry can also be unique to an individual or a cohort. One of the earliest uses of stylometry (Mosteller and Wallace, 1964) successfully identified the authors of the Federalist Papers. Stylometry has also been used to predict gender differences in writing styles (Goswami and M. Rustagi, 2009; Vel, 2000; Koppel, Argamon, and Shimoni, 2002) as well as the author's native language (Bergsma, Post, and Yarowsky, 2012).

One specific linguistic construct investigated in this study is the multi-word expression (MWE). Previous work on MWEs has posited that they are retrieved as single lexical units (Wray, 2002), rather than word by word. Further, when MWEs are produced, the words making up the expression exhibit greater phonological consistency than free expressions (Hickey, 1993). Further, speakers find pauses within a multi-word expression to be less acceptable (Pawley, 1985). This is especially valuable because "...where pauses occur they give valuable indications of possible [MWE] boundaries" (Dahlmann and Adolphs, 2007, p. 55). In a related vein of research, Erman (2007) found that pauses can be caused by cognitive demands of lexical retrieval. These studies were done using speech production rather than typing.

The experimental advantages of keystroke dynamics are two-fold: data collection is relatively low-cost and produces high accuracy. Dahlmann and Adolphs (2007) point out that accurately determining pause times in speech data can be difficult, whereas no such difficulty exists in determining pauses in typing. Further, as pointed out by Cohen Priva (2010), typing experiments,

because of their low resource requirements, are also ideal for collecting data on less studied languages.

Although the experiments in the present study are novel experiments utilizing different types of data, they have been constructed based on a number of similar studies. For instance Killourhy and Maxion (2009a) found that a Scaled Manhattan classifier, similar to that used in Chapter 4, was most effective in similar tasks. The experimental data used in this work has also been utilized in two related papers, Locklear et al. (2014) and Balagani (2013). These studies both investigate user verification, where the subject's keystrokes are used to train a template, and then tested against a subject pool to uniquely identify future typing from the same user. In addition, while both studies take advantage of some of the language production features, they do not use the full set reported below. In this work, we are investigating the generalization of these keystroke-derived features from one group of typists to another, either with respect to the type of task that is being performed, or with respect to demographic cohorts.

Chapter 3

Keystroke Dynamics of Multi-word Expressions

3.1 Introduction

The set of experiments in first chapter investigates how a specific linguistic construct, the multi-word expression, is produced in typed text. 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 chapter presents a new modality for studying MWE production, keystroke dynamics. Specifically, this chapter 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 participant is asked, in order to elicit a response, has not been taken into consideration. This chapter takes advantage of the low cost and high precision of keystroke dynamics to uncover trends in MWE production, by eliciting responses from participants using a variety of questions with very different cognitive demands. The 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 this analysis, we collected free response typing data from a large set of participants. The participants 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 production. In this chapter's main experiment, we then use these findings to analyze how multi-word expression production is affected by the cognitive demands imposed upon the participants.

This chapter advances the hypothesis that the cognitive demands of a task will impede MWE production, as the overall demands will interfere with lexical retrieval, creating a cognitive bottleneck. The 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?

3.2 Materials and Methods

The materials and methods utilized in the MWE experiments were slightly different than those utilized in experiments relating to verification and demographic identification, as outlined below.

3.2.1 Data Collection

The fundamental units of measurement in keystroke dynamics are intervals/pauses and holds (Monrose and Rubin, 2000). A pause or interval is the time elapsed between a key being released and the next key being pressed, as represented by *pause* in Figure 3.1. This is sometimes also called *latency*, *flight time* or *digraph* in related studies. In this work, we define a "pause" as the pause immediately preceding a keystroke. An "interval," on the other hand is a pause that is contextualized between both the preceding and subsequent keystroke.

The second useful metric is the elapsed time that a key is depressed, or the *key hold* span in Figure 3.1. As an example, the "digraph key hold" in Figure 3.1 is the elapsed time between the first and second arrows plus the elapsed time between the third and fourth arrows. Any pause time between key holds (labeled *pause* in Figure 3.1) is not included.

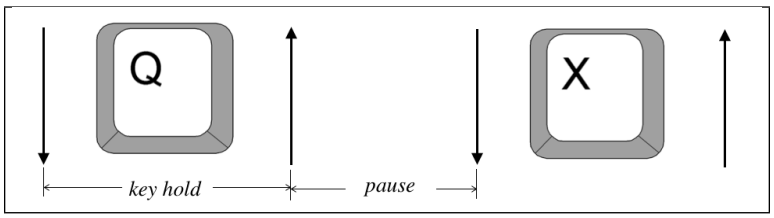


FIGURE 3.1: Timing parameters of a keystroke

The typing data was collected from 189 Louisiana Tech students (hereinafter referred to as "participants"). The participants reported themselves to be 41.3% female, 56.4% male and 88.3% right-handed and 9.1% left-handed. (As with Chapters 4 and 5, these do not sum to 100%; on each question some percentage of participants chose not to respond to one or more of the demographic questions.)

The 189 participants represent only a subset of the total participant population, which is utilized in subsequent experiments. The typing data for these MWE experiments was limited to only native English speakers and touch typists.

Only native English speakers were selected in order 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 linguistic fluency. Further, only "touch typists", or those participants who only look at the screen when typing were selected. This is in comparison to "visual typists" who look at their fingers when typing. As proposed by Johansson 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.

Similar to the experiments in Chapters 4 and 5, the participants were seated at a Dell desktop with a QWERTY keyboard and presented with a series of prompts in Standard American English selected from those listed in Appendix A. The participant was required to type at least 300 characters in response to each prompt, at which time the participant was presented with a button allowing him or her to proceed to the next prompt. Each participant responded to 10 - 12 prompts. The prompts in session 1 were completely distinct from those in session 2, though they were drawn from the same categorization of cognitive tasks. Prompts were presented in random orders, though there was an equal distribution of each task type. (However, in the second session, one of the level 3 questions was omitted and replaced by a level 5 due to an error in the preparation.) The participant was required to type 300 characters. The average response contains 921 keystrokes; the final response contained an average of 448 characters and 87 words. A keylogger with 15.625 milliseconds clock resolution was used to record text and keystroke event timestamps (Locklear et al., 2014). This collection protocol was reviewed and approved by the Louisiana Tech University IRB.

This data was collected with the intention of evaluating user verification. In order to measure the impact of the type of behavior a user is engaged in on the consistency of his or her typing, the participants were asked to perform a wide range of different tasks. In this work, we are investigating whether we can recognize what type of behavior an unseen typist is engaged in, by observing other typists performing similar tasks (though responding to different prompts).

In each session, participants were presented with a set of 12 prompts to respond to (from a set of 36). We randomized the order of the prompts before presenting them to the participants.

Each prompt was drawn from one of six tasks: REMEMBER, UNDERSTAND, APPLY, ANALYZE, EVALUATE, or CREATE. This task type was determined by the experimenters, who assessed the cognitive demands of each question as they related to Bloom's Taxonomy (Anderson, Krathwohl, and Bloom, 2001). Bloom's Taxonomy assigns a level from 1-6 to these cognitive tasks. We use this as an ordering of tasks from low to high cognitive demand. Table 3.1 contains a list of the types of tasks, the cognitive activity required to respond to each prompt and a sample question associated with each task type. The list in Appendix A contains all of the questions posed to the participants and the task associated with each question.

TABLE 3.1: Cognitive Load Definitions and Example Prompts

Task and Level	Required Activity	Example Prompts
REMEMBER - 1	Retrieve knowledge from long-term memory or explain	List the recent movies you've seen or books you've read. When did you see or read them? What were they about?
UNDERSTAND - 2	Explain, Summarize or Interpret	Where is a place that you particularly enjoy visiting? Describe what makes you happy about being at this place.
APPLY - 3	Apply, execute or implement	What would you do if you and a friend are on vacation alone and your friend's leg gets cut? Describe what procedure you would use for first aid or for finding help.
ANALYZE - 4	Organize or break material into constituent parts	Explain what you think the difference is between "communicating with" someone and "talking to" someone. How are these two terms often confused?
EVALUATE - 5	Critique or make judgments based on criteria	Do you think it's a good idea to raise tuition for students in order to have money to make improvements to the University? Why or why not?
CREATE - 6	Generate, plan or put elements together	Pretend a Hollywood executive offered to pay you to write and act in a movie. Create a movie plot with a character in it for yourself and remember that you will only be paid for creating an original plot to a movie.

We note that the cognitive level measure of a prompt is most accurately interpreted as the *expected* cognitive demands as hypothesized by Bloom's Taxonomy. It is, of course, possible or even likely that some participants may experience different cognitive loads than would be expected by a given prompt. For example, a participant may choose to create new knowledge (e.g., making up his or her favorite movie), rather than retrieve a memory in responding to a REMEMBER prompt. Additionally, a participant may have experiences that are relevant to a CREATE prompt (e.g., having written a film script); this may lead to a response which is based more on recall than creative thought. In addition to these confounds, there are environmental effects that may lead to a user experiencing a higher than expected cognitive load. For example, a cell phone might ring, or they may be distracted by other thoughts. Due to these effects, we note that there may be a significant amount of noise between the *true* cognitive demand which a participant is experiencing while responding to a prompt and the *expected* cognitive demands dictated by the prompt itself. Measuring the discrepancy between these would provide valuable information for this research, but is outside the scope of this work, and the protocol under which the data was collected.

We also note that while Bloom's Taxonomy provides a valuable way to delineate a set of tasks (e.g. *rote knowledge* vs. *knowledge creation*), the taxonomic, hierarchical nature of Bloom's system has been called into question (Paul and Binker, 1990) and heavily revised for pedagogical purposes (Anderson, Krathwohl, and Bloom, 2001). Further, it is often difficult to achieve consensus as to which label to assign, even among a group of subject-matter experts (Hoeij et al., 2004). Our findings reflect the fact that Bloom's tasks are not hierarchical or continuous in nature, but rather, reflect different, discrete tasks.

Within each valid response, we do not perform any outlier removal. However, the data was vetted to ensure that all responses are, in fact, responses to the prompt. This editing was restricted to those instances where a participant was unquestionably unresponsive – this included cases where the response included a seemingly random sequence of characters, or a word or phrase repeated multiple times, until the minimum character count was reached. However, as distinguishing between an outlier and an idiosyncrasy can be difficult and subjective, this editing was

quite conservative. Only in cases where a research assistant was completely certain that the participant was non-responsive was the response removed from the data set.

3.2.2 Features

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 Brown Concordance (Mihalcea, 1998; 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 phenomena. As a result of this exclusion, MWEs such as *ran up* in (3.1) would be included in our study, while the same non-contiguous MWE in (3.2) would not.

(3.1) *Jack ran up the bill.*

(3.2) *Jill ran the bill up.*

While keystroke dynamics is concerned with a number of timing metrics, such as key holds (*h* in Figure 3.2) and pauses between every keystroke (*p* in Figure 3.2), the current study looked only at the pause preceding a word (the second *p* in Figure 3.2). This interval consists of the time between the spacebar being released and the first key of the word being pressed. The decision to include each spaces as part of the prior word, rather than the upcoming word, was related to the findings in the next chapter (Section 4.3.3) and illustrated in Figure 4.1. Because pauses were shorter before a space, it is assumed that the word-final space is produced as "part of" the word immediately preceding it.

We also did not remove any outliers, although this is common in keystroke dynamics (Epp, Lippold, and Mandryk, 2011b; Zhong, Deng, and Jain, 2012). We feel it is difficult-to-impossible

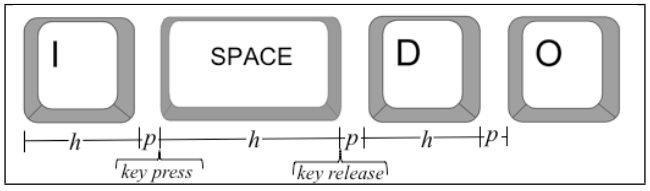


FIGURE 3.2: Timing intervals in keystroke dynamics specific to both individual keystrokes as well as entire words

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.

3.3 Experiments

3.3.1 Experiment 1: Creating A Baseline

In Experiment 2, 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 3.3, 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 pausing behavior.

As noted by Nottbusch, Weingarten, and Sahel (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 log of the pause length increased as a function of the log of the word length (see Figure 3.4). Since we expect cognitive demand to affect typing,

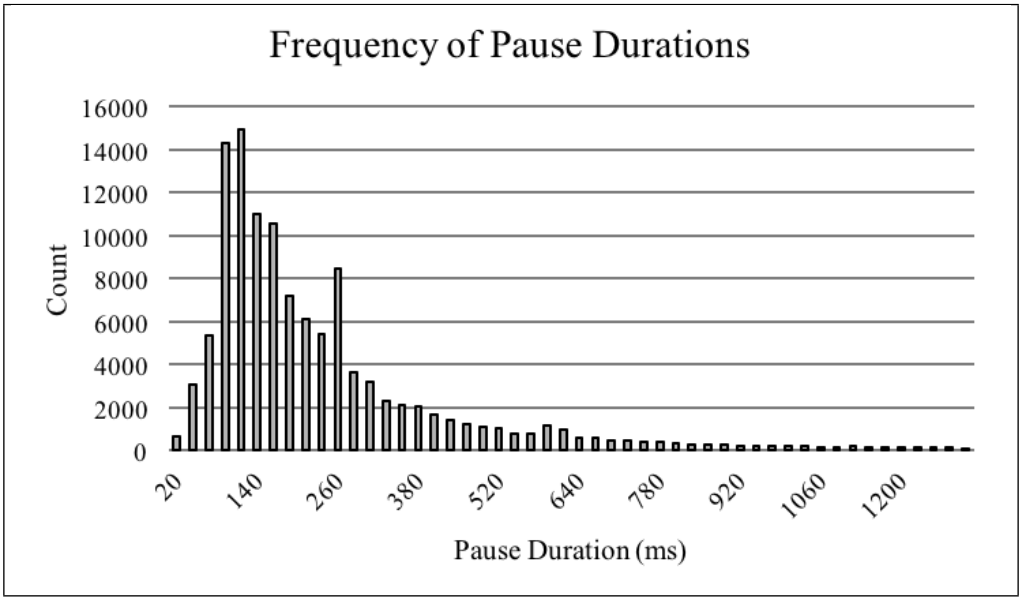


FIGURE 3.3: Distribution of all pauses

we measured this affect on each task, and created different α and β parameters for our "Expected Pause" algorithm, as described in Equation 3.3.

$$\ln[Pause_{expected}(w)] = \alpha \cdot \ln(length(w)) + \beta \tag{3.3}$$

The regression model illustrated in Equation 3.3 provided a very reliable fit for all tasks. Between each of the tasks α ranged from $0.107 - 0.112$ while β ranged from $2.20 - 2.24$. In the various implementations of Equation 3.3 R^2 ranged from $0.93 - 0.98$, yet the differences were never significant, with $0.22 < p < 0.58$.

In Experiment 2, all pauses within an MWE were quantified as a deviation from the overall expected pause (both within and outside of an MWE), where both word length and cognitive demand were taken into account in a regression analysis.

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

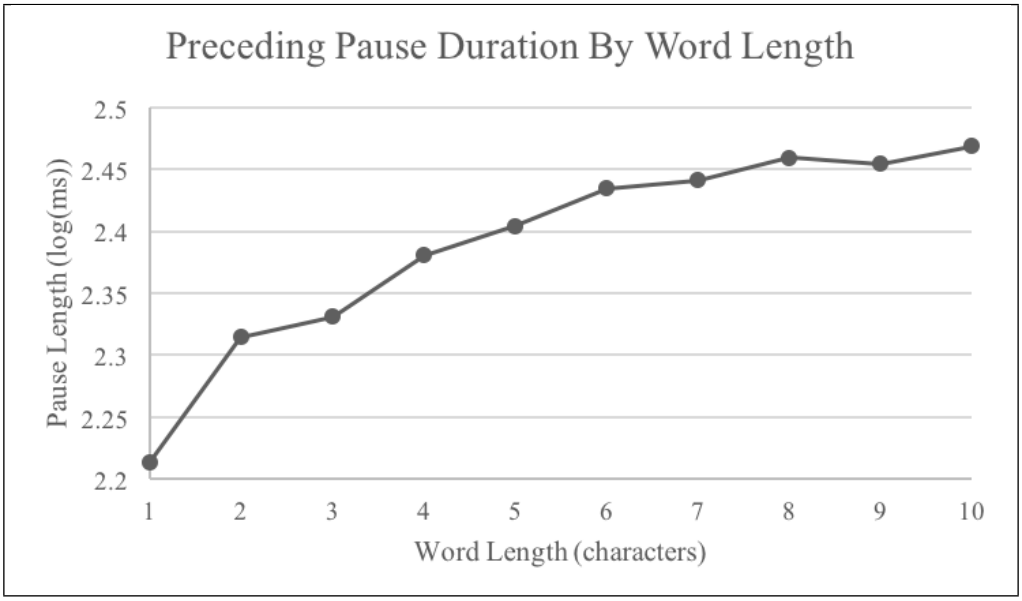


FIGURE 3.4: Duration of pre-word pause by word length

(Goldman-Eisler, 1958; Hale, 2006; Nottbusch, Weingarten, and Sahel, 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 lexical sequences.

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 Equation 3.4, 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.

$$P(w_n|w_{n-1}) = \frac{C(w_{n\text{-}1}w_n) + 1}{C(w_{n\text{-}1}) + V} \tag{3.4}$$

The grouping was done by rounding the log probability of the bigram sequence. The details

of the rounded groupings are provided in Table 3.2. 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 3.5. 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.

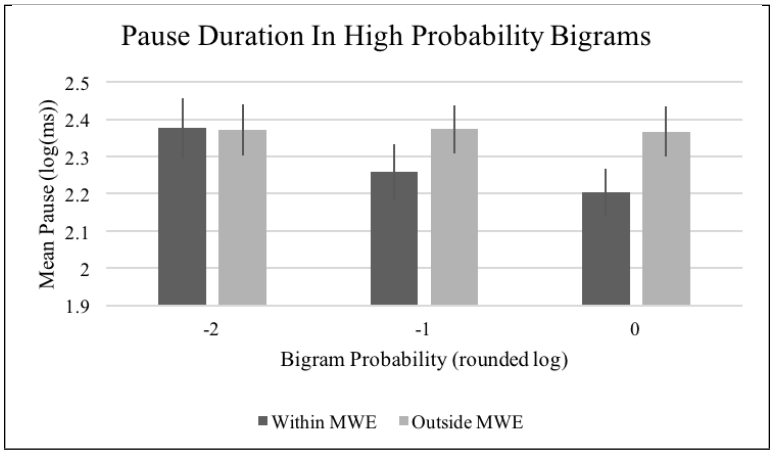


FIGURE 3.5: MWE production in high predictability sequences

	Within MWE					Outside MWE				
Rounded Log Probability	***n***	**Mean (Linear) Probability**	**Maximum (Linear) Probability**	**Minimum (Linear) Probability**	**SD**	***n***	**Mean (Linear) Probability**	**Maximum (Linear) Probability**	**Minimum (Linear) Probability**	**SD**
-2	5,801	0.012	0.032	0.003	0.008	32,548	0.013	0.032	0.003	0.008
-1	8,723	0.124	0.315	0.032	0.074	44,583	0.123	0.315	0.032	0.070
0	2,344	0.494	0.872	0.322	0.136	3,204	0.498	0.962	0.317	0.158

TABLE 3.2: Detailed parameters of rounded log probability bigram groups for MWE production in high predictability sequences

3.3.2 Experiment 2: MWEs in Varying Cognitive Tasks

MWEs were produced at a fairly consistent rate across all tasks, comprising approximately $12-13\%$ of all word tokens, as reported in Table 3.3. It should be noted that this figure is markedly

lower than often cited figures such as Erman and Warren (2000), who point out that half of spoken and written language comes from multi-word constructions. In the present case, however, we are dealing with a small subset of MWEs, namely those that were produced contiguously (cf. examples (3.1) and (3.2) 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 3.3: MWE production rates and counts by cognitive task label

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 3.6. 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.

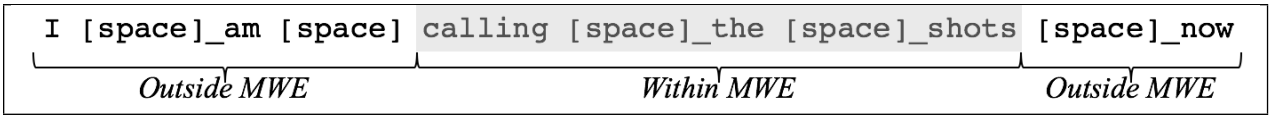


FIGURE 3.6: An example sentence. Measured pauses are represented with an underscore.

In each task, words within MWEs were consistently produced with a shorter preceding pause than were words in free expressions. As seen in Figure 3.7, pauses are shorter within MWEs across all tasks.

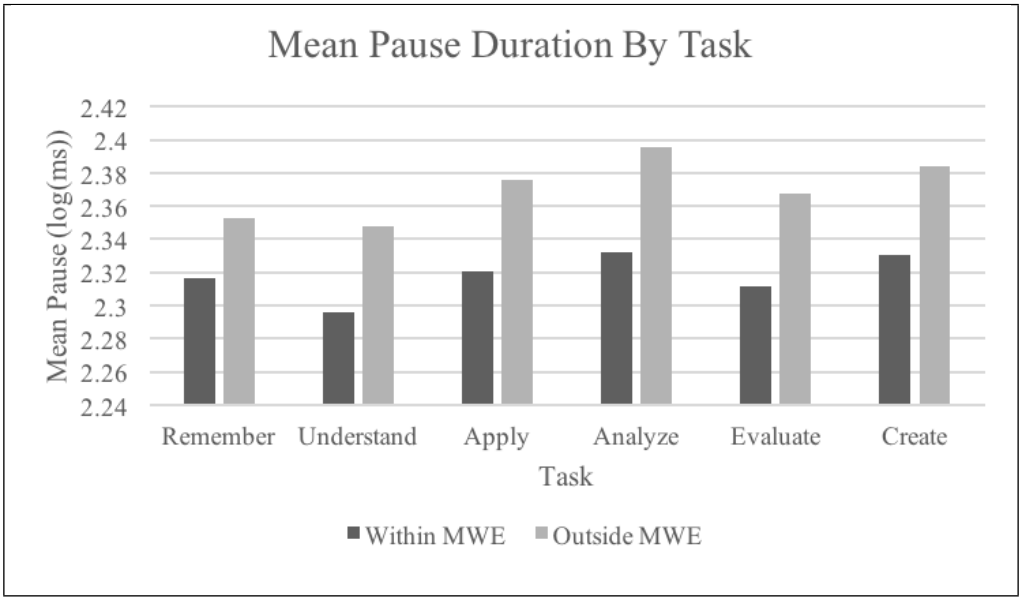


FIGURE 3.7: Pause duration by task, within and outside MWEs

However, the distributions of the means as reported in Figure 3.8 is not uniform. Within-MWE pauses are not only shorter in duration, but in addition evidence exists 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 ($IQR_{within-mwe} = 160$, $IQR_{outside-mwe} = 240$). Specifically, the Within-MWE distribution was more skewed.

Our investigation, though, aimed to look at how pausing *within MWEs* varies between cognitive loads, rather than an overall distribution. These results are illustrated in Figure 3.9. 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 complexity at the $p < 0.001$ level, $[F(5, 11796) = 4.19, p = 0.000815]$.

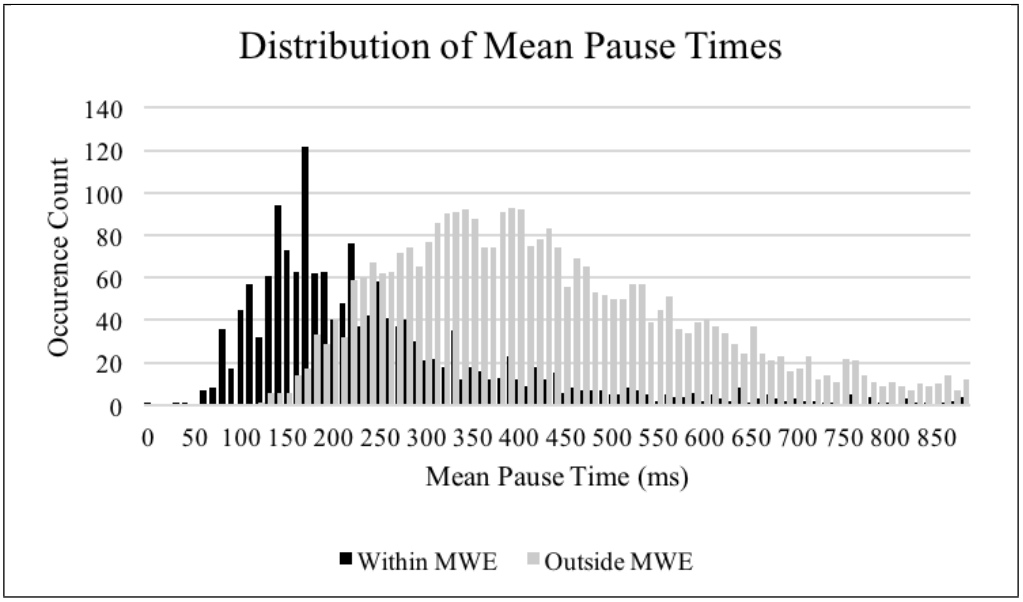


FIGURE 3.8: Distribution of mean pauses within and outside MWEs

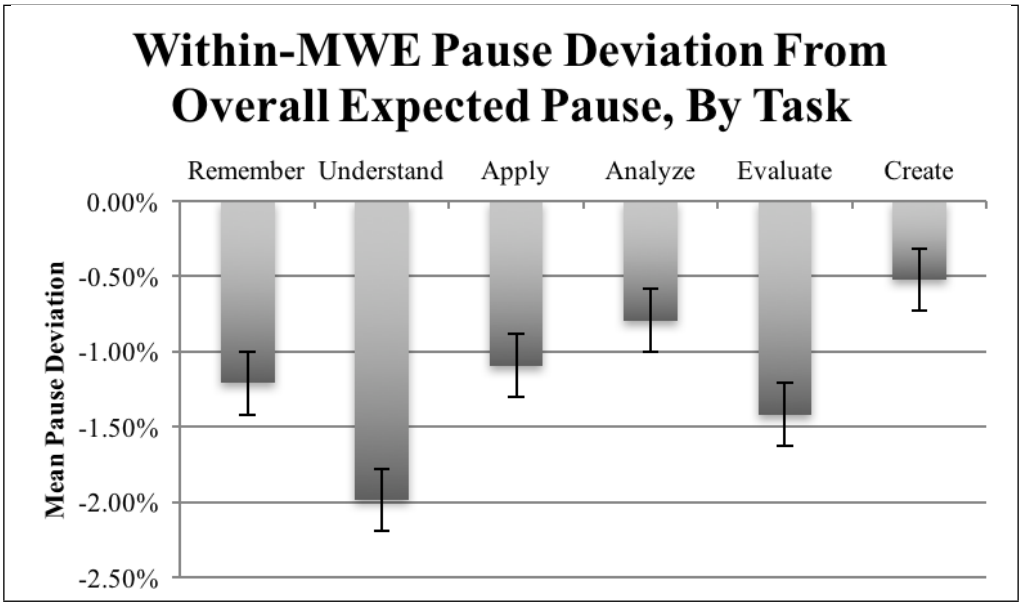


FIGURE 3.9: Within-MWE pause duration deviation by cognitive task, as calculated by the log of the pause time. In this figure, tasks are arranged from (generally) simplest to most complex.

3.4 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 (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 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 3.10. To clarify this idea, though, subsequent investigations will investigate pauses at the boundaries of MWEs.

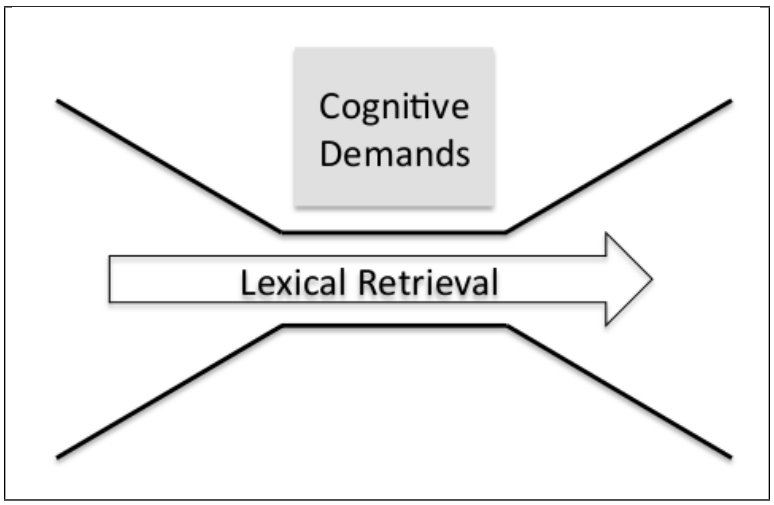


FIGURE 3.10: Model of Cognitive Bottleneck

The notion of various schemata interacting is supported by Kellogg (Kellogg, 1996), who proposes that "resources from the central executive of Baddeley's model of the working-memory, e.g., Baddeley (Baddeley and Hitch, 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, 2010), where participants 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.

Chapter 4

Linguistics and Keystroke-based Individual Authentication

4.1 Introduction

While Chapter 3 focused on a single linguistic construct, Chapter 4 moves one step further by demonstrating that a large number of linguistic constructs are relevant in understanding and identifying typing patterns. These linguistic context can be syntactic, semantic and lexical in nature. We test this notion by applying linguistic context to experiments that attempt to identify unique typists from our dataset.

Deducing elements of user behavior that are unique to the user, or identifiable, remains a persistent problem in computer security (Zheng, Paloski, and Wang, 2011). More and more, researchers are relying on biometrics such as iris scans and fingerprints over passwords as text-based passwords are becoming less reliable and less secure (Ashbourn, 2014).

The experiments in Chapter 4 propose the use of both overt and latent linguistic context to improve the uniqueness of biometric markers extracted from a user's keyboard typing. Keystroke-based authentication has typically been performed using so called "atomic" features: key holds and key intervals (Teh, Teoh, and Yue, 2013). These experiments advance the hypothesis that linguistic context impacts these atomic features. Thus, by considering an n-graph of holds and

intervals within the linguistic context in which they were produced, we will increase the accuracy of user verification.

In a set of four experiments, we answer a series of questions of interest to computer security researchers, linguists and cognitive scientists:

- Is a singleton keystroke most informative, or should keystroke events be considered in context, e.g. in groups of two or more successive keystrokes?
- Does linguistic context provide meaningful insight into a user's typing patterns? Are *n*-graphs best considered distinctly based on word context, part-of-speech, etc.?
- Is a keystroke's location relative to word boundaries important to take into consideration?
- How does feature pruning, i.e. eliminating rare features, affect verification results?

4.2 Materials and Methods

In this section we describe the keystroke data collection (Section 4.2.1) used in our user verification experiments, the specific features used for verification (Section 4.2.2), the verification algorithm (Section 4.2.3) and specifics of the verification experiment (Section 4.2.4).

4.2.1 Data Collection

The typing data was collected from 486 Louisiana Tech University students (hereinafter referred to as "participants") in two sessions, 6 months apart. Participants received unique IDs identifying them across sessions, and we collected a number of self-reported demographics for each subject. The participants were 41% female to 59% male. 82% were L1 English speakers, while 17% were non-native English speakers. Finally, 88% of participants reported being right-hand dominant, while 9% reported being left-handed.[1] The mean typing rate of the users was 168.9 intraword keystrokes per minute (Alves et al., 2007) with a standard deviation of 50.04.

[1]Note that these do not sum to 100%; on each question some participants did not report gender, native language or handedness.

All of the same experimental parameters outlined in 3.2.1 were also utilized for the experiments in Chapter 4. The only difference is that the present chapter takes advantage of nearly the full set of users.

4.2.2 Verification Features

To test our hypothesis that keystroke timing is not independent from its linguistic context, we implemented the following features into our verification experiments.

- n-graph - Pause and Hold features comprise sequences of n keystrokes. Examples of unigraphs are [*T*] or [*SHIFT-KEY*], while a digraph feature might be [*T_H*] or [*SPACEBAR_X*]. An n-graph can be a single event or an accumulation of multiple events, such as a *digraph hold* (described in Table 4.1, below). If the timing of the individual keystrokes are independent, no discriminative power will be gained by their joint representation. On the other hand, if these features are more informative than unigraph keyhold and preceding pause times, this indicates that the context in which keystrokes occur is important for verification.

Feature (Abbreviation)	Description	Example
Unigraph Hold (UH)	The length of time a single key is held down	HOLD_E
Unigraph Hold in Digraph (UH_D)	The length of a single keystroke hold, contextualized by a digraph of that keystroke and the preceding keystroke	HOLD_E_RE
Unigraph Pause (UP)	The pause length before a single keystroke	PAUSE_U
Digraph Hold (DH)	The cumulative holds times of two successive keystrokes	HOLD_Q_U
Digraph Interval (DI)	The pause length between two keystrokes	INTERVAL_B_U

TABLE 4.1: Descriptions and examples of types of n-graphs

- *n*-graph in Word The text of each session was tokenized using CoreNLP (Manning et al., 2014). Word tokens also maintained any capitalization. This tool allows us to extract distinct features for keystrokes based on the words in which they occur. Features were of the form [*UP_TH*|*THEY*] describing the keystrokes corresponding to the 'T' and 'H' in the word "THEY".
- *n*-graph in Lemma Rather than consider an *n*-graph within the bare word, we replaced each word with its lemma. For example, the words *works*, *worked*, and *working* would all be grouped into the lemma "*work*." The practice of lemmatization is also sometimes referred to as stemming, although stemming usually involves simply removing prefixes or suffixes, whereas lemmatization usually involves finding the root form of a word. This allows us to generalize features across various morphological variants of the same lemma. In addition, a lemma does not maintain capitalization since it represents an abstract or root form of the word. If a user typed "worked" one of the lemmatized features would be [*UH_ED*|*LMA_WORK*]. Even though "ed" does not appear in the lemma *work*, it does appear in the original text.
- *n*-graph in Part-of-Speech Each tokenized word was Part of Speech (POS) tagged with CoreNLP (Manning et al., 2014). Similar to extracting keystroke features based on words, we tested *n*-graphs within the part-of-speech (POS) of the word. These features had the form [*METRIC*|*POS*], where POS is some tag such as *NN* (singular noun) or *VBD* (past tense verb). This allows us to ask whether different participants type an n-graph within, e.g., nouns with consistently different timing than within, e.g., verbs. An example feature, given a user typing "car," would be [*DI_CA*|*NN*]
- *n*-graph in Lexical Category POS tags were first split into function or content tags. Function tags were then further subdivided by whether they were important "Pennebaker class" words, as defined by Chung and Pennebaker (Chung and Pennebaker, 2007). This allowed us to answer whether individual POS tags were too fine-grained. Pennebaker-class words have been found by Chung and Pennebaker (2007) to be "psychologically informative" to

a number of behaviors. Throughout the paper, we refer to this as "FCP", short for "Function/Content/Pennebaker". As an example, if a user typed "they," one feature would be [*UH_T*|*FUNCTION*].

4.2.3 Scaled Manhattan Verifier

We utilized a Scaled Manhattan (SM) verifier to verify each subject, as these have proven successful in similar tasks (Araujo et al., 2005; Killourhy and Maxion, 2009b; Killourhy and Maxion, 2010). The Scaled Manhattan verifier is a very simple but effective verification algorithm. For each user i, a template consisting of the mean $\vec{\mu_i}$ and standard deviation $\vec{\sigma_i}$of each feature $j \in J$ observed during training is constructed. During testing, a test vector $\vec{v}$ containing j features is constructed. The user whose template has the closest scaled Manhattan distance to the test vector is selected by the classifier. Scaled Manhattan distance between a test vector and template is defined as

$$d(\vec{v}, i) = \sum_{j \in J} \frac{|\vec{v}[j] - \mu_i[j]|}{\sigma_i[j]}$$

.

4.2.4 Verification Experiment Setup

In all experiments training data was taken from the first data collection session, while our testing data was drawn from the second session. Users returned six months later, and answered a new set of comparable but unique questions.

For each user, a template was created from the subject's entire first session (i.e. 12 answers). These templates were tested against separate feature vectors created from an entire individual answer regardless of the elapsed time to produce the answer or how many keystrokes were produced to compose the answer. As no modification of the test participant impostor answers was performed, this is best considered a "zero-effort" attack (Jain, Ross, and Nandakumar, 2011). Our

performance was measured by Equal Error Rate (EER). The EER is the point at which false acceptance and false rejects are equal. Similar to F_1 scores, the EER can be viewed as a tradeoff between precision and accuracy, although within biometric performance the tradeoff is between overly-stringent and overly-tolerant systems.

4.3 Results

Our experiments below measure a range of modifications to a baseline feature-set. Experiment 1 investigates the optimal atomic unit or units for authentication, Experiment 2 then adds both overt and latent forms of linguistic context to our atomic features. Both Experiments 3 and 4 attempt different methods of pruning our data: Experiment 3 attempts to prune keystrokes based on their locations relative to word boundaries, while Experiment 4 refines feature-sets based on properties of the observations comprising each feature.

4.3.1 Experiment 1: Optimal Atomic *n*-graph Feature-Set

Before testing the utility of linguistic context, Experiment 1 aimed to determine which atomic unit or combination of atomic units yielded the most accurate authentication results. We tested using key holds, key intervals (pauses contextualized between two keystrokes) and preceding pauses (pauses contextualized only by the subsequent keystroke).

The atomic features utilized for testing, in isolation and in combination with each other are described above in Table 4.1. Table 4.2 illustrates our results.

As can be seen from Table 4.2, combining multiple types of *n*-graph features substantially improves results. In addition, contextualized features such as a unigraph hold within the digraph context (UH_D) and the digraph interval (DI) perform generally more successfully than features without context. Overall, our verification experiments produced the best results by fusing three

Feature Set	EER	Feature Count
Unigraph Hold (UH)	**0.0949**	58
Unigraph Pause (UP)	0.1663	58
Unigraph Hold in Digraph (UH_D)	**0.0426**	290
Digraph Hold (DH)	0.0638	145
Digraph Interval (DI)	0.0890	145
UH, UH_D	**0.0387**	348
UP, DI	0.0774	203
UH, DI	0.0483	203
DH, UH, DI	0.0396	324
UP, UH_D, DI	0.0406	324
UH, UH_D, DI	**0.0368**	261

TABLE 4.2: Results of Experiment 1: Investigating atomic feature sets described in Section 4.2.2

feature-sets: unigraph holds (no surrounding context), unigraph holds contextualized by the preceding keystroke, and digraph intervals (which take into account both the preceding and subsequent keystroke).

4.3.2 Experiment 2: Adding Linguistic Context

The goal of Experiment 2 was to test whether more explicit *linguistic* context would improve our results. This is based on a hypothesis that typing patterns are influenced by the lexical and syntactic structure of the linguistic content being produced, even if this content is not visible from a surface-level reading.

We hypothesize that the same letter combination, when produced within different linguistic constraints, will be produced with a consistent difference. For example, a user may consistently type "ED" more quickly within a past-tense verb than she will when "ED" appears in a noun. Therefore, a feature such as *[DI_ED]* is too coarse-grained, and should be further broken up by the linguistic structures in which "ED" appears, e.g. *[DI_ED|RED]*.

Based on our previous findings (Section 4.3.1) all linguistic context experiments are performed by adding context to the feature-set made up of unigraph holds, unigraph holds in digraph context

and digraph intervals.

We tested combinations of three forms of linguistic context described in Section 4.2.2, word (WRD), lemma (LMA), part-of-speech (POS), and Function/Content/Pennebaker words (FCP) contexts. In each experiment, we augment the base feature-set with features that incorporate linguistic context, rather than replacing the base features.

As can be seen in Table 4.3, adding the lemmatized features provided the most accurate linguistic information.

Feature Set	EER	Feature Count
Word (WRD)	0.0329	1810
Lemma (LMA)	**0.0309**	1762
Part-of-speech (POS)	0.0426	1036
Function/content (FCP)	0.0368	716
WRD, FCP	0.0359	1714
LMA, POS	0.0464	1759
LMA, FCP	0.0329	2120

TABLE 4.3: Results of Experiment 2: Adding linguistic context to experimental feature-sets

4.3.3 Experiment 3: Omitting Word-Liminal Intervals

Prior research has suggested that the most accurate assessment of a typist's proficiency should come from the rate at which only intra-word keystrokes are produced (Alves et al., 2007). This implies that pauses between words are caused by factors outside of "typing proficiency," such as sentence planning or lexical retrieval.

These findings are partially corroborated by an investigation of the variance surrounding intra- and inter-word pauses as well as pause lengths before and after sentence-final and clause-final punctuation within our training data set.

The box plots in Figure 4.1 not only provide a picture of differing means, but also provide a glimpse of differences in variance, as well. The first chart compares the distribution of pauses that occur intra-word, i.e. between two alphanumeric keystrokes with the distribution of pauses that

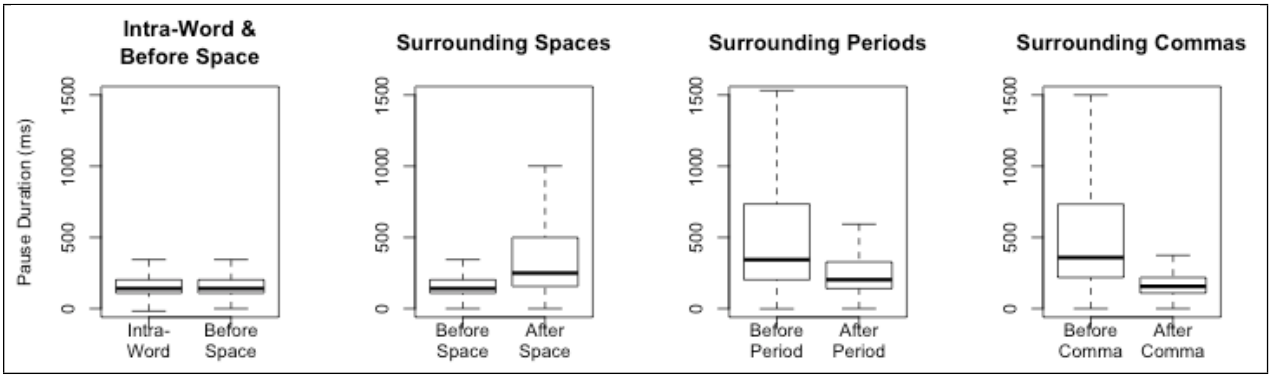


FIGURE 4.1: Distributions of pauses in intra-word intervals, intervals surrounding spaces and intervals surrounding punctuation. Pauses in intra-word intervals are nearly identical to pauses in intervals preceding spaces. Distributions in pauses surrounding spaces and punctuation display significant differences depending on whether they precede or trail the given keystroke.

occur after a word-final keystroke and before a space. The second chart compares the distributions of pauses preceding and following a spacebar. The third chart compares pauses surrounding a period. Finally, the fourth chart compares pauses surrounding commas.

As illustrated in Figure 4.1, pauses that occur within words are nearly identical to pauses that occur between the last keystroke in a word and the trailing space. This implies that the spacebar is struck almost as "part of" the word it follows, with little hesitation between concluding the typing of a word and entering a space. On the other hand, *after* a space is produced, the distribution of pauses displays more variance and a greater median pause time, or more uncertainty.

Conversely, more hesitation and variance exists *before* producing a period or a comma. Once a clause or sentence is concluded, though, with its appropriate punctuation, a typist moves much more rapidly and reliably on to the subsequent clause or sentence, as illustrated by the relatively more narrow distributions following periods and commas.

Given that Alves et al. (2007) finds intra-word typing rate to be the most consistent metric of typist proficiency and that word-liminal intervals provide a great deal of variance, we ran a series of experiments investigating whether removing these possible sources of variance could improve

authentication results.

Excluded Intervals	EER	Feature Count
No exclusions	**0.0368**	426
Pre-word	0.0406	335
Pre-punctuation	0.0386	353
Pre-word and -punctuation	0.0406	330

TABLE 4.4: Results of Experiment 3: Removing Word-Liminal Keystrokes

As shown in Table 4.4, the carte blanche removal of word-liminal intervals did not improve results. It is possible that removing these intervals simply removed too much overall information.

4.3.4 Experiment 4: Feature Pruning

We next investigated whether eliminating rare features would have an impact on classifier performance. Keeping the number of features small leads to a fast and robust verifier, but pruning too aggressively will result in less informative features, which could negatively impact performance.

Figure 4.2 illustrates the mean number of observations within each feature in a baseline feature-set. The baseline feature-set was made up only of unigraph holds (UH), unigraph holds in digraph context (UH_D) and digraph intervals (DI). Most features were made up of only a handful of observations: The mean size of a feature was 11.6 observations while the median size was 4 observations. The interquartile range (IQR) of the feature-set was between 2.6 and 6.0 observations. As seen in Figure 4.2 the decrease in counts of features with additional observations follows an exponential distribution.

We investigate four pruning methodologies. The methodologies are described below. The pruning threshold is represented by θ.

- Minimum-maximum Pruning - A feature is included if the size of only a single subject's set of observations is greater than θ. Even if the feature was absent from every other subject's feature-set, a single subject's production would be sufficient for inclusion of the feature in the overall feature-set.

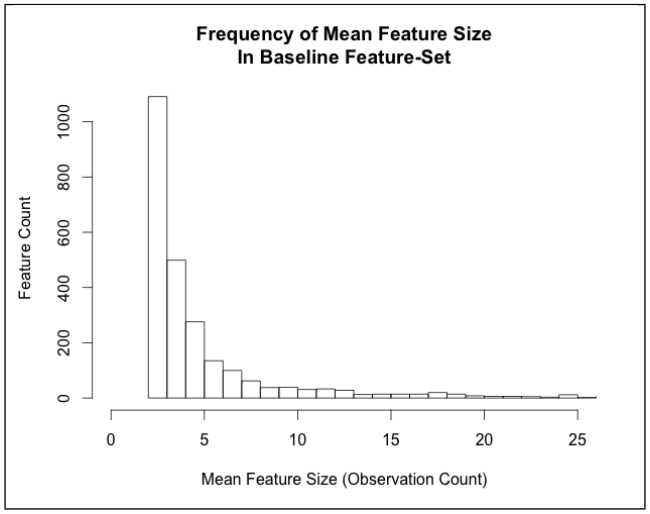


FIGURE 4.2: Feature-set sparsity, as illustrated by exponentially decreasing counts of observations within a feature

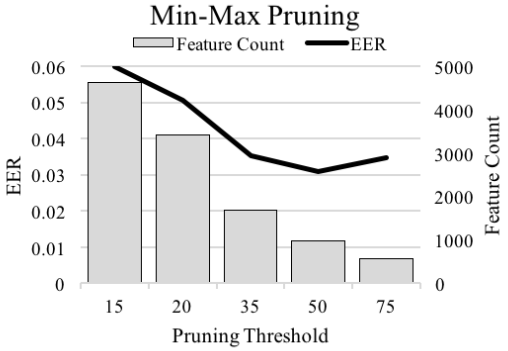
- Maximum-minimum Pruning - This type of feature pruning represents the converse of minimum-maximum pruning. In maximum-minimum pruning, every participant must produce a set of observations greater in size than θ. If even one participant fails to produce a set of observations of sufficient size, the feature will not be included for any participants in the feature-set.
- Top Count Pruning - For each feature, the number of observations was totaled across all participants. The features were then ordered from greatest total number to least, and the top θ features were selected.
- Z-score Pruning - For each feature, we calculate a z-score by the following formula,

$$z = \frac{x - \mu}{\sigma}$$

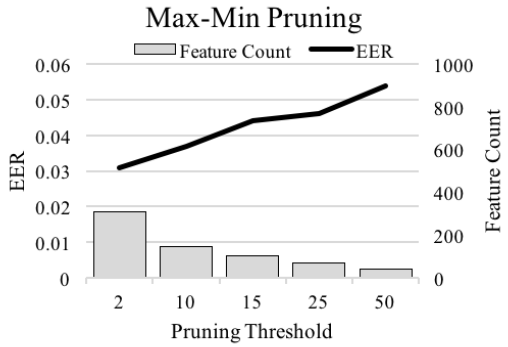
where x is the feature value, and μ and σ are the mean and standard deviation of the feature, respectively. Unlike the other pruning methodologies, which are a function of the *number* of the set of observations, z-score pruning is a function of a property of the *values* of the observations themselves. Specifically, if the mean $|z| < \theta$ for each occurrence of the feature,

the feature is eliminated from the overall feature-set.

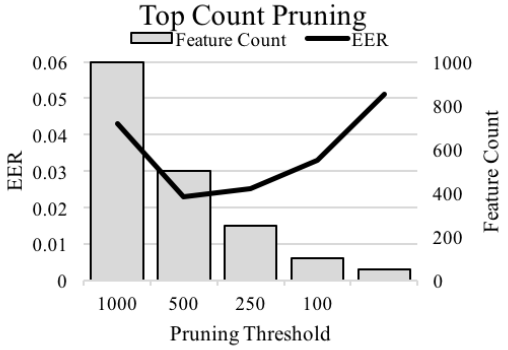
Figure 4.3 compares the various pruning methodologies, and provides both the resulting EER as well as the size of the feature-sets.



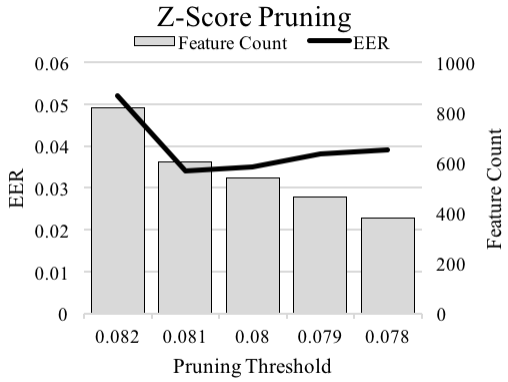
(A) Minimum-maximum pruning results



(B) Maximum-minimum pruning results



(C) Top feature count pruning results



(D) Z-score pruning results

FIGURE 4.3: Results of various pruning methodologies. Note that minimum-maximum pruning results in a substantially larger feature-set, resulting in a different y-axis scale for Figure 4.3a

Our best results came from taking only the 500 features with the most observations (Figure 4.3c). Most pruning methodologies exhibit a local minima in their results (Figures 4.3a, 4.3c, and 4.3d). However, maximum-minimum pruning (Figure 4.3b) proved to be too stringent, as illustrated by the fact that increasing the θ threshold never improved results.

4.4 Discussion

Our results investigate the hypothesis that language production as exhibited by keystroke timing is affected not only by neighboring keystrokes and other surface-level factors but also by latent linguistic factors. Taking these factors into account can improve the performance of a Scaled Manhattan verifier.

Throughout keystroke dynamic-based authentication studies, the most common temporal metrics utilized are unigraph hold (UH in Table 4.2) and digraph interval (DI in Table 4.2). If we take the EER of the fusion of these features to be a baseline (0.0483), then our most optimally pruned linguistic feature set improved authentication results by approximately 52% (Top 500 features contextualized by lemmatization), with an EER of 0.0232.

The fact that the features contextualized by their lemmas rather than by their raw word has important implications. A lemma represents the underlying abstract concept of a word Warren, 2012. By utilizing this abstract version of a word, our feature-sets actually become more informative. This may point to the influence of abstract lexical content on language production, although further experiments would be required to test that this is not an artifact of lemma data being more sparse than specific word data.

As seen in Experiment 1, even the most minor contextualizing improves authentication results. For key holds, a unigraph hold (UH) when considered within the context of the preceding keystroke (UH_D) improves EER by over 50% from 0.0949 to 0.0426. A keyhold measurement can also be improved by adding together the holds of two neighboring keystrokes. This form of contextualization improved EER by 33% from 0.0949 for unigraph holds to 0,0638 for digraph holds.

A similar improvement is seen for intervals. A unigraph pause (UP) only takes the subsequent keystroke into consideration when creating features. However, if a pause or interval is considered within the context of both its preceding and subsequent keystrokes (DI), then EER is also improved by nearly 50% from 0.1663 to 0.0890.

Further improvements were observed by fusing both holds and intervals, rather than considering each in isolation. This allows for features to provide a more full picture of the subject's typing, as it illustrates the time spent both between keystrokes and during a keystroke. By fusing unigraph holds, unigraph holds in digraph context and digraph intervals, EER was improved by 14% from 0.0426 to 0.0368 over the best performing feature-set made up of a single type of feature (UH_D).

Surprisingly, our findings in Experiment 3 seem to show that accuracy in measuring typing proficiency does not directly translate into accurate metrics for user authentication. While the findings in studies such as Alves et al. (2007), relating to typist skill level, were improved by focusing only on intra-word keystrokes, we did not find a similar phenomenon in user authentication.

We did find that intervals following a spacebar and preceding clause-final punctuation displayed more variance than intervals preceding a spacebar or following clause-final punctuation. However, eliminating the types of features that displayed greater overall variance did not improve results. Similarly, pruning features based on mean z-score did not improve upon results. These findings suggest that outliers are informative for user authentication rather than serving as a source of noise.

Finally, the results of our pruning experiments in Experiment 4 show that some measure of flexibility is essential when pruning a feature-set. The most narrowly defined methodologies for pruning did not provide optimal results. Maximum-minimum pruning allowed for the inclusion of any feature that had to exceed a threshold only once to be included. This resulted in too large of a feature-set. On the other hand, minimum-maximum pruning required that every observation of the feature exceed a threshold to be included. This resulted in too small of a feature-set, as just one participant *not* producing a feature would eliminate that feature from the feature-set.

Rather, top count pruning resulted in optimal results. By considering only the cumulative count over all features, any extremes would not have a significant impact on a feature being included or excluded. If a rare feature was produced in abundance by a single user, this would most

likely not be enough to have the feature included as an overall most-frequently occurring feature. Similarly, if only one user failed to produce a commonly occurring feature, this would most likely not have enough impact on the overall count to drop the feature from the most-frequently occurring features.

The results of the experiments relating to variance and pruning suggest the importance of two facets of language production and therefore importance to user authentication:

- A single subject's language production expresses a large amount of variation. Trying to systematically eliminate outliers will eliminate data points that are valuable to characterize and authenticate a user.
- Language production can vary greatly between participants even if certain features occur fairly frequently. Any authentication technique that does not allow for some flexibility in handling outlier participants will not produce optimal results.

Chapter 5

Linguistics and Demographic Cohort and Task Complexity Prediction

5.1 Introduction

Chapter 5 generalizes beyond individuals and single linguistic constructs to investigate if typing and linguistic production patterns maintain similarities across entire populations. We describe two applications of combining keystroke dynamics, stylometry and a new set of language production features: to identify the type of cognitive task a typist is performing and to identify three demographic cohort attributes.

In the case of predicting the type of cognitive task, we aim to determine whether the user is performing a cognitively simple task, such as recalling known information, or performing a more cognitively taxing task such as analyzing an argument or creating a new idea. The research presented in this chapter rests on two assumptions. First, we assume that different types of tasks will have different cognitive demands. This assumption is based on the Bloom's Taxonomy of Learning (Anderson, Krathwohl, and Bloom, 2001), widely used in education to categorize the cognitive demands of instructional activities. Second, we assume that the cognitive activity of a typist, particularly when performing a language production task, is reflected in his or her typing behavior. This assumption is supported by the findings of Vizer, Zhou, and Sears (2009), which

observed that a user's typing patterns vary based on the cognitive demands of a task.

Specifically, we hypothesize that the cognitive demands of performing a task will have an observable impact on a typist's behavior that can be measured through features related to keystroke dynamics, stylometry, and language production. In the first set of experiments presented in this chapter, participants respond to prompts which are drawn from different types of cognitive tasks. We then predict the type of task a participant is performing given the typing patterns and final static text.

For demographic prediction, we divide our participants along three broad demographic dimensions: gender, dominant hand and primary language (native vs. non-native speakers of English). Each of these demographic divisions may be viewed as a cohort with a different set of keystroke dynamics when compared to its counterpart. We aim to be able to place a user in a cohort based on the user's typing patterns and language use, such as "left-handed, female, native English-speaker". In the context of user identification and verification, this can be used as a filter to eliminate some candidates from further consideration enabling more focused downstream analysis.

This work employs a number of novel features for keystroke dynamics and stylometry. In addition to measuring hold and interval times of each key individually, we explore aggregations of keys based on their keyboard position, which distinguishes, for example, keys typed by the left and right hand. By performing stylometric analysis on streams of typed data, we are able to develop features measuring revision behavior in addition to the final, static text. Moreover we develop a number of language production features which extend traditional stylometric measures with information about their timing.

The most important contributions of our study are:

- We demonstrate how the type of task a typist is performing – based on the expected cognitive demand – affects typing output. Previous studies have centered around a homogeneous task type, whereas we can show the effects of varying cognitive demands.

- We propose and implement a new class of features, keystroke language production. These features take advantage of both keystroke dynamics and stylometry, to capture the dynamics, or prosody, of a typist's language production.
- The text being analyzed in this work is entered freely, with minimal constraints as to length or content. Moreover, predictions are made using much less data per answer than comparable studies, and with significantly more participants.
- Typical studies of this kind attempt to model the behavior of a typist and compare subsequent samples of the same person's typing to this model. In this work, we demonstrate the value of typing behavior to generalize to unknown typists, i.e. those not seen during training.

This chapter is structured as follows: Section 5.2 describes the methods that are in common between the two sets of experiments including details of the data collection (Section 5.2.1) and a description of the features we analyze (Section 5.2.2). Sections 5.3.1 and 5.3.2 describe experiments in predicting cognitive task and demography from an unknown typist, respectively.

5.2 Materials and Methods

In this section we describe the experimental methods shared by the experiments in predicting cognitive task (cf. Section 5.3.1), and recognizing demography (cf. Section 5.3.2). We first describe the data collection in Section 5.2.1. In Section 5.2.2, we describe the features used for the analysis and classification in both subsequent experiments.

5.2.1 Data Collection

As in Chapter 4, the typing data was collected from 486 Louisiana Tech University students. In addition, we asked participants whether they look at their hands when typing (visual typing) or look at the screen (touch typing); 64.7% of participants use touch typing, while 31.3% use practice visual typing. We note that this was self-reported, and not objectively verified. In other words, a

participant might *want to* believe that he does not look at his hands when typing, but in fact, he usually does.

All experiments below were performed on the entire population, both touch- and visual-typists, whereas some previous studies (cf. Song, Wagner, and Tian, 2001) report results from only touch typists. We believe this makes the reported results more robust, as we did not restrict our participant set to only a subset of the population based on typing style.[1]

Finally, as noted in Table 3.1, each level of Bloom's Taxonomy has an associated numerical level. Whereas in Chapter 3 we make no assumptions about the relationship among tasks, and consider each task as a discrete label, the current chapter assumes a more continuous relationship between the tasks or cognitive levels.

5.2.2 Features

The features used to analyze each subject's answers fall under 3 broad categories: keystroke dynamics, stylometry, and language production. There were a total of 2,381 features extracted. Table 5.1 describes the broad classes of features used in this work.

Due to the short length of participant responses, many features are not present in a given response. If a certain feature was not present in the test set, we replaced missing features with the mean feature value calculated over the training data. This guarantees minimal impact on classification performance due to unseen features. However, the 'count' of a particular event is never a 'missing' feature, a count is simply 0. An example of a missing feature would be the mean key interval between keys 'Q' and 'Z'. If a user has never typed these two characters in sequence, there is no mean interval that can be calculated. No outlier removal or modification was applied to the observed features. While this may improve performance, we are hesitant to modify any observed data, as what may appear to be an outlier, may be an important signal to some population or uniquely indicative of a task.

[1] In unreported results, we also repeated all experiments on only touch-typists. We found that the results were only marginally different, and not consistently better nor worse.

Feature Group (Count of Features)	Feature Type	Example Feature
Keystroke Dynamics (2098)	Key Hold	`Mean Shift Key Hold, Mean H Key Hold`
	Preceding Pause	`Variance of Pause Before Spacebar, Mean Pause Before T Key`
	Hand-based	`Mean Left Hand Key Hold`
	Finger-based	`Variance of Index Finger Key Holds`
	Keyboard Row	`Mean Pause Preceding Home Row Keys`
	Common/Rare Consonant/Vowel	`Mean Pause Preceding Common Consonants`
	Common to Rare Character Ratio	`Ratio of Mean Consonant to Vowel Key Hold`
Stylometry (89)	Sentence Metrics	`Mean Sentence Length, Total Sentence Count`
	Word Metrics	`Median Word Length`
	Character-Type Metrics	`Alphabetic to Numeric Character Ratio`
	Capitalization Metrics	`Capital to Lowercase Character Ratio`
	Type-Token Ratio	`Lexical Diversity, Lexical Density`
Language Production (194)	Part-of-Speech Timing	`Variance of Pause Preceding Noun`
	Punctuation Timing	`Mean Pause Preceding Comma`
	Misspelling Metrics	`Ratio of Misspelled to Correctly Spelled Words`
	Revision Metrics	`Mean Time Spent Revising Text`
	Lexical Units Within Burst	`Mean Words Produced in Typing Burst`

TABLE 5.1: Feature List for Cognitive Complexity and Demographics Experiments

Keystroke Dynamics Features

Keystroke dynamics looks at the speed at which a user's hands move across a keyboard (Bergadano, Gunetti, and Picardi, 2003) and the timing between keystrokes. The features analyzed in the present study capture rate and rhythm qualities including the overall user typing speed, durations and frequencies of pauses in typing, and pauses before specific keys.

As noted in Table 5.1, we created a large amount of keystroke dynamics features. This is to be expected, as the size of a feature set capturing every key combination would be equal to the number of keys squared. This leads to many empty features. Thus the number of features reported in the above table represents an upper bound on the effective dimensionality of the feature vectors. We note that all empty features are ignored by all classifiers.

With greater than 90% accuracy, a typist can be identified by the rate and rhythm of their typing see (see Baaijen, Galbraith, and Glopper, 2012; Bergadano, Gunetti, and Picardi, 2003; Canales et al., 2011; Gunetti and Picardi, 2005; Killourhy and Maxion, 2010; Killourhy and Maxion, 2009b). The rate of keystroke production may be indicative of familiarity with the typed material. Taking the latter notion one step further, familiarity with the typed material may have a multitude of underlying causes, from the physical presence of the typed text or availability in memory (affecting

cognitive demand) to the native language of the typist. Here we investigate whether these measures are consistent across a population by virtue of their demographics, or by the task which they are engaged in.

Key Intervals and Key Holds: One of the staple metrics of keystroke dynamics, digraph rates measure the latency between any two keystrokes and the duration that each key is depressed (Joyce and Gupta, 1990). For the present study, we utilize the mean latency between any two keystrokes, including punctuation, symbols and numbers, and the mean "hold", the duration of depression, for each key.

Consonant timing: If a user is more familiar with English, and the distribution of letters in the English language, it stands to reason that he or she is also more adept at quickly depressing the more common letter keys. On the other hand, if a user is newer to the English language, he or she may be equally adept at striking any keyboard key (Gunetti, Picardi, and Ruffo, 2005). By looking at the timing surrounding common and rare characters, we hoped to determine language familiarity.

- Common consonant timing: Mean time between pressing any key and then pressing a common consonant. Common consonants are defined as elements of the set (h, n, r, s, t) (Stewart et al., 2011).
- Common consonant timing ratio: Ratio of pause between any two keystrokes and pause between a keystroke of any key to a keystroke of a common consonant key.
- Rare consonant timing: Mean time between pressing any key and then pressing a rare consonant key. Rare consonants are defined as elements of the set (j, k, q, v, x, z) (Stewart et al., 2011).
- Rare consonant timing ratio: Ratio of pause between any two keystrokes and pause between a keystroke of any key to a keystroke of a rare consonant key.
- Common to rare consonant timing ratio: Ratio of common consonant timing to rare consonant timing.

Hand- and Finger-specific timing: Traditional keystroke dynamics performs measurements based on each key individually. In this set of features we group keys by their canonical hand and finger that would be used to type them based on "touch-typing" norms. For example, the 'A' key is indicated by LEFT HAND, LITTLE FINGER. This categorization enables us to measure the key holds intervals grouped by hand (LEFT, RIGHT) and finger INDEX, MIDDLE, RING, and LITTLE FINGER). We do not include the thumb in these measurements. The space bar is the only canonical key for the thumb, and our data is not able to distinguish which thumb is used to hit the spacebar. Based on these broad classes, we extract key holds and key intervals based on 1) hand, 2) finger, and 3) finger and hand.

Keyboard row production rates: Similar to the features mentioned above we group keys by canonical row. For example, every key in the HOME ROW, i.e. CAPS LOCK though ENTER, is grouped under one category. Given these categories, we extract key hold and interval measures based on the key's row.

Stylometric Features

Stylometry incorporates syntactic, lexical and semantic analyses of a given text. Every aspect from average sentence length to part-of-speech frequency falls under the purview of stylometric analysis. Stylometry analyzes static text rather than the dynamic features of text production (cf. Juola, 2006).

Stylometry represents measurements of linguistic information to quantify the individual "style" of a writer. The development of authorship attribution has been markedly improved by the development of more sophisticated Natural Language Processing techniques (Stamatatos, Fakotakis, and Kokkinakis, 2001). We hypothesize that the writing style of a typist is impacted by the cognitive demands of the given task, and moreover, that stylometric features reveal key demographic information.

Linguistic unit lengths: The most basic set of stylometric features counts the number and length of linguistic units, words and sentences.

- Sentence count: Utilizing Apache OpenNLP's Sentence Detector (Baldridge, 2005), rather than rely on common sentence-terminating punctuation, we counted the number of sentences per response.
- Mean sentence length: Utilizing the same resources as sentence count, the mean number of word tokens per sentence was determined. As a language user gains better command of a language, their mean sentence length also becomes longer (Stamatatos, Fakotakis, and Kokkinakis, 2001). This has repercussions to native language, and cognitive load classification.
- Word token count: The number of word tokens in a response. The tokenizer divides words such as "that's" into the two words, "that" and "is".
- Mean word token length: A measure of the mean word token length, in number of characters.

Character type: Use of, or disuse of, specific character types can aid in the identification of typist (Grieve, 2007). Certain users may prefer to write out numerals, e.g. "four", versus Arabic numerals, e.g. "4." Further, users tend to exhibit patterns in capitalization, e.g. "Soccer Champion" versus "soccer champion" (Vel, 2000). We contrast these linguistic types to the keyboard position types described previously.

- Alphabetical character ratio: Ratio of alphabetical characters (a-z, non-numeric) to total number of keystrokes
- Numeric character ratio: Ratio of numeric characters (0-9) to total number of keystrokes
- Uppercase character ratio: Ratio of uppercase characters to total alphabetical characters
- Spacebar ratio: Ratio of total depressions of space bar to total keystrokes
- Vowel ratio: Ratio of vowels to total number of alphabetical characters

Consonant frequency We hypothesize that the use of common or rare consonants (cf. Stewart et al., 2011) are indicative of lexical complexity. If a participant uses a high ratio of rare consonants, this could be a marker of more sophisticated word use, and/or a more advanced vocabulary.

- Common consonant ratio: Ratio of h, n, r, s, t to total alphabetical characters
- Rare consonant ratio: Ratio of j, k, q, v, x, z to total alphabetical characters

Lexical Diversity: Lexical diversity measures the ratio of unique words to total words. As a metric of writing style, lexical diversity is one of the oldest methods for authorship attribution, preceding many more complicated analyses (Holmes, 1985).

- Type-token ratio: This is the most basic measurement of lexical diversity, in which the number of unique word tokens is divided by the total number of words. This generally reflects the size of a typist's vocabulary, as a larger vocabulary results in the use of a greater number of different words.
- Moving-average type-token ratio (MATTR): The primary shortcoming of the type-token ratio is that it does not control for length, i.e. longer texts will usually have a lower ratio (Covington and McFall, 2010). MATTR, on the other hand, only considers a fixed number of words at a time, and increments through the text, e.g. words 1-50, 2-51, 3-52, etc. This produces more informative results when comparing texts of different lengths (Covington and McFall, 2010).

Lexical Density: Lexical density measures the number of unique parts of speech divided by the total number of word tokens. Higher lexical density is used as a measure of language complexity. (Ure, 1971)

Language Production Features

Production features are a hybrid of the above two categories, incorporating elements from both linguistic analyses and keystroke rate and timing. While stylometric and keystroke dynamic features are measured independently of one another, language production features use elements of both to create unique categories of features. For example, while stylometric features may look only at the frequency of verbs to nouns, and keystroke dynamics may look only at average keystroke typing speed, language production features may measure the average typing speed of verbs versus nouns.

Our investigation is informed by linguistic meta-information. Given that languages exhibit certain predictable linguistic patterns, we hypothesize that these patterns are also borne out during the typing process (Bergsma, Post, and Yarowsky, 2012). By exploiting this fact, we hope to gain a more nuanced understanding of typing patterns, based on lexical and syntactic patterns.

The features we created, based on a hybridization of keystroke dynamics and stylometry, are as follows:

Part-of-speech pauses: We measure the mean length of a pause before and after each word as represented by its part of speech. By combining pause data with this syntactic data, we hope to deduce underlying features of a typist's habits. The parts of speech used in this study were: a) Nouns: singular, plural, proper, gerund; b) Verbs: verbs of all tenses and persons, past participles, modals; c) Modifiers: numbers, determiners, adjectives, adverbs, wh-determiners, wh-adverbs.

Punctuation pauses: This metric measured the pause time before and after punctuation marks that break up phrases or units of thought. These include sentence ending punctuation (periods, exclamation points, question marks), commas, and semicolons (Vizer, Zhou, and Sears, 2009). We hypothesize that a user who pauses for a greater length of time around a phrase terminating punctuation mark is engaging in planning behavior indicating greater cognitive load.

Misspelling pauses: This metric used data from Jazzy Spell Checker (Idzelis, 2013) to identify the correct spelling and common misspellings of individual words. From this we calculate whether a user pauses longer or more often before and after misspelled words.

Revision Features: We hypothesize that a typist's behavior when revising previously typed text is influenced by cognitive load and language familiarity. We define "in revision" as any delete or backspace keystroke and any time at which the typist is not at the leading edge of the buffer, but rather has gone back and made a revision. This allows us to characterize a user's typing as being "in revision" or not. Using this distinction, we extract a number of features to capture a typist's revision behavior.

- Mean character length of revisions: A typist's behavior or demography may be characterized by whether he or she makes long or short revisions.
- Mean length of time in revisions: We hypothesize that a typist may pause and think more during a revision if he or she is performing a task with greater cognitive demands, while typists executing less demanding tasks will make brief, immediate corrections.
- Revision ratio: This feature measures the length of time in revision to the time of the overall typing session. This can be loosely considered as how "confidently" a typist is behaving, i.e. whether he or she is constantly backtracking and making corrections, or spending more time in the production of novel text.

Typing Burst Features: In addition to segmenting lexical events by traditional grammatical units, e.g. sentences, we also divide typing sessions according to when a user paused during his or her typing. We defined a pause as a cessation in typing greater than 250 milliseconds. This number is based on findings in Baaijen, Galbraith, and Glopper (2012), which found similar timing to be indicative of a suspension in typing activity, as opposed to a keystroke interval. After breaking down typing sessions in this manner, we analyze the events that took place between pauses.

- Mean word count between pauses: This feature provided a baseline measurement of the number of complete words that occurred between each pause. If a typist is producing in a more "stream of consciousness" mode, we predict that this is indicative of cognitive task, as well as greater language familiarity.
- Mean word count between sentence start and pause: We measured the number of words that were produced between the beginning of a new sentence and the typist's first pause.
- Mean character count between sentence start and pause: Similar to the previous feature, this feature counts the number of characters. By counting characters as opposed to words, we can control for word length, and detect features such as typist fatigue, which may be mediated by character count rather than word count.

5.3 Experiments

In this section, we present two sets of experiments: 1) predicting the cognitive task (Section 5.3.1), and 2) predicting the demographic indicators: gender, handedness and native language (Section 5.3.2). Both experiments use the data and features described in Section 5.2.2. In the following subsections we describe any methods that are specific to only one of the experiments. Following this, we present and discuss experimental results.

5.3.1 Experiment 1: Prediction of Cognitive Task

We conduct four experiments to predict the type of cognitive task a typist is engaged in as described in Section 5.2.

5.3.1.1 Experiment 1: Methods

We divide participants between one of two distinct sets, training and test. The same participant pool was used for all experiments, where each testing and training set included 352 distinct participants, for a total participant pool of 704 participants. These subsets are constructed such that that no participant in the training set was included in the test set, and vice versa.

Moreover, we use responses to prompts from session 1 for training, and prompts from session 2 for testing. Thus no specific prompt was used for both training and testing. This allows us to recognize the type of task rather than recognizing specifically the prompt to which a participant was responding.

Training and testing sets contained approximately equal numbers of each type of cognitive task. We used Weka (Hall et al., 2009) to predict which type of task was being performed, using four classifiers: Naive Bayes, AdaBoost with single split decision trees, SVM with an RBF Kernel, and SVM with a Linear Kernel. For each experiment, available parameters are tuned using ten-fold cross validation on the training data. All of the classifiers, save Naive Bayes, include protections against overfitting in their training procedure (either via a sensitive objective function, or explicitly

training criteria). They are well motivated for use in situations where the number of available features is large with respect to the number of data points.

Our choice of classifiers centered around the types of features each could best work with. For example, SVM was used because it is defined by a convex optimization problem (no local minima). AdaBoost was selected for its ability to ensemble multiple (weak) classifiers. For all classifiers such as SVM, Logistic and Naive Bayes, missing features were imputed as the feature mean. These represent a set of classifiers that have shown to be effective on a variety of other classification tasks.

Moreover, with the exception of Naive Bayes, all are well motivated in classification contexts where the dimensionality of the feature vector is large, relative to the number of training data points; they each include protection against overfitting via regularization, effectively eliminating irrelevant features from the model. By exploring a range of classification routines we are able to measure the difficulty of the classification task, rather than making any assumptions about which specific classifier will be most effective *a priori*.

We explore a number of different ways to represent cognitive task. First, we attempt to classify each task individually. However, since assigning cognitive labels can be subjective, we also attempted broader classifications. We divided the tasks in groups of 2 and groups of 3, and also looked at the polar extremes, comparing only the most demanding tasks to the simplest tasks under Bloom's Taxonomy, to remove any noise from moderately demanding tasks. Finally, we attempt a regression analysis, assigning each task a number from 1 to 6, to determine if a linear relationship between the tasks is helpful (cf. the ordering of Bloom's original taxonomy).

5.3.1.2 Experiment 1: Results

The results of experiments predicting cognitive task are reported in Table 5.2. Baseline accuracy values appear in parentheses below the class granularity identifier along with the total size of the experiment. For the sake of conciseness, we used the numeric labels (1-6) for each task (cf. Table 3.1). However, for classification experiments, each task was considered a discrete class. The best

classifier for each experiment is bolded. P-values based on a one-tailed binomial proportion tests are included for all results.

Class granularity		Naive Bayes	AdaBoost	SVM-RBF	SVM-Linear
6-way (16.67%)	Acc.	17.71 %	22.14%	**33.14%**	31.61%
N=4236		p=0.036	$p< 10^{-19}$	$p< 10^{-150}$	$p< 10^{-125}$
3-way (33.33%)	Acc.	35.17%	**49.66%**	48.11%	47.26%
N=4236		p=0.0059	$p< 10^{-105}$	$p< 10^{-87}$	$p< 10^{-77}$
2-way (1,2,3 vs. 4,5,6) (50.00%)	Acc.	47.07%	58.55%	59.56%	**59.70%**
N=4236		p=0.99	$p< 10^{-28}$	$p< 10^{-35}$	$p< 10^{-36}$
2-way (1,2 vs. 5,6) (50.00%)	Acc.	49.22%	66.66%	69.42%	**69.68%**
N=3179		p=0.81	$p< 10^{-79}$	$p< 10^{-108}$	$p< 10^{-111}$
2-way (1 vs. 6) (50.00%)	Acc.	53.95%	64.97%	70.55%	**72.39%**
N=1416		p=0.0016	$p< 10^{-29}$	$p< 10^{-54}$	$p< 10^{-65}$

TABLE 5.2: Results of cognitive demand identification experiments

We also reran the cognitive task recognition experiment using only the subset of participants who were native English speakers. By running the experiment with only this subset, we aimed to elucidate whether non-native speakers were a source of noise, as their cognitive demands would include not only responding to the prompt but responding in a non-native language. Surprisingly, the results were only marginally improved. Correctly classified instances increased by 1.3%, and mean F-score only increased from 0.286 to 0.307. Perhaps this speaks to the robustness of our feature-set, in that it cuts through any noise not related to cognitive demand. However, before drawing any strong conclusions, a more thorough analysis is required.

Moreover, Table 5.3 lists how many answers were misclassified by 1, 2, 3, 4 and 5 levels. These results closely parallel the above classification results.

5.3.1.3 Experiment 1: Discussion

For each of the experiments and for each classifier (with the exception of Naive Bayes on 1,2,3 vs 4,5,6), we find that we are able to predict which type of task is being completed at or above baseline

Margin of Error	Accuracy	Random Baseline
0	16.48%	16.67%
1	50.19%	44.44%
2	80.71%	61.11%
3	96.46%	77.78%
4	99.48%	94.44%
5	100%	100%

TABLE 5.3: Margin of error over a random baseline for cognitive demand predictions in Experiment 1.

regardless of how we group the tasks. The independence of both participant and prompt across training and test sets demonstrates that these features, based on keystroke dynamics, stylometry and language production are able, to some degree, to capture differences in the cognitive demands of an unknown prompt being undertaken by an unknown subject.

We observe that while there is not a single best classifier to predict cognitive demands, the SVM variants perform consistently and reliably above chance. The choice of kernel–RBF or Linear–never leads to significant differences in accuracy at the 0.01 level. Differences between these two classifiers are minimal.

Through inspection of the results of the three binary classification experiments, distinguishing HIGH (CREATE and EVALUATE tasks) and LOW (REMEMBER and UNDERSTAND) cognitive demand tasks we can draw some conclusions about the differences between these categories. We find that as we restrict data points to instances of more extreme examples of cognitive demand, the overall accuracy increases. This is to be expected.

We also observe that the ability to differentiate HIGH from LOW cognitive demands when the intermediate tasks are included, achieves a performance not exceeding 60%. By omitting these data points – the REMEMBER, UNDERSTAND vs. CREATE, EVALUATE – we see a $\sim$10% absolute improvement to accuracy on both the SVM and AdaBoost classifiers. This suggests that the inclusion of the APPLY and ANALYZE tasks is a source of noise. Despite having fewer data points, we are seeing greater differentiation in the smaller data set. In contrast, the difference from omitting

the level 2 and 5 data points is much more modest; the absolute change to accuracy is at most 2.71%.

When examining the results of finer distinctions of cognitive demand, we still achieve performance that exceeds the chance baselines. In the 6-way classification, SVM classification with RBF Kernels can predict the type of task with 33.14% accuracy over a baseline of 16.67%. The performance on 3-way classification, distinguishing HIGH (CREATE and EVALUATE), MID (APPLY and ANALYZE) and LOW (REMEMBER and UNDERSTAND) achieves 49.66% accuracy over a 33.33% chance baseline ($p < 10^{-105}$).

These results suggest that it is possible to recognize what sort of task an unknown person is performing based on inspection of a relatively short observation (roughly 450 characters) of typing behavior. It is not yet determined what should be considered the upper bound for performance on this task. Some of the errors are due to individual differences between typists, other errors are due to the discrepancy between different questions labeled as the same type of task. Another source of error is the use of a label for cognitive demand based on how we *expect* a typist to respond cognitively to a type of prompt instead of using a more empirical measure of the cognitive demand the participants are, in fact, experiencing.

A number of features proved to be especially useful in these experiments. Up to 12 of the most useful feature names are listed in Table 5.4 by the experiment performed along with each feature's Information Gain Ratio, and its relationship to Bloom's Taxonomic cognitive level of the task. For the relationship to cognitive demand, we determine one of five relationships based on the relationship between the feature value and the cognitive level associated with a task:

- Ascending ($\nearrow$): Increasing with higher cognitive level
- Descending ($\searrow$): Decreasing with higher cognitive level
- Bimodal ($\vee$): Strictly lower with middle values of cognitive level
- Unimodal ($\wedge$): Strictly higher with middle values of cognitive level
- Multimodal ($\sim$): Any relationship other than the above

We find that for the 6-way classification, the relevant features do not show consistent directions in their relationship with task.

The three-way classification of cognitive demand includes many groupings of keys by keyboard region, as opposed to specific bigrams. It is unclear for each of these which specific words lead to increased or decreased observances of these features, but it suggests that there is benefit in looking at broad classes of letters, rather than individual key intervals, when there is enough data to generalize from.

The binary classification results show some consistent differences between high and low cognitive demand tasks. Responses to prompts that are likely to require higher cognitive demand tend to include more modal verbs (e.g. "could", "may"). Responses to prompts requiring greater cognitive demand also contain a decreased upper case ratio – possibly due to longer sentences as well as fewer proper names, fewer space characters – indicating fewer words – increased lexical density – a measure of the complexity of the sentence – and a significant amount of differentiation between groups of keys both by keyboard position and the distinction of rare and common consonants.

Class Granularity	Info Gain	Feature
6-way	0.3724 $\sim$	Modal Counts (Mean, Median)
	0.2200 $\sim$	Lexical Density (Mean, Median)
	0.2054 $\sim$	"D" unigram Count
	0.1819 $\sim$	Right Index to Right Index (Count)
	0.1746 $\sim$	Right Ring to Left Middle (Count)
	0.1674 $\sim$	Characters Per Word (Median)
	0.1631 $\sim$	Right Ring to Right Index (Count)
	0.1568 $\sim$	Middle Finger to Thumb/Space (Count)
	0.1563 $\sim$	Characters Per Word (Mean)
3-way	0.1306 $\nearrow$	Modal Counts (Mean, Median)

Class Granularity	Info Gain	Feature
	0.1120 ∨	Right Ring to Right Index (Count)
	0.1101 ↘	Right Index to Right Index (Count)
	0.0920 ∧	"U" unigram Count
	0.0873 ∨	Characters Per Word (Median)
	0.0873 ∨	Top Row Right to Bottom Row Right (Count)
	0.0821 ↗	Right Index to Left Little Finger (Count)
	0.0795 ∨	Characters Per Word (Mean)
	0.0765 ∧	Bottom Row Left to Top Row Left (Count)
	0.0753 ∨	Uppercase Ratio (Mean, Median)
2-way	0.0864 ↘	Characters Per Word (Median)
{1,2,3 vs. 4,5,6}	0.0791 ↘	Characters Per Word (Mean)
	0.0572 ↗	Middle to Ring (Count)
	0.0530 ↗	Right Index to Left Ring (Count)
	0.0514 ↘	Top Row to Thumb (Count)
	0.0509 ↘	Top Row to Space (Count)
	0.0497 ↘	Function Word to Vowel (Count)
	0.0493 ↗	Right Index to Home Row (Count)
	0.0490 ↘	None to Left (Count)
	0.0487 ↗	Top Row to Left Ring (Count)
	0.0421 ↗	Right Index to Left Little Finger (Count)
	0.0418 ↘	Speed Thumb
2-way	0.1468 ↗	Modal Counts (Mean, Median)
{1,2 vs. 5,6}	0.0855 ↗	Top Row to Ring (Count)
	0.0730 ↗	Right Index to Home Row (Count)
	0.0667 ↗	Right Ring to Left Middle (Count)

Class Granularity	Info Gain	Feature
	0.0601 ↗	Right Index to Left Little Finger (Count)
	0.0535 ↗	Top Row to Right Ring (Count)
	0.0506 ↗	Ring to Middle (Count)
	0.0489 ↗	Right to Left Home Row (Count)
	0.0454 ↗	"OU" bigram Count
	0.0452 ↗	Top Row Right to Home Row Right (Count)
2-way	0.43652 ↗	Modal Counts (Mean, Median)
{1 vs. 6}	0.20148 ↗	U Count
	0.19811 ↗	"OU" bigram Count
	0.1859 ↗	Right Ring to Left Middle (Count)
	0.18245 ↘	Lexical Density (Mean, Median)
	0.17017 ↗	"D" unigram Count
	0.15441 ↗	Right Index to Home Row (Count)
	0.15031 ↗	Top Row Left to Top Row Right (Count)
	0.14758 ↗	Left Middle to Space Row (Count)

TABLE 5.4: Cognitive Demand Feature Relevance Measured by Information Gain

5.3.2 Experiment 2: Prediction of Demography

In this section we describe a number of experiments to predict three demographic indicators: gender (male vs. female), handedness (left vs. right) and primary language (English vs. non-English).

5.3.2.1 Experiment 2: Methods

The demographic labels were extracted according to how the participants identified themselves (cf. Section 5.3.1.1). Although we recognize there may be differences between the way participants

identify themselves and their actual demographics, we nonetheless identify the participants based on these labels in these experiments. As in the cognitive load prediction, demographic prediction experiments are all performed on training and test data which contain both different participants and different prompts. As some participants did not respond to demographic questions, the number of distinct participants in the train and test partitions for these experiments were 329, 344, and 348 for handedness, gender, and primary language, respectively. For consistency we kept the size of train and test sets equal in all experiments.

Again, we use Weka (Hall et al., 2009) to perform classification experiments. Unlike the cognitive load classification experiments, the distribution of males and females and English and non-English speaking participants are not even. Only 41.3% of participants are female, 9.1% are left-handed and 17.0% have a native language that is not English.

Thus, we treat these tasks as *detection* tasks, where the challenge is identifying the minority class, either Female, Left-handed or "non-English". As detection tasks we evaluate performance using F_β-measure with $\beta = 1$ in detecting the minority class and ROC area of the demographic classification experiments using the features described in the previous section. We conduct our experiments with four classifiers: LogitBoost, Naive Bayes, SMO (RBFKernel) and SimpleLogistic, tuning hyper-parameters through ten-fold cross validation on the training data. These classifiers were selected for their effective detection of minority classes when the training data has a skewed class distribution.

5.3.2.2 Experiment 2: Results

One pair of experiments uses all training data ("Unbalanced"), wherein the majority of demographic labels belongs to one class. Another pair ("Balanced") is conducted on a downsampled set in which the number of labels of the majority set matches the number in the minority set for the training material only. Downsampling is a standard technique used to address imbalanced data sets (Japkowicz and Stephen, 2002). A summary of the results appears in Table 5.5. Baseline

F-measure values appear in parentheses below the task. This baseline is based on random class assignment based on the unmodified training distribution.

The F_1 measure, or the harmonic mean of precision and recall, is a way to measure accuracy. *Precision* is the number of true positives (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of instances labeled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class). *Recall* is defined as the number of true positives divided by the total number of instances that actually belong to the positive class, i.e. the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been (Wikipedia, 2014). Higher F-measure indicates better performance. F-scores are used heavily in the Information Retrieval community, but has been applied to other detection tasks.

The F_1-measure does not lend itself to parametric distribution for statistical significance testing. To generate a p-value for these results, we use a randomization approach. We assign each data point a random class based on the training class distribution, and measure the F_1 score of this random assignment. We repeat this random process 50,000 times, and calculate the rate at which the random process yields a higher F_1 score than the classifier. This results in a measure of p-value for the null hypothesis.

5.3.2.3 Experiment 2: Discussion

We have varying levels of success in recognizing the demography of an unknown typist, although we do consistently classify with greater than chance performance. We are encouraged that we are able to predict with performance greater than chance using as little as 300 characters of data, however, there is still room for improvement here.

In predicting primary language, we see a clear benefit from balancing the distribution of the training data; performance increases on all classifiers. We do not see the same benefit in predicting gender nor handedness. For gender we find performance to decrease, and handedness has mixed

Demographic	Training		LogitBoost	Naive Bayes	SVM-RBF	Logistic
Gender	Unbalanced	F_1	0.518	0.473	0.485	**0.524**
(0.447)		p-value	0.0000	0.0022	0.0000	0.0000
Gender	Balanced	F_1	**0.516**	0.468	0.462	0.513
(0.447)		p-value	0.0000	0.01046	0.0484	0.0000
Handedness	Unbalanced	F_1	0.010	**0.183**	0.043	0.113
(0.100)		p-value	1.0	0.0000	1.0	0.1917
Handedness	Balanced	F_1	0.050	**0.223**	0.009	0.097
(0.100)		p-value	0.9999	0.0000	1.0	0.5871
Primary Language	Unbalanced	F_1	0.170	**0.355**	0.000	0.254
(0.166)		p-value	0.3818	0.0000	1.0	0.0000
Primary Language	Balanced	F_1	0.037	**0.462**	0.455	0.387
(0.166)		p-value	1.0	0.0000	0.0000	0.0000

TABLE 5.5: Results of prediction of demographic recognition experiments

results. This is due to the fact that there is greater imbalance in the language and handedness labels compared to gender. The gender classification suffers from the reduced size of the training data while not benefiting sufficiently from class balance. For gender classification we do not find a consistently best performing classifier, though Simple Logistic yields the best overall results. For primary language and handedness classification, we find the Naive Bayes classifier to generate the best results. This is somewhat remarkable, as Naive Bayes includes no protection against overfitting when dealing with large (and potentially sparse) feature vectors.

We also divided demographic prediction results by cognitive task to see if any task was particularly better or worse for predicting a certain demographic. We found the results to be consistent across tasks, with the accuracy varying by less than 1%.

In addition, we compiled prediction results for each answer. For 55% of the answers, we correctly predicted all 3 demographics. For 95% of the answers, we correctly predicted at least 2 of the 3 demographics. Being able to correctly predict demographics is of great interest to user verification application by providing a mechanism to pare down or prune a pool of participants.

We find a number of features are especially useful in predicting these demographic labels. For gender classification, three features are especially useful:

1. the timing between a punctuation symbol and the spacebar is faster for male participants than for female participants,
2. the timing before and after function keys is faster for males,
3. the timing before and after common digraphs (such as "ou" and "er") is faster for females.

We hesitate to speculate as to the source of these differences. We do, however, find it interesting that the differences are in language production and keystroke dynamics, but not in traditional stylometric indicators. This novel result is in contrast with previous studies which have found gender to be reliably predicted based on stylometry (Goswami and M. Rustagi, 2009; Vel et al., 2002; Koppel, Argamon, and Shimoni, 2002). One source of this disparity is the length of the analyzed text; we are using a few sentences, while these studies have used full blog posts, emails, and full length texts averaging over 30,000 words, respectively.

For primary language, we find non-native English participants are typically slower typists than those who identify as native speakers. Three features are especially telling:

1. the timing before and after function keys including shift and backspace keys,
2. the timing before and after the "." key,
3. the timing before and after common digraphs ("ou" and "er" for example).

On each of these, non-native English typists are slower than native English speaking typists. While we expected to see faster typing in the native English participants, the differences are not so dramatic as to make it trivial to distinguish these groups. It is possible that our participants, all students in an American university with experience typing, are more familiar with the QWERTY keyboard and with the language than an average person whose primary language is not English. Each of the three features mentioned previously contribute more to effective classification of a typist as a native or non-native English speaker than overall typing rate. The increased pauses around function keys and the period (".") key may be evidence of increased sentence planning

time, suggesting that non-native English speakers type most words at a similar rate as native English speakers, but take extra time planning a sentence or making a revision (longer pauses around backspace and delete keystrokes). These are novel observations which warrant additional exploration to understand the impact of a speaker's native tongue on their typing behavior.

Chapter 6

Conclusion

The typing process is affected by myriad pressures, ranging from motor skills and limitations to cognitive and psychological constraints. This thesis purports to demonstrate the significant effect of linguistic structure on typing dynamics.

Chapter 3 demonstrated that linguistic structure, *per se*, can have a noticeable affect on typing. These effects can be seen independent of the probabilistic structure of the word sequences, e.g. n-gram conditional probability. Further, these linguistic patterns are able to provide insight into higher-level cognitive processes, which might be more difficult to discern in noisier data.

The linguistic effects observed in Chapter 3 are then expanded in Chapter 4 to a broad array of linguistic phenomena including syntactic, semantic and lexical structure. Not only do these linguistic properties affect the typing process, they also create consistent and reliable changes from individual to individual. This allows for more informative and fine-grained feature-sets, which can be used as a biometric for more accurate identity verification.

Finally, the prominence of linguistic pressures is expanded to large populations in Chapter 5. The unique affects of linguistic structure are not only reliable from individual to individual, but also create consistent changes over entire demographic cohorts, as well as reliable changes when the cognitive complexity of a task is altered. Chapter 5 also provides novel synthetic features which combine the most informative aspects of both keystroke and linguistic production.

Overall, these experiments also demonstrate the utility of keystroke dynamics as a high-accuracy,

low-resource form of investigation. The experiments in Chapter 5 are more robust because they demonstrate that informative features can be created independently of a specific typist or a specific type of prompt. This avoids the costly and time-consuming constraint of many biometrics which require a "templating" process to form an outline of an individual. Moreover, all of the experiments use orders of magnitude less data per person to achieve our results, compared to previous studies (cf. Vel et al., 2002; Koppel, Argamon, and Shimoni, 2002). Further, we use a much larger participant pool than comparable studies.

In sum, keystroke dynamics can provide rich information about cognition and linguistic structure. This information can be collected and annotated with relatively low overhead compared to dynamic speech production, but can still provide stark insights. The complexities of language are immense; perhaps further studying typing dynamics will aid in scaling down this immensity.

Appendix A

Appendix A: Essay Prompts Posed to Subjects in Data Collection

In this appendix, we include a full listing of the prompts used in the collection of typing data. The collection was divided into two sessions, containing distinct prompts with similar expected cognitive loads. Table A.1 contains the cognitive load, session number and prompt text for all prompts.

Cognitive Load	Session	Prompt
1	1	List the recent movies you've seen or books you've read. When did you see or read them? What were they about? Please use complete sentences.
1	1	Which sport(s) do you like to watch/play?
1	1	What made you decide to join Louisiana Tech University?
1	2	What are some things that you like about Ruston?
1	2	What are your favorite things about winter?
1	2	What is the best thing you ever ate at a restaurant? Describe it.

Cognitive Load	Session	Prompt
2	1	Where is a place that you particularly enjoy visiting? Describe what makes you happy about being at this place.
2	1	What is your favorite place to go out for a meal? What do you like about this place?
2	2	What would you say has been the best college class you have taken and what did you enjoy about that class?
2	2	What is something that you dread talking to your family about? Why do you not like to talk to them about this?
2	2	Can you describe the process of applying to college?
3	1	What would you do if you and a friend are on vacation alone and your friend's leg gets cut? Describe what the procedure you would use for first aid or for finding help.
3	1	What would you do if you were home alone and a fire started?
3	2	Suppose you were in NYC and had a very important presentation to give at 8AM the next morning at Louisiana Tech. You get to the airport in New York to discover that your flight has been delayed and will likely cause you to miss your layover in Atlanta. What steps would you take to ensure that you are at Louisiana Tech in time for your presentation?
3	2	What would you do if you woke up and realized your car would not start?

Cognitive Load	Session	Prompt
4	1	Explain what you think the difference is between "communicating with" someone and "talking to" someone. How are these two terms often confused?
4	1	Compare and contrast two genres of music.
4	2	Compare and contrast two sources you use for news and current events.
4	2	Give step-by-step driving directions to your favorite place in or around Ruston.
4	2	Explain what the saying "Not all that glitters is gold" means.
5	1	What email provider do you think is the best?
5	1	What social networking web-sites do you use?
5	1	Do you think it's a good idea to raise tuition for students in order to have money to make improvements to the University? Why or why not?
5	2	Do you think that capital punishment should be legal? Why or why not?
5	2	Do you think people should be required to have car insurance? Defend your decision.
6	1	Pretend a Hollywood executive offered to pay you to write and act in a movie. Create a movie plot with a character in it for yourself and remember that you will only be paid for creating an original plot to a movie.

Cognitive Load	Session	Prompt
6	1	If you were to create a picture of any type of landscape you wanted what objects would you include in it? How would you go about creating the landscape?
6	1	How would you design your class if you were the teacher? What subject would you teach? How would you structure your tests?
6	2	Decide on a party or event that you want to have and write details as to how you would plan this event. Write only about the planning you would do before the day of the event.

TABLE A.1: List of all Prompts along with their expected cognitive load and session number

Bibliography

Alves, Rui Alexandre, Sao Luis Castro, and Thierry Olive (2008). "Execution and pauses in writing narratives: Processing time, cognitive effort and typing skill". In: *International journal of psychology* 43.6, pp. 969–979.

Alves, Rui Alexandre et al. (2007). "Influence of typing skill on pause-execution cycles in written composition". In: *Studies in writing* 20, p. 55.

Anderson, Lorin W, David R Krathwohl, and Benjamin Samuel Bloom (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Allyn & Bacon.

Araujo, L.C.F. et al. (2005). "User Authentication Through Typing Biometrics Features". In: *Trans. Sig. Proc.* 53.2, pp. 851–855. ISSN: 1053-587X. DOI: 10.1109/TSP.2004.839903(410)53. URL: http://dx.doi.org/10.1109/TSP.2004.839903(410)53.

Ashbourn, Julian (2014). *Biometrics: Advanced identity verification: the complete guide*. Springer.

Baaijen, V. M., D. Galbraith, and K. de Glopper (2012). "Keystroke Analysis Reflections on Procedures and Measures". In: *Written Communication* 29.3, pp. 246–277.

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

Balagani, Kiran S (2013). *Investigating Cognitive Rhythms as a New Modality for Continuous Authentication*. Tech. rep. DTIC Document.

Baldridge, J. (2005). *The OpenNLP project*. www.opennlp.sourceforge.net.

Banerjee, Salil P and Damon L Woodard (2012). "Biometric authentication and identification using keystroke dynamics: A survey". In: *Journal of Pattern Recognition Research* 7.1, pp. 116–139.

Bartlow, N. and B. Cukic (2006). "Evaluating the Reliability of Credential Hardening through Keystroke Dynamics". In: *17th International Symposium on Software Reliability Engineering (ISSRE '06)*.

Bartmann, Dieter, Idir Bakdi, and Michael Achatz (2007). "On the design of an authentication system based on keystroke dynamics using a predefined input text". In: *International Journal of Information Security and stabilitycy* 1.2, p. 1.

Bergadano, F., D. Gunetti, and C. Picardi (2003). "Identity verification through dynamic keystroke analysis". In: *Intelligent Data Analysis* 7.5, pp. 469–496.

Bergsma, S., M. Post, and D. Yarowsky (2012). "Stylometric analysis of scientific articles". In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 327–337.

Canales, Omar et al. (2011). "A Stylometry System for Authenticating Students Taking Online Tests". In: *Proceedings of Student-Faculty Research Day, CSIS, Pace University*.

Choi, Yejin (2014). "Keystroke Patterns as Prosody in Digital Writings: A Case Study with Deceptive Reviews and Essays". In: Empirical Methods on Natural Language Processing (EMNLP).

Chung, Cindy and James W Pennebaker (2007). "The psychological functions of function words". In: *Social communication*, pp. 343–359.

Cohen Priva, Uriel (2010). "Constructing Typing-Time Corpora: A New Way to Answer Old Questions". In: *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*, pp. 43–48.

Coover, John E (1923). "A method of teaching typewriting based upon a psychological analysis of expert typing". In: *National Education Association* 61, pp. 561–567.

Covington, M. A. and J. D. McFall (2010). "Cutting the Gordian knot: The moving-average type–token ratio (MATTR)". In: *Journal of Quantatiative Linguistics* 17.2, pp. 94–100.

Curtin, Mary et al. (2006). "Keystroke biometric recognition on long-text input: A feasibility study". In: *Proc. Int. MultiConf. Engineers & Computer Scientists (IMECS)*.

Dahlmann, Irina 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*. Association for Computational Linguistics, pp. 49–56.

Darabseh, Alaa and Akbar Siami Namin (2014). "The accuracy of user authentication through keystroke features using the most frequent words". In: *Proceedings of the 9th Annual Cyber and Information Security Research Conference*. ACM, pp. 85–88.

Epp, C., M. Lippold, and R. Mandryk (2011a). "Identifying Emotional States Using Keystroke Dynamics". In: *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems (CHI 2011)*. Vancouver, BC, Canada, pp. 715–724.

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

Erman, Britt (2007). "Cognitive processes as evidence of the idiom principle". In: *International Journal of Corpus Linguistics* 12.1, pp. 25–53. URL: http://web.a.ebscohost.com/ehost/pdfviewer/pdfviewer?sid=b6807aa4-eef6-4f5e-b2f1-7f12880fbf18%40sessionmgr4004&vid=0&hid=4214.

Erman, Britt and Beatrice Warren (2000). "The idiom principle and the open choice principle". In: *TEXT-THE HAGUE THEN AMSTERDAM THEN BERLIN-* 20.1, pp. 29–62.

Fairhurst, M. and M. Da Costa-Abreu (2011). "Using Keystroke Dynamics for Gender Identification in Social Network Environment". In: *4th International Conference on Imaging for Crime Detection and Prevention (ICDP 2011)*.

Finlayson, Mark Alan 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*. Association for Computational Linguistics, pp. 20–24.

Fodor, Janet Dean (2002). "Prosodic disambiguation in silent reading". In: *PROCEEDINGS-NELS*. Vol. 1. 32; VOL 1, pp. 113–132.

Gentner, Donald R, Serge Larochelle, and Jonathan Grudin (1988). "Lexical, sublexical, and peripheral effects in skilled typewriting". In: *Cognitive Psychology* 20.4, pp. 524–548.

Giot, R. and C. Rosenberger (2012). "A New Soft Biometric Approach For Keystroke Dynamics Based On Gender Recognition". In: *International Journal of Information Technology and Management (IJITM)* 11.1/2, pp. 35–49.

Gleick, James (2012). *The Information: A History, A Theory, A Flood, 2011*. Londỳn: Fourth Estate.

Goldman-Eisler, Frieda (1958). "Speech production and the predictability of words in context". In: *Quarterly Journal of Experimental Psychology* 10.2, pp. 96–106.

Goswami, S. and S. Sarkar adn M. Rustagi (2009). "Stylometric Analysis of Bloggers' Age and Gender". In: *Proceedings of the Third International AAAI Conference on Weblogs and Social Media (ICWSM 2009)*.

Grieve, J. (2007). "Quantitative authorship attribution: An evaluation of techniques". In: *Literary and linguistic computing* 22.3, pp. 251–270.

Gunetti, D., C. Picardi, and G. Ruffo (2005). "Keystroke analysis of different languages: A case study". In: *Advances in Intelligent Data Analysis VI*. Springer Berlin / Heidelberg, pp. 133–144.

Gunetti, Daniele and Claudia Picardi (2005). "Keystroke Analysis of Free Text". In: *ACM Trans. Inf. Syst. Secur.* 8.3, pp. 312–347. ISSN: 1094-9224. DOI: 10.1145/1085126.1085129. URL: http://doi.acm.org/10.1145/1085126.1085129.

Hale, John (2006). "Uncertainty about the rest of the sentence". In: *Cognitive Science* 30.4, pp. 643–672.

Hall, M. et al. (2009). "The WEKA Data Mining Software: An Update". In: *SIGKDD Explorations* 11.1.

Hickey, Tina (1993). "Identifying formulas in first language acquisition". In: *Journal of Child Language* 20.01, pp. 27–41.

Hoeij, Maggy JW van et al. (2004). "Developing a classification tool based on Bloom's taxonomy to assess the cognitive level of short essay questions". In: *Journal of veterinary medical education* 31, pp. 261–267.

Holmes, David I (1985). "The analysis of literary style–a review". In: *Journal of the Royal Statistical Society. Series A (General)*, pp. 328–341.

Idrus, Syed Zulkarnain Syed et al. (2014). "Soft biometrics for keystroke dynamics: Profiling individuals while typing passwords". In: *Computers & Security* 45, pp. 147–155.

Idzelis, Mindaugas (2013). *Jazzy: The Java Open Source Spell Checker*. URL: http://jazzy.sourceforge.net.

Jain, Anil K, Arun A Ross, and Karthik Nandakumar (2011). *Introduction to biometrics*. Springer Science & Business Media.

Japkowicz, N. and S. Stephen (2002). "The class imbalance problem: A systematic study". In: *Intelligent Data Analysis* 6.5, pp. 429–450.

Johansson, Roger et al. (2010). "Looking at the keyboard or the monitor: relationship with text production processes". In: *Reading and writing* 23.7, pp. 835–851.

Joyce, Rick and Gopal Gupta (1990). "Identity authentication based on keystroke latencies". In: *Communications of the ACM* 33.2, pp. 168–176.

Juola, P. (2006). "Authorship Attribution". In: *Foundations and Trends in information Retrieval* 1.3, pp. 233–334.

Kellogg, Ronald T (1996). "A model of working memory in writing". In: *The science of writig*. Ed. by C.M. Levy and S.E. Ransdell. Lawrence Erlbaum Associates, Inc, pp. 57–71.

Killourhy, Kevin and Roy Maxion (2010). "Why Did My Detector Do That?!: Predicting Keystroke-dynamics Error Rates". In: *Proceedings of the 13th International Conference on Recent Advances in Intrusion Detection*. RAID'10. Ottawa, Ontario, Canada: Springer-Verlag, pp. 256–276. ISBN: 3-642-15511-1, 978-3-642-15511-6. URL: http://dl.acm.org/citation.cfm?id=1894166.1894184.

Killourhy, Kevin S and Roy A Maxion (2009a). "Comparing anomaly-detection algorithms for keystroke dynamics". In: *Dependable Systems & Networks, 2009. DSN'09. IEEE/IFIP International Conference on*. IEEE, pp. 125–134.

Killourhy, Kevin S. and Roy A. Maxion (2009b). "Comparing anomaly-detection algorithms for keystroke dynamics." In: *DSN*. IEEE, pp. 125–134.

Koppel, M., S. Argamon, and A. R. Shimoni (2002). "Automatically Categorizing Written Texts by Author Gender". In: *Literary and Linguistic Computing* 17.4, pp. 401–412.

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

Levy, Roger (2008). "Expectation-based syntactic comprehension". In: *Cognition* 106.3, pp. 1126–1177.

Locklear, Hilbert et al. (2014). "Continuous authentication with cognition-centric text production and revision features". In: *Biometrics (IJCB), 2014 IEEE International Joint Conference on*. IEEE, pp. 1–8.

Manning, Christopher D. et al. (2014). "The Stanford CoreNLP Natural Language Processing Toolkit". In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Baltimore, Maryland: Association for Computational Linguistics, pp. 55–60. URL: http://www.aclweb.org/anthology/P/P14/P14-5010.

Mihalcea, Rada (1998). "Semcor semantically tagged corpus". In: *Unpublished manuscript*.

Monrose, Fabian, Michael K Reiter, and Susanne Wetzel (2002). "Password hardening based on keystroke dynamics". In: *International Journal of Information Security* 1.2, pp. 69–83.

Monrose, Fabian and Aviel Rubin (1997). "Authentication via keystroke dynamics". In: *Proceedings of the 4th ACM conference on Computer and communications security*. ACM, pp. 48–56.

Monrose, Fabian and Aviel D Rubin (2000). "Keystroke dynamics as a biometric for authentication". In: *Future Generation computer systems* 16.4, pp. 351–359.

Montalvao, Jugurta, Carlos Augusto S Almeida, and Eduardo O Freire (2006). "Equalization of keystroke timing histograms for improved identification performance". In: *Telecommunications Symposium, 2006 International*. IEEE, pp. 560–565.

Mosteller, Frederick and David L Wallace (1964). *Applied Bayesian and classical inference: the case of the Federalist papers*. Addison-Wesley.

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

Paul, Richard W and AJA Binker (1990). *Critical thinking: What every person needs to survive in a rapidly changing world.* ERIC.

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

Riggenbach, Heidi (1991). "Toward an understanding of fluency: A microanalysis of nonnative speaker conversations". In: *Discourse processes* 14.4, pp. 423–441.

Rumelhart, David E and Donald A Norman (1982). "Simulating a skilled typist: A study of skilled cognitive-motor performance". In: *Cognitive Science* 6.1, pp. 1–36.

Saevanee, Hataichanok, Nathan L Clarke, and Steven M Furnell (2012). "Multi-modal behavioural biometric authentication for mobile devices". In: *Information Security and Privacy Research*. Springer, pp. 465–474.

Sag, Ivan A et al. (2002). "Multiword expressions: A pain in the neck for NLP". In: *Computational Linguistics and Intelligent Text Processing*. Springer, pp. 1–15.

Salthouse, T. A. (1986). "Perceptual, cognitive, and motoric aspects of transcription typing". In: *Psychological bulletin* 99.3, p. 303.

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

Shaffer, LH (1978). "Timing in the motor programming of typing". In: *The Quarterly Journal of Experimental Psychology* 30.2, pp. 333–345.

Sim, Terence and Rajkumar Janakiraman (2007). "Are digraphs good for free-text keystroke dynamics?" In: *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, pp. 1–6.

Smith, Nathaniel J and Roger Levy (2013). "The effect of word predictability on reading time is logarithmic". In: *Cognition* 128.3, pp. 302–319.

Snow, Rion et al. (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*. Association for Computational Linguistics, pp. 254–263.

Song, Dawn Xiaodong, David Wagner, and Xuqing Tian (2001). "Timing Analysis of Keystrokes and Timing Attacks on SSH." In: *USENIX Security Symposium*. Vol. 2001.

Stamatatos, E., N. Fakotakis, and G. Kokkinakis (2001). "Computer-based authorship attribution without lexical measures". In: *Computers and the Humanities* 35.2, pp. 193–214.

Stewart, John C et al. (2011). "An investigation of keystroke and stylometry traits for authenticating online test takers". In: *Biometrics (IJCB), 2011 International Joint Conference on*. IEEE, pp. 1–7.

Teh, Pin Shen, Andrew Beng Jin Teoh, and Shigang Yue (2013). "A survey of keystroke dynamics biometrics". In: *The Scientific World Journal* 2013.

Ure, Jean (1971). "Lexical density and register differentiation". In: *Applications of linguistics*, pp. 443–452.

Vel, O. de (2000). "Mining e-mail authorship". In: *Proc. Workshop on Text Mining, ACM International Conference on Knowledge Discovery and Data Mining (KDD'2000)*.

Vel, O. de et al. (2002). "Language and Gender Author Cohort Analysis of e-mail for Computer Forensics". In: *Proceedings Digital Forensics Research Workshop*. Syracuse, NY, USA.

Villani, M. et al. (2006). "Keystroke Biometric Recognition Studies on Long-Text Input under Ideal and Application-Oriented Conditions". In: *Computer Vision and Pattern Recognition Workshop*, pp. 17–22.

Vizer, L. M., L. Zhou, and A. Sears (2009). "Automated stress detection using keystroke and linguistic features: An exploratory study". In: *International Journal of Human-Computer Studies* 67.10, pp. 870–886.

Warren, Paul (2012). *Introducing psycholinguistics*. Cambridge University Press.

Wikipedia (2014). *Precision and recall — Wikipedia, The Free Encyclopedia*. [Online; accessed 10-February-2015]. URL: http://en.wikipedia.org/w/index.php?title=Precision_and_recall\&oldid=635850583.

Wray, Alison (2002). *Formulaic language and the lexicon*. Cambridge University Press.

Zheng, Nan, Aaron Paloski, and Haining Wang (2011). "An efficient user verification system via mouse movements". In: *Proceedings of the 18th ACM conference on Computer and communications security*. ACM, pp. 139–150.

Zhong, Yu, 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*. IEEE, pp. 117–123.