FAKULTA
PŘÍRODOVĚDNĚ-HUMANITNÍ
A PEDAGOGICKÁ TUL

Bakalářská práce

English Orthography in Retrospect

Studijní program: B0114A300068 Anglický jazyk se zaměřením na vzdělávání
Studijní obory: Anglický jazyk se zaměřením na vzdělávání
Německý jazyk se zaměřením na vzdělávání

Autor práce: **Dietrich Marchenko**
Vedoucí práce: Mgr. Jaromír Haupt, Ph.D.
Katedra anglického jazyka

Liberec 2023

FAKULTA
PŘÍRODOVĚDNĚ-HUMANITNÍ
A PEDAGOGICKÁ TUL

Zadání bakalářské práce

English Orthography in Retrospect

Jméno a příjmení:	**Dietrich Marchenko**
Osobní číslo:	P20000287
Studijní program:	B0114A300068 Anglický jazyk se zaměřením na vzdělávání
Specializace:	Anglický jazyk se zaměřením na vzdělávání Německý jazyk se zaměřením na vzdělávání
Zadávající katedra:	Katedra anglického jazyka
Akademický rok:	2021/2022

Zásady pro vypracování:

In the public eye and scientific world alike, English Orthography is known to be devoid of sound logic and simplicity. Centuries of vowel shifts and obsessive word borrowing have led to unpredictable changes in the way people read, write, and arguably think in English.
This thesis will consist of three parts, and will concern itself with English orthography in retrospect, how it influences the speakers around the world and how it can become more logical in the least destructive way.
The first part of the thesis will deal with the social aspects of a spelling reform. Chiefly about the role of English in todays world, how it shapes people's perception and linguistic abilities and how it is represented in the media.
The second part of the thesis constitutes work with scientific publications to better understand the logic and history of English spelling as a system, as well as a thorough analysis of the potential benefits and issues of a spelling reform, both for L2 and L1 speakers. This will be done by utilizing both corpora, linguistic handbooks and social network chatrooms to collect the needed data.
The third and last part of the thesis will consist of a measure to address the most troublesome parts of the orthography. This measure will be based on the research of linguists and scientists before the date of this publication, and will try to achieve the equilibrium between usefulness and backwards compatibility with the written language of the past.

Rozsah grafických prací:
Rozsah pracovní zprávy:
Forma zpracování práce: tištěná/elektronická
Jazyk práce: angličtina

Seznam odborné literatury:

Cook, V.J. ”L2 Users and English Spelling.” *Journal of Multilingual and Multicultural Development* 18, no. 6 (1997): 474–88. https://doi.org/10.1080/01434639708666335.
Scholes, Robert J., and Linnea C Ehri. ”How English Orthography Influences Phonological Knowledge as Children Learn to Read and Spell.” Essay. In *Literacy and Language Analysis*. New York: Routledge Taylor & Francis Group, 2016.
Seymour, Philip H., Mikko Aro, and Jane M. Erskine. ”Foundation Literacy Acquisition in European Orthographies.” *British Journal of Psychology* 94, no. 2 (2003): 143–74. https://doi.org/10.1348/000712603321661859.
Upward, Christopher, and George Davidson. *The History of English Spelling*. Malden, MA: Wiley-Blackwell, 2011.

Vedoucí práce: Mgr. Jaromír Haupt, Ph.D.
Katedra anglického jazyka

Datum zadání práce: 20. dubna 2022
Předpokládaný termín odevzdání: 26. dubna 2023

L.S.

prof. RNDr. Jan Picek, CSc.
děkan

Mgr. Zénó Vernyik, Ph.D.
vedoucí katedry

V Liberci dne 20. dubna 2022

Prohlášení

Prohlašuji, že svou bakalářskou práci jsem vypracoval samostatně jako původní dílo s použitím uvedené literatury a na základě konzultací s vedoucím mé bakalářské práce a konzultantem.

Jsem si vědom toho, že na mou bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména § 60 – školní dílo.

Beru na vědomí, že Technická univerzita v Liberci nezasahuje do mých autorských práv užitím mé bakalářské práce pro vnitřní potřebu Technické univerzity v Liberci.

Užiji-li bakalářskou práci nebo poskytnu-li licenci k jejímu využití, jsem si vědom povinnosti informovat o této skutečnosti Technickou univerzitu v Liberci; v tomto případě má Technická univerzita v Liberci právo ode mne požadovat úhradu nákladů, které vynaložila na vytvoření díla, až do jejich skutečné výše.

Současně čestně prohlašuji, že text elektronické podoby práce vložený do IS/STAG se shoduje s textem tištěné podoby práce.

Beru na vědomí, že má bakalářská práce bude zveřejněna Technickou univerzitou v Liberci v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších předpisů.

Jsem si vědom následků, které podle zákona o vysokých školách mohou vyplývat z porušení tohoto prohlášení.

14. července 2023 Dietrich Marchenko

Acknowledgements

I would like to first thank Dr. Haupt for his expert critiques, which have helped make this thesis take a superior shape, and Dr. Madsen for his initial guidance, which helped me gain a new perspective on certain lingustic topics. Then, I would like to thank my girlfriend for being around during these trying times.

Anotace

Tato práce podrobně zkoumá obtížnosti spojené s anglickým pravopisem a probíhající debatu ohledně jeho možné reformy. Dále se zabývá proveditelností a potenciálními dopady reformy anglického pravopisu. Byly provedeny dva samostatné praktické experimenty, které se snaží kvantifikovat a prokázat komplexnost anglického pravopisu, rovněž také analyzovat efektivitu preskriptivního a přirozeně se vyvíjejícího respellingu. Bylo zjištěno, že přirozený respelling, lépe definovaný jako přirozená evoluce hláskování v důsledku mnoha faktorů, by mohl mít slušný potenciál pro reformu pravopisu bez potřeby nařízené intervence. Je nutné zmínit, že během psaní práce bylo rozhodnuto o úpravě původního abstraktu práce, neboť pořadí, v němž výzkum probíhá, bylo po kritice vedoucích práce revidováno a vylepšeno.

Klíčová slova: anglický pravopis, reforma pravopisu, hloubka pravopisu, OCI, grafémy, fonémy, preskriptivismus v pravopisu, přirozené přepravopisování, korpus, analýza

Abstract

This thesis provides a comprehensive exploration of the complexities inherent in English orthography and the ongoing debate surrounding potential reform, scrutinizing the feasibility and potential impacts of English spelling reform. Two separate practical experiments have been conducted, attempting to quantify, and prove the complexity of the English spelling, as well as analyze the efficacy of prescriptive, and naturally occurring respelling. It has been concluded that natural respelling, better defined as a natural evolution of spelling, due to numerous factors, could possess decent potential to reform the orthography without the need for prescriptive intervention. It is necessary to point out that during the writing of the paper it was decided to alter the original abstract of the thesis, which was included in the STAG system, since the order in which the research takes place was revised and improved after the critiques of the thesis supervisors.

Keywords: English orthography, spelling reform, orthographic depth, OCI, graphemes, phonemes, spelling prescriptivism, natural respelling, corpus, analysis

Contents

List of tables

List of abbreviations

ASCII – *American Standard Code for Information*

EME – *Early Middle English*

ESL - *English Second Language*

GVS – *Great Vowel Shift*

IPA – *International Phonetic Alphabet*

L2 – *Second Language*

ME – *Middle English*

NHS – *National Health Service*

NR – *Natural Respelling*

OCI – *Orthographic Complexity Index*

OE – *Old English*

PR – *Prescriptive Respelling*

RAE – *Real Academia Española (Royal Spanish Academy)*

SS – *Standard Spelling*

Introduction

English orthography has been a subject of linguistic interest, and debate among scientists, educators, and language learners. The unique position of English as a global *lingua franca*, and the peculiar inconsistency of its spelling that, at first glance, is devoid of logic, has raised numerous questions about the impact of such orthography on speakers around the world, and whether a standardized English spelling reform should, or even *could* ever become implemented.

This thesis aims to analyze the issues of English spelling from the historical, social, and reformatory standpoints, and conclude, whether a spelling reform should ever be implemented, providing at the same time an alternative possible course of action based on the research done. The structure of this thesis is non-standard when compared to other bachelor's theses. The first theoretical part begins with a review of the most important definitions that would appear throughout the whole work. It then focuses on the problems of orthographic depth, and its impact on the learners, perusing some of the better-known studies, and academic opinions on the subject matter. Next, the thesis delves into English orthography as a linguistic system, with a specific emphasis on its historic development. It illustrates the chief driving forces behind the irregular spelling of English and proceeds to compare it to languages with both shallow and deep orthographies, Spanish and Danish. This concludes with the first experiment: a grapheme-to-phoneme relationship formula, which would allow for calculating the OCI (Orthographic Complexity Index), i.e the orthographic difficulty of any given phrase or sentence, which is then followed with a comparative calculation for English, as well as Spanish and Danish languages. With this, the experiment attempts to quantify the difficulty of English orthography, putting it in contrast with other systems. The second

theoretical part then proceeds to scrutinize the need for English spelling reform. It analyzes the historical attempts to amend the spelling, as well as assesses the potential benefits and negatives of spelling reform. The obstacles in the way of an English spelling reform—theoretical or practical—are also extensively studied. The chapter concludes with the second experiment, which constitutes a comparative measurement of the difficulty of various types of reformed spelling, be it prescriptively (Simplified Spelling Board), or naturally respelled (online usage of written language). The study utilizes the aforementioned OCI formula, custom corpora with approximately 1,000,000 lemmata to gather the unofficial spellings, as well as other formulas to provide relevant statistics that would aid in the better elaboration of the findings. The thesis concludes with a discussion, and some additional thoughts on the topic of English spelling reformation after having gone through the research.

Since one of the primary goals for this thesis was to scrutinize the concept of reforming the spelling as well as contrasting it with a natural way of respelling, newly coined terms below were provided in order to simplify the reuse of these concepts further. Therefore, *prescriptive respelling* refers in this thesis to a deliberate attempt—by a scientific body, institution, or a single entity—to amend the spelling as they themselves see fit, motivated by various reasons that are grounded on common logic, or their own presuppositions (*for instance, a proposition by the Simplified Spelling Board in 1906, which was never fully adopted).* On the other hand, *natural respelling* refers to a frequent naturally occurring deviation in spelling, i.e. misspelling that has become frequent enough for it to be used by a variety of people, and with the potential to become normalized for official use. This is especially true in the age of the Internet, where spelling prescriptivism might seem much less enforced compared to any other

medium. These respellings could be utilized in a wide variety of situations (*for instance*, *"boiz," "cuz," "ur," "wat"*).

1. Important definitions

1.1. Orthography, and spelling

Consulting the Merriam-Webster (2023) dictionary, it offers the following definitions: "The art of writing words with the proper letters according to standard usage," and "the representation of the sounds of a language by written or printed symbols," as well as "a part of language study that deals with letters and spelling."

Of course, orthography and spelling, while related, are distinct concepts. David Crystal (2003) supplies the following definitions: Orthography is a broader term that encompasses the "standardized system for writing a language." It includes "not only spelling, but also punctuation, capitalization, word breaks, emphasis, and certain typographical features" (4-5). Spelling, on the other hand, refers specifically to the "convention that determines how words are formed from individual letters and diacritics." In other words, spelling is a component of orthography, but not its entirety (5). *Both terms* (spelling and orthography) are going to be *used interchangeably* throughout the thesis, as the main concern of this paper lies in studying both the conventions of letter formations and their accordance with the phonemic content of English, as well as more broad aspects of written language.

Additionally, in his book "Spell it Out" (2012) David Crystal argues that orthography is essential for effective communication, as it provides consistency and clarity in written language (158), adding "we need an orthography to be predictable.

There has to be a systematic relationship between sounds and letters. In a perfectly phonetic spelling system, the relationship is one-to-one: each sound is represented by one letter, so that it can be easily written, and each letter is pronounced with one sound, so that it can be easily read" (9).

1.2. Dyslexia

According to Mayo Clinic (2023), dyslexia is a learning difficulty that can lead to problems with reading, writing, and spelling. NHS (2023) explains that dyslexia is a specific learning difficulty that affects certain abilities used for learning, such as reading and writing. Intelligence, however, is not affected by dyslexia. In studies conducted by Aro and Wimmer (2003), and Ziegler and Goswami (2005), links between dyslexia and difficulty reading were found (621-622, and 8-10).

1.3. Deep, and shallow orthographies

The distinction between shallow and deep orthographies is a central concept in the study of reading and writing systems. Shallow orthographies are characterized by a highly consistent correspondence between graphemes (written symbols) and phonemes (speech sounds), whereas deep orthographies exhibit more complex relationships between these elements (Seymour, Aro, and Erskine, 2003, 144).

Shallow orthographies, such as the Italian, Spanish or Finnish writing systems, have a more transparent and predictable mapping between graphemes and phonemes (Ziegler and Goswami, 2005, 9-10). In these systems, each letter consistently represents a specific phoneme, and each phoneme is typically represented by a single letter (Landerl, Wimmer, and Frith, 1997, 316). This one-to-one correspondence simplifies

the process of decoding, making it easier for readers to map written symbols onto their corresponding speech sounds (Perfetti, 2007, 358). On the other hand, deep orthographies, such as English and French, involve more complex and less predictable relationships between graphemes and phonemes (Seymour, Aro, and Erskine, 2003, 144-146). In these systems, a single grapheme can represent multiple phonemes and a phoneme can be represented by different graphemes. Moreover, the pronunciation of a grapheme may also be influenced by its surrounding context (Treiman, 1997, 327). This opacity in the grapheme-phoneme correspondence makes decoding words more challenging for readers, as they must rely on additional information to determine the correct pronunciation of written words (Ziegler and Goswami, 2005, 28-29).

1.4. Graphemes and phonemes

Graphemes and phonembes constitute *the fundamental units of writing and speech* respectively, the understanding of which is pivotal in the realm of linguistics. Graphemes represent the smallest units of a writing system (Scragg, 1974, 14). In English, for instance, letters such as **"a"**, **"b"**, or **"c"** are graphemes, as are combinations like **"sh"** and **"ch"** in "ship" and "chip", respectively. This is in line with the concept of an alphabet, where each symbol or group of symbols corresponds to a distinct sound or set of sounds (Coulmas, 2003, 61). Phonemes, on the other hand, are the smallest units of sound in a language that can distinguish one word from another. In English, the words "bat" and "pat" are differentiated solely by the initial phonemes **/b/** and **/p/**. It is crucial to note that phonemes *are abstract representations of speech sounds* and do not correspond directly to the physical sounds produced in speech (Gussenhoven & Jacobs, 2011, 11-15).

Thus, the interaction between *graphemes and phonemes* provides the backbone for understanding the structure, development, and complexity of different writing systems.

1.5. Spelling reforms

Merriam-Webster (2023) writes that these refer to "a movement to modify conventional spelling so as to lessen or remove the differences between the orthography and the pronunciation of words". Other sources, like the Etymological Dictionary of the English Language by Ernest Klein (1987) suggest that these are "the deliberate modification and standardization of a language's orthography to enhance consistency, simplicity, and ease of learning" (11). Additionally, spelling reform is an *amendment of the established spelling rules* of a language, with the aim of rendering the system more consistent, logical, or easier to learn. It can involve changes in individual word spelling, alphabetic letters, and overall orthographic patterns (Venezky, 1999, 12-14).

2. Impact of orthography on reading, and spelling acquisition

The cause for a slower acquisition of both spelling and reading skills have been aptly studied throughout the years, with numerous papers *supporting* the idea that orthographic depth does indeed slow down the aforementioned skill acquisition.

2.1. Overview of the related studies

Certain studies have concluded that the lack of consistency in spelling rules can make *it more difficult* for individuals with dyslexia or other learning disabilities to read

and write in English (Norton and Wolf, 2012, 447-448). Research also indicates that readers of shallow orthographies tend to rely more on *phonological decoding strategies*, as the transparent correspondence between graphemes and phonemes facilitates a rapid and efficient mapping process (Ziegler and Goswami, 2005, 556).

In contrast, readers of deep orthographies often employ a combination of phonological and morphological strategies, due to the inconsistent nature of grapheme-phoneme relationships in these systems (Casalis and Louis-Alexandre, 2000, 325). These differences in orthographic depth *have implications* for literacy development. Empirical studies have consistently shown that children learning to read in shallow orthographies tend to acquire reading skills *more quickly* and experience fewer difficulties in comparison to their counterparts learning to read in deep orthographies (Aro and Wimmer, 2003, 628-629).

However, it is important to note that the advantage of shallow orthographies in the early stages of reading development *may diminish* over time as readers become more proficient and develop compensatory strategies for dealing with the complexities of deep orthographies (Share, 2008, 587).

2.2. Examples of shallow and deep orthographies, and their comparison to English

To better illustrate the position of English in the context of other European languages and their orthography, two other languages were chosen as candidates for *deep*, and *shallow* orthography examples, *Danish* and *Spanish* respectively. Both

languages resemble English either on the lexical level (Spanish, with a Latin connection), or Danish (a Germanic language).

2.2.1. Spanish

Spanish orthography is a *largely shallow system*, with a consistent one-to-one correspondence between phonemes and graphemes (Harris, 2006, 12). Carreiras, Alvarez, and De Vega, (1993) talk about how the orthographic principles governing Spanish make it a relatively straightforward language for learners. The consistent letter-sound relationship greatly reduces the cognitive load for spelling and reading (32-33). Spanish employs the Latin alphabet with the addition of the letter "**ñ**" and uses diacritic accents on vowels in specific circumstances. Spanish orthography *underwent a reform* in the 18th century under the guidance of the Royal Spanish Academy (Real Academia Española, RAE) to align spelling more closely with pronunciation, leading to its present phonemic nature (RAE, 2010).

2.2.2. Danish

Danish orthography represents a *significant deviation* from the phonemic principle, aligning more towards the deep end of the orthographic spectrum, though not quite as much as English. Danish orthography employs a variant of the Latin alphabet, exhibiting extensive utilization of diacritics and digraphs. It is characterized by *silent letters, numerous vowel sounds, and a significant degree of irregularity* (Daugaard, et al., 2020, 10-13). Authors of this publication also describe the fact that several letters may represent the same sound in Danish, and a single letter might represent different sounds depending on its position or surrounding letters. It is worth noting that Danish *has not undergone a major orthographic reform* since the adoption of the modern alphabet in the 19th century.

The phoneme-grapheme correspondence in Danish is therefore intricate due to the disparity between spoken and written language, making it particularly challenging for some learners (Elbro, Nielsen, and Petersen, 1994, 174-175).

2.1.3. Comparison with English

English orthography poses considerable difficulty to non-native speakers and learners (Aro and Wimmer, 2003, 628-629). In sharp contrast, Spanish orthography exhibits an almost one-to-one correspondence between sounds and letters, a feature characterizing shallow orthographies (Carreiras, Alvarez, and De Vega, 1993, 32-34). The RAE has maintained a *strict* normative policy, resulting in a spelling system where *pronunciation reliably indicates spelling* (34). Danish, on the other hand, possesses a deeper orthography with less predictable sound-letter correspondences (Juul and Sigurdsson, 2005, 10-11). However, Danish orthography is *less irregular* than English *due to a historical spelling reform* in the 19th century which made the orthography somewhat more phonetic (Elbro, Nielsen, and Petersen, 1994, 109-112).

3. Causes of the perceived difficulty of English orthography

Some linguists are of the opinion that English Orthography has a *profound historical reason* for having a complicated orthography. Linguists like Noam Chomsky and Morris Halle (1968) argue that English orthography *is not a purely phonemic system*, as there are many cases where spelling does not correspond directly to pronunciation, but *an etymological system*, where the historical context plays an important part in the word's spelling (1-3). Steven Pinker (1994), a cognitive psychologist and linguist, in his book called "The Language Instinct" agrees with Noam Chomsky that orthography is a window into the human mind, and understanding the principles of orthography can provide insight into the cognitive processes involved in reading and writing (16). It was therefore crucial to study the historical *driving forces* behind the intricacies of English orthography.

3.1. The Germanic roots (OE) and the arrival of Latin script (OE – EME)

Various Germanic tribes, namely the Anglo-Saxons, brought their language to Britain in the 5th century, laying the foundation for Old English (Crystal, 2003, 11). Before the arrival of Christian missionaries, the Anglo-Saxons employed the *runic futhorc script* for writing Old English. This script consisted of a set of angular letters, *designed for carving on wood or stone*, and was used for inscriptions and short texts, such as memorial stones and amulets (Page, 1999, 8). However, the runic script was *ill-suited* for representing the full range of Old English phonology, and due to various reasons, its use was limited in scope compared to the Latin script that would soon supplant it (Hogg and Denison, 2006, 13-14).

The Latin script was *adapted to accommodate the distinct phonological features of Old English*, which included sounds not found in Latin (Page, 1999, 9). To represent these unique sounds, scribes modified the Latin alphabet by incorporating new characters and employing digraphs (Hogg and Denison, 2006, 14). One notable adaptation involved the introduction of two additional characters**, "þ"** (thorn) and **"ð"** (eth), to represent *the voiceless and voiced dental fricatives*, respectively (Fulk, 2012, 21-23). Unexpectedly, the later transformation of the Old English letter thorn **"þ"** to the letter **"Y"** in the printed text was a result of the introduction of movable type printing, has led to the emergence of the common **"ye"** as seen in phrases like "Ye Olde Curiositie Shoppe" (Anderson, 1969, 22). Another adaptation was the use of digraphs, such as **"cg"** for the velarized **[j]** sound, **"sc"** for the *voiceless palatal fricative* **[ʃ]**, and **"hw"** for the *voiceless labiovelar approximant* **[ʍ]** These digraphs allowed scribes to represent Old English sounds using combinations of existing Latin characters (Hogg and Denison, 2006, 15-16).

3.2. The Norman Conquest (EME)

The Norman Conquest of 1066 marked a significant turning point in the history of the English language, with profound effects on both its orthography and vocabulary. "The Norman Conquest made English for two centuries the language mainly of the lower classes while the nobles and those associated with them used French on almost all occasions" (Baugh and Cable, 1978, 4).

The invasion led by William the Conqueror brought about the *establishment of Norman French as the language of the ruling class and the church*, which influenced

English orthography and introduced an extensive amount of French loanwords into the English lexicon (Crystal, 2003, 35). For instance, the Old English **"cw"** combination was replaced with **"qu"** (as in "queen"), and the Old English **"sc"** became **"sh"** (as in "ship") (Fisher, 1977, 860-880). Fisher also writes how French orthographic conventions, such as the use of the digraph **"ch"** to represent the **[ʧ]** sound, were integrated into English spelling or the **"ou"** spelling for the **[u]** sound (in words like "hous" and "mous") which can also be traced back to the influence of Norman French (860-872).

The Norman Conquest also led to the adoption of numerous French loanwords, which dramatically expanded the English vocabulary (Millward and Hayes, 2011, 192-194). "The thousands of loanwords that poured into English after the Norman Conquest had an effect beyond that of merely adding new terms and synonyms to the language. They also provided the raw material for an intricate system of levels of vocabulary ranging from the colloquial through the formal, from the every day to the highly technical, from the general to the highly specialized." These loanwords were primarily related to *law, government, art, literature, and religion*, reflecting the domains in which the Normans wielded power and influence (Baugh and Cable, 1978, 110).

3.3. The Printing Press and Standardization (ME)

Before the invention of the printing press in the 15th century, English orthography was characterized by significant inconsistencies and a lack of standardization (Wright, 2000, 20). One of the primary reasons for the inconsistencies in pre-printing press English orthography was the presence of numerous regional dialects (Hogg, 1992, 292). However, the *introduction of the printing press* into

England by William Caxton in the 15th century, who is also credited with the standardization of English to a great extent, has *fundamentally influenced orthographic norms*. Upon the introduction of Caxton's printing press in 1476, a critical instrument for the dissemination of a standardized language became accessible (Millward and Hayes, 2011, 166). The printing press fostered the propagation of a *relatively consistent* written form of the English language. This process significantly shaped a unified written form, despite the multitude of spoken dialects across the regions (Crystal, 2003, 11).

A *striking* outcome of this incongruity between written and spoken English is the presence of silent letters and non-intuitive spellings in contemporary English. Words like "knight" and "gnaw," which pronounced their initial consonants during the Middle English period, no longer do so in today's English (Lass, 1984, 6). It is crucial to underline that this standardization occurred synchronously with a significant linguistic phenomenon, the Great Vowel Shift (Crystal, 2003, 12).

3.4. The Great Vowel Shift (EME – ME)

The GVS, a *major change* in the *phonetics of the English language*, occurred between the late 14th and early 18th centuries (Lass, 1984, 128). The *systematic shift in the pronunciation of long vowels* significantly impacted the development of the English language and has been the subject of extensive research in historical linguistics (Stockwell and Minkova, 2001, 80-81).

The GVS involved a *series of changes in the pronunciation of long vowels*, characterized by the raising and diphthongization of certain vowels (Lass, 1984, 129).

For instance, the **[i:]** sound, as in "bite," shifted to the modern English **[aɪ]** sound, while the **[e:]** sound, as in "bēte," shifted to the modern English **[i:]** sound. The Middle English **[u:]** sound, as in "hūs" (/huːs/), shifted to the modern English **[aʊ]** sound, as in "house" (/haʊs/) The **[o:]** sound, as in "bōt" (/boːt/), shifted to the modern English **[u:]** sound, as in "boot" (/buːt/) (Lass, 1984, 129-130). These and numerous other changes occurred gradually and affected all long vowels in the language, resulting in a reorganization of the entire English vowel system (Stockwell and Minkova, 2001, 81).

McMahon (1994) argues that the GVS had *profound* consequences for the English language, particularly in terms of phonetics. Because of its *temporal coincidence* with the introduction of the printing press, coupled with a drastic altering of the pronunciation of long vowels, the GVS created some *large discrepancies* between English orthography and pronunciation, which persist to this day (16). Additionally, the shift contributed to the development of new vowel distinctions and phonological patterns, which influenced the subsequent evolution of the English language and its dialects (Stockwell and Minkova, 2001, 82).

4. A comparative formulaic approach to prove English spelling difficulty

A natural goal was to devise a theoretical approach to quantify the purported difficulty of English spelling. Looking into the subject of quantifying orthographic depth, certain studies might come to attention, namely Schmalz, Marinus, Coltheart et al. (2015) where the linguists have taken into account such criteria of Dutch, English, French, German and Italian orthographies as "Total number of rules (DRC)", "Single-letter rules (DRC)", "Multi-letter rules (DRC)", yielding results for "Irregular words (%)" while accounting for "Parsing" and "Generalization" accuracy (%) (1).

4.1. Description

Objective: the formula described below capitalizes on the quantity of graphemic representation for each phoneme, and vice-versa, thus attempting to explain a significant aspect of a language's orthographic complexity while contrasting it with orthographies of Spanish and Danish in the same way

The general hypothesis: English orthography is purported to be more complex than other European orthographies (Spanish and Danish)

The operational hypothesis: English orthography dictates a wider variety of phonemes per a single graphemic representation than in other European languages. English spelling has more explicit exceptions.

4.2. Methodology

The approach denotes an arbitrarily called ***Orthographic Complexity Index (OCI)*** as a measurable unit of orthographic difficulty. The formula to calculate this index would then appear to be: $\mathbf{OCI = \sum(PG) + (GP) / 2N}$, where: **PG =** the number of different phonemes a single grapheme can represent in a given language variety, **GP =** the number of different graphemes a single phoneme can represent in the given language variety, **N =** the total number of phonemes in the text.

The thought process behind this formula encapsulates the fundamental intricacies of English spelling, characterized by a high degree of both grapheme-to-phoneme and phoneme-to-grapheme inconsistency. These two phenomena can significantly increase orthographic complexity and, consequently, the difficulty of reading and spelling acquisition (the more ways there is to read any given word, the more ambiguous it is to infer spelling to the reader). The sum of the total of these "ways" is then divided by N, the total number of phonemes, and multiplied by 2, which combined serves as a vital denominator in the formula, creating a proportional representation that normalizes the data.

4.2.1 Calculation procedure

The procedure for calculating the OCI value for a given string of text will be undertaken for each grapheme and phoneme of the word:

- Generate phonemic transcription of the string using the phonemic inventory of a standard dialect, sourcing it from Wikipedia
- Map each phoneme to its corresponding grapheme in the string.

- Calculate the number of different graphemes for each phoneme (GP) and vice versa (PG).
- Compute the final OCI.

Table 1: Illustrative calculations of PG and GP values for any given string

Target word	"The"
Grapheme	<t>
PG calculation	This grapheme represents the phoneme /t/ in Standard American English.
PG value	1
GP calculation	The phoneme /t/ could be represented by graphemes "t" (as in "top"), "tt" (as in "butter"), "ed" (as in "jumped"), and "th" (as in "Thomas") in Standard American English
GP value	4

4.2.2 Prerequisites

For the formula to be utmost precise, it is necessary to keep in mind the following:

- The variety of the analyzed language needs to be of a single entity, as for example American Standard English, Mexican Spanish and Standard Danish, with an established amount of phonemes
- The source for information on phonemes and graphemes of a given language and/or dialect was fetched from Wikipedia, or any other credible source.

4.2.3 Data collection

For this specific study, a *pangram of comparable length* (around 93 symbols) from each candidate language was taken into consideration. A pangram, according to Merriam-Webster, (2023) also known as a holoalphabetic sentence, is a "sentence that uses every letter of a given alphabet at least once" which has allowed for a wide variety of phonemes and their corresponding graphemes to be considered. This approach would also eliminate the need for studying lengthy strings or corpora to improve efficacy, as it accounts for a broader range of phoneme/grapheme relationships in the language. Additionally, to improve the precision of the formula, the length of the strings has to be of a comparable length, as the division by N times 2 normalizes the values only in a single-layer manner, and might not account for fringe comparisons (comparing a single word of three letters with a string of three hundred letters).

4.3. Analysis, results and their interpretations

In order to shorten the length of the thesis, the calculations for each given word were cached. The necessary steps to verify the information below were outlined in Chapter 5.2.1.

4.3.1. Spanish

Table 2: Calculations for the Spanish pangram of 97 characters

String	Benjamín pidió una bebida de kiwi y fresa. Noé, sin vergüenza, la más exquisita champaña del menú

IPA Transcription	[beɲaˈmin piˈðjo ˈuna beˈβiða de ˈkiwi i ˈfresa. noˈe, sin berˈɣwenθa, la ˈmas eksˈkisita tʃamˈpaɲa del ˈmenu]
Phonological specifics	Here, the **/ɣ/** symbol represents the voiced velar fricative, **/ɲ/** represents the palatal nasal consonant
N	77
∑(PG + GP)	78
Final score	**~0.51**

This rather low OCI for Spanish, a language known for its relatively shallow orthographic system, might reflect a higher degree of phoneme-grapheme correspondence. This result is consistent with the linguistic structure of Spanish, which, due to its historical evolution, relies on a relatively consistent phonemic orthography as per Carreiras, Alvarez, and De Vega (1993).

4.3.2. Danish

Table 3: Calculations for the Danish pangram of 87 characters

String	**Quizdeltagerne spiste jordbær med fløde, mens cirkusklovnen Walther spillede på xylofon**
IPA Transcription	/kvɪsdɛltɑgɐnə spistə joˌɐbɛˌɐ mɛ ðˌ fløðˌə, mɛns kɛʁkʰuskləvnən valtəʁ spiləðˌə pɔ ksyləfɔn/

Phonological specifics	Here, the **/ʁ/** is a uvular approximant, **/ð̩/** is a voiced dental fricative, **/kʰ/** is an aspirated voiceless velar plosive, and **/ɐ̯/** represents a near-open central vowel
N	79
∑(PG + GP)	194
Final score	**~1.23**

With this OCI, Danish might appear to possess a rather deep orthographic system, matching the scientific description by the words of which it encompasses a significant number of grapheme-phoneme combinations, reflecting the influence of historical spelling conventions and the effects of phonetic changes that have not been represented in the orthography, as per Elbro, Nielsen, and Petersen, (1994).

4.3.3. English

Table 4: Calculations for the English pangram of 79 characters

String	The quick brown fox jumps over a lazy dog, while I am in the process of writing
IPA Transcription	[ðə kwɪk braʊn fɑks dʒʌmps oʊvər ə leɪzi dɔg, waɪl aɪ æm ɪn ðə prɑsɛs ɑv ˈraɪtɪŋ]
N	61
∑(PG + GP)	226

Final score	**~1.85**

This could be interpreted in different ways. To begin, the findings presented thus far suggest a rather high Orthographic Complexity Index (OCI) for English. This figure, considering the limitations described in the next chapter, represents the substantial degree of inconsistency between written and spoken English. An OCI of 1.85 signifies that, on average, nearly two phoneme-grapheme combinations exist for each phoneme present in the sentence under consideration. This multiplicity might underline the inherent ambiguity of English orthography, with a single grapheme potentially representing multiple phonemes and vice versa, as per Crystal (2012). This complexity also relates to the non-phonemic nature of English orthography, which tends to prioritize morphological and etymological consistency over phonemic representation. This approach to orthography results in a system that preserves the history and relatedness of words, concurring with the conclusions of Chomsky and Halle (1968) but is not necessarily intuitive to spellers or learners, particularly those who are non-native speakers or those with reading difficulties such as dyslexia, as per Ziegler and Goswami (2005).

4.4. Approach limitations

It is necessary to address certain inherent limitations within the Orthographic Complexity Index (OCI) approach, highlighting the uncertainties involved. Naturally, it is advised to see the results as an illustrative proof of work for the concepts described in the theoretical parts of the thesis. Firstly, precision in the calculations remains a contentious issue, directly hinged upon the accuracy and sufficiency of the phonemic inventory of the dialect under study. Inadequate or erroneous phonemic information

may compromise the reliability of the OCI computation, leading to misrepresentations of a language's orthographic depth. Secondly, the assumption underlying the OCI approach posits an equal difficulty level across all phoneme-grapheme correspondences, an oversimplification that may not hold true in practice. Phoneme-grapheme correspondences can vary significantly in their degree of complexity and the cognitive effort required for their mastery, nuances which the OCI, in its current form, does not account for. Finally, the computational labor involved in calculating the OCI for text strings exceeding 100 symbols is formidable. Given the scale and complexity of this task, the execution and analysis of such an expansive data set lie beyond the purview of this thesis. Thus, the OCI, while a valuable measure, carries inherent limitations that temper its applicability and precision in orthographic complexity analysis.

4.5. Conclusions

The OCI in itself, being a novel approach, might offer a useful metric for measuring and comparing orthographic depth across languages. Through this lens, the idiosyncrasies of English, Spanish, and Danish orthographies become readily apparent, highlighting the considerable variations in phoneme-grapheme correspondences across these languages. The comparative analysis of the Orthographic Complexity Index (OCI) across English, Spanish, and Danish might offer insightful revelations about the idiosyncrasies of these languages" orthographic systems. The calculated OCI for Spanish stands at 0.51, a significantly lower figure compared to Englishes 1.85. Conversely, Danish presents an OCI of 1.23, closer to English, but still considerably lower. The calculations and subsequent interpretations could demonstrate that English, with its high OCI, might exhibit a deeper orthographic system compared to Spanish and

Danish. A higher OCI signifies a higher level of orthographic complexity, which can pose learning and literacy challenges discussed in the previous chapters.

5. Case for English spelling reformation

This chapter serves as a review of what constitutes the purported benefits of spelling reforms, what are some successful examples of those, and whether it could be deemed viable to reform English spelling in the context of today.

5.1. Aspects and purported benefits of spelling reform

As per Carney (1994), benefits could be *multi-faceted*. Firstly, a spelling reform can *reduce* the cognitive load on learners, especially for languages with complex orthographic systems where the pronunciation of a word cannot be reliably predicted from its spelling and vice versa (57-58). A reformed spelling system, which is more phonetic and has a one-to-one correspondence between sounds and letters, can potentially *ease* the learning process for both native speakers and second-language learners (62).

Spelling reforms can also play a *vital role* in promoting literacy and *improving* the overall efficiency of a language. By simplifying orthography and increasing the consistency of spelling-to-sound correspondences, reforms can make it easier for learners to acquire reading and writing skills (Seymour et al., 2003, 2). Reforms can also improve the *consistency of spelling-to-sound* correspondences, which may facilitate language processing and communication (Ziegler and Goswami, 2005, 2).

The notion that spelling reform could promote literacy is *supported* by empirical research. One such instance is the German spelling reform of 1996, which endeavored

to simplify the spelling system and rectify inconsistencies. Notably, Röber-Siekmeyer (2001) found that *reformed* German spelling was associated with *improved reading speeds* among students, signifying that spelling reform could potentially facilitate quicker acquisition of reading skills (2). Similarly, the Turkish language reform of the 1920s and 1930s which was aimed at replacing Arabic script with the Latin alphabet, in a bid to increase literacy rates in the country *has largely succeeded*. According to the official statistics, this dramatic shift indeed resulted in a *significant improvement in literacy rates,* from 9% in the 1920s to around 90% by the end of the 20th century (Lewis, 2002, 16).

5.1.1. Examples of successful spelling reforms

The German Orthographic Conference of 1991 culminated in a systematic spelling reform in the German language aimed at simplifying and standardizing its orthography. The reformed guidelines marked a shift from etymological to phonetic spelling principles in German. The reform initially met with some opposition but was ultimately widely adopted, substantially enhancing the uniformity of German orthography (Russ, 2002, 105-106).

In the Turkish Alphabet Reform of 1928, as part of Mustafa Kemal Atatürk's modernization campaign, the Ottoman Turkish script, which was Arabic-based, was supplanted by the Latin-based Turkish alphabet. This reform had significant implications for literacy, rendering the language easier to learn and teach, and facilitating a dramatic increase in literacy rates (Lewis, 2002, 97).

5.1.2. Examples of unsuccessful and partly successful spelling reforms

The French Spelling Reform, a proposed alteration to the French language in the 1990s, which suggested changes to approximately 2000 words, was largely unadopted due to resistance from the public and educational institutions. The implementation of the reform has been slow and inconsistent, with numerous French speakers and institutions opting not to use the proposed new spellings (Jouannin, 2008, 14-16).

In the English language, during the 18th and 19th centuries, some efforts were made to standardize the spelling, with proposals by individuals like Benjamin Franklin and Noah Webster. Franklin, for instance, developed a reformed English spelling system, which was published posthumously in "A Scheme for a new Alphabet and a Reformed Mode of Spelling" (Franklin, 1779). This *was not widely accepted*, however, and is viewed by many as an unsuccessful attempt at spelling reform. Webster had greater success with his American English dictionary, implementing some spelling simplifications such as removing the **"u"** in words like "colour" to become "color", but many of his other proposed changes *did not gain* widespread acceptance (Crystal, 2012, 14-16). Then, in the 20th century, the Simplified Spelling Board was created in the United States with the aim of simplifying the spelling of English words. The board, financially backed by Andrew Carnegie, proposed a list *of 300 reformed spellings* in 1906 (Simplified Spelling Board, 1906). However, despite the initial enthusiasm, this reform was also *largely unsuccessful* due to widespread public resistance and the complexity of implementing the changes (Crystal, 2012, 16).

5.2. Obstacles of spelling reformation in English

Numerous factors inherent to the English language and its position in the world might render any attempt to amend its spelling *not feasible*. The following chapter attempts to illustrate this point in depth while providing these factors, and their influence.

5.2.1 Obstacle: Diversity of English and its phonetics

The diversity of the language can be a *major hurdle* in its standardization, as there would be less and less common ground to build a new standard on. English, as per David Crystal (2012) is now an official language in *over 50* countries, with *approximately 1.5 billion* speakers encompassing both native (L1) and non-native speakers (L2 or ESL). Coupled with the overarching usage of English by L2 and ESL speakers in countries such as India, Nigeria, the Philippines and others, English has been made into a de-facto default language for international and intercultural use (1). According to Kachru (1992) "English becomes a global medium with local identities and messages" and as a global *lingua franca*, English has transcended geographical boundaries, resulting in the emergence of distinct dialects and variations across different regions of the world. These variations, driven by sociopolitical, historical, and cultural factors, have given rise to unique forms of English that are integral to the identity of their respective communities (3).

The diversity of English phonetics is also a testament to the language's richness and complexity. These occur due to various reasons, and lead to *very different* accents and dialects. Some of the most notable dialects of L1 English speakers might include Southern American Accent, Australian accent, New Zealand accent, and various British

accents. Moreso the L2 accents, including Indian accents, Nigerian accents, Singapore accents and so on. (Wells, 1982, 144-156). These variables can alter phonetic representations of a single word, leading to *a multiplicity* of possible transcriptions. A classic example is the word "schedule," pronounced as /ˈʃɛdjuːl/ in British English and /ˈskɛdʒuːl/ in American English (Roach, 2009, 16-21).

5.2.2. Obstacle: Absence of a centralized regulatory institution

The English language presents *a distinctive case* within the linguistic world, primarily due to the lack of a formal regulatory institution governing its use, evolution, and orthography. This distinctiveness inherently affects the language, as Venezky (1999) elucidates in "The American Way of Spelling: The Structure and Origins of American English Orthography." Unlike languages such as French, which have academies like *Académie Française*, or Spanish, with *Real Academia Española*, that enforce guidelines and rules for spelling and grammatical structures, English lacks such a governing body. This absence *might have lead* to a distinct flexibility and variability in English, distinguishing it from other languages with more centralized regulation (Venezky, 1999, 13). In contrast, languages such as German and Turkish, where spelling reforms have been instituted by a centralized authority, present *a different* picture.

5.2.3. Obstacle: Unsatisfactory relationship between reform efficacy, and its societal acceptance

A phenomenon of innate human resistance to change, especially one that is sudden and radical, which has been *extensively* studied across various fields, such as psychology and organizational studies. The resistance is often compared to the

anecdotal experiment *of slowly boiling a frog*, which purports that if a frog is placed in boiling water, it will immediately jump out, but if it is placed in cold water that is slowly heated, it will not perceive the danger and will be boiled alive. The analogy illustrates how individuals and organizations *may fail* to recognize gradual, yet significant changes until it is too late (Carlson and Sherk, 2002, 2-3). Behavioral and cognitive biases are the primary explanations for this phenomenon. The Status Quo Bias, a cognitive bias that prefers current states over changes (Samuelson and Zeckhauser, 1988, 517-518), and the Endowment Effect, where people place more value on things because they own them (Kahneman, Knetsch and Thaler, 1991, 193), make people *resistant* to change. Further, *sudden and radical* changes can induce psychological discomfort and fear of the unknown, thereby fueling resistance (Oreg, Vakola and Armenakis, 2011, 14).

However, it is crucial to note that the "frog metaphor" has been *widely debunked*. Zoologists have proven that frogs will indeed attempt to escape as the water heats up (Drummond, 2014, 3-5). Despite this, the analogy continues to hold metaphorical value in illustrating human behavior concerning *gradual and radical* changes.

6. Providing an alternative to prescriptive respelling, and comparing those with OCI

As argued in the previous chapter, a highly efficient reform—an example of a perfectly efficient reform would be the one to achieve a total 1:1 correspondence between spelling and the International Phonetic Alphabet (IPA) transcription of a given dialect—would face formidable opposition, substantially diminishing its chances of implementation. Conversely, a less efficient reform—perhaps one that merely amends problematic spellings with minimal interventions—could garner less resistance, thereby enhancing the likelihood of its realization. Building on this thought, it was only fair to study the natural respelling examples more closely, in order to shine light on their probable efficacy in decreasing the difficulty of English orthography.

6.1. Description

Objective: attempt to compare the orthographic difficulty using the OCI (Orthographic Complexity Index) of words extracted from corpora (natural respelling), to those taken from linguistic literature (prescripted respelling) and those exemplary of standard American spelling.

The general hypothesis: Prescriptive, rather than natural respelling should yield better results because it would be based on a linguistic framework, and should, in theory, get widespread acceptance.

The operational hypothesis: Spelling reforms, i.e. prescriptive respelling, come from the standpoint of linguistics, ergo their efficacy and their substantiation are

reasonable in nature. Misspelling, ie. natural respelling is erroneous and damages the language.

6.2. Methodology

Three (3) sets of ten (10) identical in meaning, but different in spelling words were studied, and compared. The first set was exemplary of natural respelling, the second of prescriptive respelling, and the third of standard American spelling. The source for the first set was an online chatroom corpus of ~1,000,000 lemmata. The source for the second set was "Handbook of Simplified Spelling" by the Simplified Spelling Board. The final tables contain information on the purported orthographic difficulty for each word, which was measured utilizing the OCI formula from the approach outlined in Chapter 5. Additionally, each table contains information about the usage frequency in the studied corpus, as well as percentage differences between their relative OCIs, and the percentage of usage per total cases, if applicable. For simplicity reasons, the "total use cases" value was represented by the sum of the frequency values for each word.

6.2.1. Calculation procedure

The procedure to calculate values was identical to the procedure outlined in Chapter 5, with two exceptions:

- The calculation procedure was done for single-word strings only
- The final tables includes the percentage change between the different OCI values, which was calculated with a standard formula: $(V2 - V1) / V1 \times 100$, where $V1$ = the first value, $V2$ = second value

- The final table also includes the percentage of usage per total cases. These percentage values were calculated with a standard formula $|V1 / V2| \times 100$, where $V1$ = the value of a given word, $V2$ = total cases.

In order to shorten the length of the thesis, the calculations for each given word were cached. Below is an illustration of calculations done for each word pair.

Table 5: Illustrative calculations for the word “dogs”

Target word with standard spelling	"Dogs"
N	**4**
PG calculation	[d] = 1, [o] = 15, [g] = 2, [s] = 7, in Standard American English.
PG value	**25**
GP calculation	[d] = 1, [o] = 7, [g] = 2, [s] = 2, in Standard American English
GP value	**12**
Total OCI	**~9.25**

Table 6: Illustrative calculations for the word “dogz”

Target word with reformed spelling	"Dogz"

N	**4**
PG calculation	[d] = 1, [o] = 15, [g] = 2, [z] = 1 in Standard American English.
PG value	**19**
GP calculation	[d] = 1, [o] = 7, [g] = 2, [z] = 1, in Standard American English
GP value	**11**
Total OCI	**~7.5**

6.2.2. Prerequisites

For the OCI formula, identical prerequisites to Chapter 5 need to be taken into consideration.

6.2.3. Data collection

For the first array of 10 *"natural respelling"* candidate words, an online chat-room corpus of approximately 1,000,000 lemmata was processed. These words are frequent misspellings, which might be either intentional (significant reduction of graphemes in the most logical way) or not intentional (using a single different grapheme for the phonemic representation). The source for these words was general-purpose chat rooms on the messenger Telegram. This has facilitated the data collection process, which came down to exporting the messaging history as a machine-readable *.json file, which was then processed with a custom Python script that would remove strings of text that would contain less than 8 words in them (to exclude short, nonsensical messages). The script would then also remove any non-ASCII characters, as well as

most of the characters that do not belong to English orthography. The processed strings were then transferred to Sketch Engine, where they were tokenized, becoming available for further analysis. With the help of Sketch Engine's API, the data extracted from these corpora have been then used to create a list of words for further research. The following were considered upon creating the list:

- The Keyword API in Sketch Engine first ensured that the words that were selected are of non-standard spelling.
- The word added to the primary list was neither an abbreviation nor an acronym.
- The total number of words in the primary list was 50.
- The frequency of the word's occurrence clearly signified the non-accidental nature. Any words that were encountered to have **<5%** of the frequency of their standard counterpart were discarded (i.e if the word "dogs" has only occurred 100 times, while its counterpart "doga" 3 times, which is only **3%** of the standard frequency, the word "doga" would be discarded, since its usage would be considered accidental (typo), as well as illogical)
- After the selection, 10 final candidates were fetched in random order.

For the second array of 10 "prescriptive respelling" words, the task was facilitated by the "Handbook of Simplified Spelling," an openly available source of suggestions that were devised by linguists with a clear intent of fixing certain aspects of English spelling. These suggestions included lists of example reformed words, as well as general guidelines for spelling and respelling certain phonemes. These guidelines and lists were deployed to create a matching pairing for the naturally respelled word, which served as a sure way to contrast these suggestions, comparing their OCI scores, and if applicable their frequencies, drawing conclusions from that.

Lastly, the source for the standard American spelling was Mariam-Webster Online Dictionary.

6.3. Analysis, results and their interpretations

6.3.1. Set "though" vs. "tho" and "tho"

As apparent from the table, the word "though" shares its reformed spelling with the "Handbook of Simplified" counterpart. Together, they share an OCI of 1.67, which might signify an approximately 69% decrease in orthographic difficulty, a rather substantial difference when compared to the baseline value of the word in standard spelling. Additionally, the frequency of the reformed spelling in the studied corpus shows that users might utilize a reformed spelling in ~34% of total cases, which is a rather high value for a non-standard spelling.

Table 7: Lemma “though”

	Word	**OCI**	**OCI difference**	**Frequency (pmw)**	**Percent per total use cases**
S.S.	though	5.5	baseline	1,546	~65,43%
P.R.	tho	1.67	-69.63%	817	~34.57%
N.R.	tho				

6.3.2. Set "through" vs. "thru" and "thru"

Together with the lemma "though," "through" also shares its reformed spelling with the "Handbook of Simplified" counterpart. Together, they share an OCI of 1.625,

which, when compared to an OCI of 4.625 of the standard American spelling might signify a substantial ~65% decrease in orthographic difficulty. Additionally, the frequency of the reformed spelling in the studied corpus shows that users might utilize a reformed spelling in ~12% of total cases.

Table 8: Lemma “through”

	Word	**OCI**	**OCI difference**	**Frequency (pmw)**	**Percent per total use cases**
S.S.	through	4.625	baseline	535	~88,00%
P.R.	thru	1.625	-64.86%	73	~12,00%
N.R.	thru				

6.3.3. Set "you" vs. "yu" and "u"

This particular set is evident of the repeating trend of the below-threshold usage frequency of prescriptively respelled words, "yu" in this case. However, the usage of a naturally reformed counterpart for "you" seems to appear approximately every 6th time per total use cases, or ~15%, potentially signifying both a moderate to substantial decrease in orthographic difficulty, and decent adoption rate. The mode of employment for the naturally reformed spelling "u" might be of a more relaxed, nonchalant nature.

Table 9: Lemma “you”

	Word	**OCI**	**OCI difference**	**Frequency (pmw)**	**Percent per total use cases**
S.S.	you	3.67	baseline	14,770	~85.55%
P.R.	yu	2.5	-31.88%	negligible	negligible
N.R.	u	1.5	-59.12%	2,701	~15.45%

6.3.4. Set "your" vs. "yur" and "ur"

The prescriptively reformed "yur" achieves a lower OCI of 2.17, potentially representing an estimated reduction of 30.56% in orthographic complexity compared to the baseline. The naturally reformed spelling "ur" exhibits an even further reduction in OCI, standing at 1.3, indicating a decrease of 58.4% in orthographic complexity from the original. From the frequency standpoint, the prescriptive spelling reform "yur" again demonstrates negligible presence in the studied corpus. Although constituting a minor percentage of ~17.4%, this incidence rate is not insignificant, given the non-standard nature of the spelling. Furthermore, the use of "ur" may hint towards a more casual, relaxed mode of communication, similar to Set 3.

Table 10: Lemma “your”

	Word	**OCI**	**OCI difference**	**Frequency (pmw)**	**Percent per total use cases**
S.S.	your	3.125	baseline	2,937	~82.60%
P.R.	yur	2.17	-30.56%	negligible	negligible
N.R.	ur	1.3	-58.40%	619	~17.40%

6.3.5. Set "release" vs. "releas" and "relese"

As measured by the OCI, the prescriptively reformed version "releas" exhibits a lower OCI value of 1.83, reflecting a moderate decrease of 26.79% in orthographic complexity from the baseline, while the naturally reformed version "relese" achieves an OCI of 1.34, signifying an even greater reduction of 46.4% in orthographic complexity relative to the standard spelling. The prescriptive reform "releas" has an

incidence that can be classified as negligible. Conversely, "relese", while showing a low frequency in usage (9.57%), indicates a non-trivial, although likely idiosyncratic presence in the corpus.

Table 11: Lemma “release”

	Word	OCI	OCI difference	Frequency (pmw)	Percent per total use cases
S.S.	release	2.5	baseline	255	~90.43%
P.R.	releas	1.83	-26.79%	negligible	negligible
N.R.	relese	1.34	-46.40%	27	~9.57%

6.3.6. Set "weird" vs. "wierd" and "wierd"

Both prescriptively and naturally reformed words possess an OCI that is 14.28% lower than that of the baseline, potentially signifying a low to moderate complexity decrease. With a total use case frequency of approximately 12%, or every 8th use, these reformed versions might be indicative of the potentially idiosyncratic nature of use.

Table 12: Lemma “weird”

	Word	OCI	OCI difference	Frequency (pmw)	Percent per total use cases
S.S.	weird	2.1	baseline	303	~89.09%
P.R.	wierd	1.8	-14.28%	41	~11.91%
N.R.	wierd				

6.3.7. Set "because" vs. "becaus" and "cuz"

This set is once again indicative of the negligible incidence rate of a prescriptively reformed word, coupled with a rather high adoption rate of the naturally respelled word. The latter also possessed a substantially lower OCI of approximately ~65%, while being preferred by 18% of usage cases in the corpus, which could hint at a low to moderate adoption rate. The style of employment of such a spelling however might probably also be considered relaxed and nonchalant.

Notwithstanding, it is necessary to mention that this comparison is not ideal, since the reformed version "cuz" seems to be a contraction, stemming from the root "cause" and not "because." However, the table was left as it is, given that there appears to be no naturally respelled version of "because" where the prefix "-be" would not be discarded.

Table 13: Lemma “because”

	Word	**OCI**	**OCI difference**	**Frequency (pmw)**	**Percent per total use cases**
S.S.	because	4.625	baseline	2,181	~81.94%
P.R.	becaus	2.3	-50.27%	negligible	negligible
N.R.	cuz	1.62	-64.97%	481	~18.06%

6.3.8. Set "uncanny" vs. "uncanney"

As apparent from the table, the standard spelling of the word "uncanny" seems to have no counterpart in the reformed spelling. However, the OCI difference between the spellings might appear interesting, as the respelled version "uncanney" has an extra letter "Y" added as an ending, which, naturally, might increase the orthographic complexity. This could be explained by hypercorrectness, perhaps influenced by the

words like "valley," "money" etc. All things considered, the total frequency could hint at a high idiosyncratic nature of the reformed spelling

Table 14: Lemma “uncanny”

	Word	OCI	OCI difference	Frequency (pmw)	Percent per total use cases
S.S.	uncanny	1.5	baseline	16	~85.00%
P.R.	-	-	-	-	-
N.R.	uncanney	1.66	+10.66%	3	~15.00%

6.3.9. Set "some" vs. "som" and "sum"

The prescriptively reformed spelling "releas" registers a diminished OCI at 1.5, marking a reduction of 35.62% in orthographic complexity when juxtaposed with the baseline. The naturally reformed spelling "relese" indicates an even further reduction, its OCI standing at 1.2, reflecting a 48.49% decrease in orthographic complexity from the original term. Upon investigating usage frequency, the prescribed reform "releas" returns a negligible incidence rate in the studied corpus, pointing to its underutilization, while "relese" emerges every 25th time per total use cases, contributing to a relatively minor portion of ~4.07%. At first glance, this could be indicative of a very high idiosyncratic nature, but considering a rather high number of uses throughout the corpus (115 per million words,) the usage "sum," might be considered a fringe spelling used for joking, irony, or other methods of communication, similar to "u," and "ur."

Table 15: Lemma “some”

	Word	OCI	OCI difference	Frequency (pmw)	Percent per total use cases

S.S.	some	2.33	baseline	2,709	~95.93%
P.R.	som	1.5	-35.62%	negligible	negligible
N.R.	sum	1.2	-48.49%	115	~4.07%

6.3.10. Set "sucks" vs. "sucks" and "sux"

Once again, this set is evident of nonexistent prescriptively respelled alternative, whereas a naturally respelled alternative, albeit of a rather low adoption rate of approximately every 10th use case, is being used throughout the studied corpus. This spelling might be very likely an example of the same fringe spelling, designed for joking, irony, or other communication purposes.

Table 16: Lemma “sucks”

	Word	**OCI**	**OCI difference**	**Frequency (pmw)**	**Percent per total use cases**
S.S.	sucks	2.125	baseline	163	~90.06%
P.R.	-	-	-	-	-
N.R.	sux	1.67	-21.41%	18	~9.94%

6.4. Limitations of this study

Certain limitations may have affected the quality and precision of the information provided, therefore it is advised to see the results as an illustrative proof of concept for the points described in the theoretical parts of the thesis. To begin, one significant constraint was the limitations of the Sketch Engine software, in particular its maximum allowance of 1,000,000 lemmata per account. Despite this limitation, however, this corpus size was more than adequate for the illustrative purposes of the

paper. Another challenge lay in attributing misspellings to certain social group idiosyncrasies. Establishing such a connection with certainty proved to be a daunting task, as it was inherently difficult to definitively prove that these orthographic discrepancies were not mere singularities of a specific social group, therefore a mention of the potential for that was added to the interpretations. Another noteworthy caveat concerned the underlying assumption of the calculation methodology. This research assumed that all phoneme-grapheme correspondences carried an equal degree of learning difficulty, a supposition that might not have held true in reality, given the nuanced complexity of language acquisition and use. Finally, the possibility of human error in the calculations had to be acknowledged. Despite careful consideration and scrutiny, there remained the inherent fallibility that came with human involvement.

6.5. Conclusions

The study has revealed several salient points in the discourse of English spelling reform. It is evident from the various sets studied that the orthographic complexity index (OCI) of both naturally and prescriptively respelled words tends to be lower than that of the standard American spelling, indicating a possible decrease in orthographic difficulty. However, a noteworthy discrepancy exists between the adoption rates of naturally and prescriptively respelled alternatives, with the latter consistently showing negligible presence in the studied corpus. The naturally respelled alternatives, on the other hand, show varying degrees of adoption, suggesting a range of applications from casual, relaxed communication to potentially fringe or idiosyncratic uses, such as for humor, irony, or other stylistic purposes. These findings illuminate the complex

dynamics of spelling reform in English, shedding light on a largely new concept of natural respelling.

Discussion

Undoubtedly, English orthography's unpredictability and irregularities stem from a complex confluence of historical influences. To redress these inconsistencies with meaningful impact, a wholesale renovation of the spelling system appears necessary, rather than a mere amendment of the most troublesome areas. However, this proposal is bound to encounter insurmountable obstacles. The scope of such an undertaking—requiring widespread changes in a globally-used, multifaceted language—invites considerable resistance and imposes implementation challenges of considerable magnitude.

As such, alternative paradigms of thought beckon consideration: allowing for greater variety in spelling (or *zero orthographic prescriptivism policy*). where only people's consistent failure to recall correct spellings, as evidenced in the second experiment of this study (analyzing a corpus drawn from online web chats), is the force to suggest potential candidates for respelling. Notable instances nowadays might include simplifications such as "thru" and "tho," a reduced double consonant in "accomodate," and the blurring of homophonic distinctions in "they're/their/there," "you're/your," "could of/could've," and "accept/except." Such an approach to spelling better reflects current usage patterns (given the ubiquity of the Internet, and its relaxed, quick-paced style of communication), might lower barriers to learning and usage, and may, theoretically, encounter less resistance than more radical orthographic alterations. However, the effectiveness, practicality, and wider acceptance of this new paradigm warrant further investigation and rigorous testing before any formal proposal for English spelling reform could be made.

References

ANDERSON, Donald. 1969. *The Art of Written Forms.*

ARO, Mikko, and WIMMER, Heinz. 2003. "Learning to Read: English in Comparison to Six More Regular Orthographies." Applied Psycholinguistics 24, no. 4: 621-635.

BAUGH, Albert C., and CABLE, Thomas. 1978. *A History of the English Language. 3rd ed.* Englewood Cliffs, NJ: Prentice-Hall.

CARLSON, Robert., and SHERK, James. 2002. *If You Don't See It Coming, You May Have Been Stuck in the Slow Lane.* The Journal of Business Strategy 23 (3): 36-42.

CARNEY, Edward. 1994. *A Survey of English Spelling.* London: Routledge.

CASALIS, Sébastien., and LOUIS-ALEXANDRE, Marie-France. 2000. *Morphological Analysis, Phonological Analysis and Learning to Read French: A Longitudinal Study.* Reading and Writing 13, no. 3: 303-335.

CHOMSKY, Noam. 1965. *Aspects of the Theory of Syntax.* Cambridge, MA: MIT Press.

CHOMSKY, Noam., and HALLE, Morris. 1968. *The Sound Pattern of English.* New York: Harper and Row.

CHOMSKY, Noam. 2006. *Language and Mind.* United Kingdom: Cambridge University Press.

COULMAS, Florian. 2003. Writing Systems: An Introduction to their Linguistic Analysis. Cambridge: Cambridge University Press.

CRYSTAL, David. 2003. *The Cambridge Encyclopedia of the English Language.* Cambridge: Cambridge University Press.

CRYSTAL, David. 2012a. *English as a Global Language. 2nd ed.* Cambridge: Cambridge University Press.

CRYSTAL, David. 2012b. *Spell it Out.* London: Profile Books.

DATAREPORTAL, We Are Social, and MELTWATER. 2023. *Languages most frequently used for web content as of January 2023, by share of websites. Chart.* Accessed April 24, 2023. https://www.statista.com/statistics/262946/most-common-languages-on-the-internet/.

DAUGAARD, Helle Trebbien, JUUL, Holger, POUYENNE, Line Engel Clasen, ELBRO, Carsten, and SIGURDSSON, Baldur. 2020. "The Complexities of Danish Orthography." Nordic Journal of Literacy Research 6, no. 1: 1-21.

DRUMMOND, David C. 2014. *The Truth about the Boiled Frog. Did He Jump Out?* Psychological Reports 114 (1): 296-298.

ELBRO, Carsten, NIELSEN, Ida, and PETERSEN, Dorthe Klint. 1994. "Dyslexia in Adults: Evidence for Deficits in Non-word Reading and in the Phonological Representation of Lexical Items." Annals of Dyslexia 44: 205-226.

HARRIS, James W. 2006. *Syllable Structure and Stress in Spanish: A Nonlinear Analysis (Vol. 8).* The MIT Press.

HOGG, Richard M., and DENISON, David. 2006. A History of the English Language. Cambridge: Cambridge University Press.

FISHER, John H. 1977. *Chancery and the Emergence of Standard Written English in the Fifteenth Century*. Speculum 52, no. 4 (October).

FRANKLIN, Benjamin. 1779. *A Scheme for a New Alphabet and a Reformed Mode of Spelling.* In The Papers of Benjamin Franklin, edited by Leonard W. Labaree, vol. 29, 1-16. New Haven, CT: Yale University Press.

FROST, Ram. 2005. *Orthographic Systems and Skilled Word Recognition Processes in Reading.* The Science of Reading: A Handbook, edited by M. J. Snowling and C. Hulme, 272-295. Malden, MA: Blackwell.

FULK, Robert D., and JURASINSKI, Stefan. 2012. *The Old English Canons of Theodore*. Cambridge: Cambridge University Press.

GUSSENHOVEN, Carlos., & JACOBS, Haike. 2011. *Understanding Phonology.* London: Hodder Education.

JOUANNIN, Christine. 2008. *Pourquoi la Réforme de l'Orthographe de 1990 n'a pas Réussi.* Revue Internationale d'Éducation de Sèvres.

JUUL, Holger, and SIGURDSSON, Baldur. 2005. *Orthography as a handicap? A direct comparison of spelling acquisition in Danish and Icelandic.* Scandinavian Journal of Psychology, 46(3), 263-272.

KACHRU, Braj B. 2005. *Asian Englishes: Beyond the Canon.* Hong Kong: Hong Kong University Press.

KAHNEMAN, Daniel., KNETSCH, Jack L., and THALER, Richard H. 1991. *Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias.* Journal of Economic Perspectives 5 (1): 193-206.

KLEIN, Ernest. 1987. *A Comprehensive Etymological Dictionary of the English Language.* Amsterdam: Elsevier Science Publishers.

LANDERL, Karin, WIMMER, Heinz, and FRITH, Uta. 1997. *The Impact of Orthographic Consistency on Dyslexia: A German-English Comparison.* Cognition 63, no. 3: 315-334.

LASS, Roger. 1984. *Phonology: An Introduction to Basic Concepts.* Cambridge: Cambridge University Press.

LEITH, Dick. 1997. *A Social History of English.* London: Routledge.

LEWIS, Geoffrey. 2002. *The Turkish Language Reform: A Catastrophic Success.* Oxford University Press.

MERRIAM-WEBSTER. 2023a. *Orthography*. Accessed April 16, 2023. https://www.merriam-webster.com/dictionary/orthography.

MERRIAM-WEBSTER. 2023b. *Pangram.* Accessed May 9, 2023. https://www.merriam-webster.com/dictionary/pangram.

MERRIAM-WEBSTER. 2023c. *Spelling.* reform. Accessed May 9, 2023. https://www.merriam-webster.com/dictionary/spelling-reform.

MAYO CLINIC. 2023. *Dyslexia.* Accessed June 1, 2023. https://www.mayoclinic.org/diseases-conditions/dyslexia/symptoms-causes/syc-20353552.

MCMAHON, April M. S. 1994. *Understanding Language Change.* Cambridge: Cambridge University Press.

MILLWARD, Celia M., and HAYES, Mary. 1996. *A Biography of the English Language*. 2nd ed. Fort Worth, TX: Harcourt Brace.

NHS. 2023. *Dyslexia.* Accessed June 1, 2023. https://www.nhs.uk/conditions/dyslexia/.

NORTON, Elizabeth S., and WOLF, Maryanne. 2012. *Rapid Automatized Naming (RAN) and Reading Fluency: Implications for Understanding and Treatment of Reading Disabilities.* Annual Review of Psychology 63: 427-452.

NORTON, Elizabeth S., and TREIMAN, Rebecca. 2012. *Spelling in Kindergarten Children: A Within-Subjects Study of the Effects of Quantity and Distribution of Practice.* Learning and Instruction 22, no. 6: 484-495. Accessed April 23, 2023. https://journals.sagepub.com/doi/epub/10.1177/2332858416675346.

OREG, Shaul, VAKOLA, Maria, and ARMENAKIS, Achilles. 2011. *Change Recipients" Reactions to Organizational Change: A 60-Year Review of Quantitative Studies.* The Journal of Applied Behavioral Science 47 (4): 461-524.

PAGE, R. I. 1999. An Introduction to English Runes. 2nd ed. Woodbridge, Suffolk, UK: The Boydell Press.

PERFETTI, Charles A. 2007. *Reading Ability: Lexical Quality to Comprehension.* Scientific Studies of Reading 11, no. 4: 357-383.

PINKER, Steven. 1994. *The Language Instinct: How the Mind Creates Language.* New York: HarperCollins.

REAL ACADEMIA ESPAÑOLA (RAE). 2010. Ortografía de la lengua española. Madrid: Espasa.

ROACH, Peter. 2009. English Phonetics and Phonology: A Practical Course. 4th ed. Cambridge: Cambridge University Press.

RÖBER-SIEKMEYER, Cordula. 2001. *The Impact of the German Spelling Reform on Reading Processes.* Applied Psycholinguistics 22 (4): 513-534.

RUSS, Charles V. J. 2002. *German Language*. In The New Encyclopædia Britannica: Macropædia. 15th ed.

SAMUELSON, William, and ZECKHAUSER, Richard. 1988. *Status Quo Bias in Decision Making*. Journal of Risk and Uncertainty 1 (1): 7-59.

SCHMALZ, Xenia, MARINUS, Eva, COLTHEART, Max et al. 2015. *Getting to the bottom of orthographic depth.* Psychonomic Bulletin & Review 22: 1614-1629. Accessed July 5, 2023. https://doi.org/10.3758/s13423-015-0835-2.

SCRAGG, Donald G. 1974. *A History of English Spelling.* Manchester: Manchester University Press.

SEYMOUR, Philip H. K., ARO, Mikko, and ERSKINE, Jane M. 2003. *Foundation Literacy Acquisition in European Orthographies.* British Journal of Psychology 94, no. 2: 143-174.

SHARE, David L. 2008. *On the Anglocentricities of Current Reading Research and Practice: The Perils of Overreliance on an "Outlier" Orthography*. Psychological Bulletin 134, no. 4: 584-615.

SIMPLIFIED SPELLING BOARD. 1920. *Handbook of Simplified Spelling.*

STOCKWELL, Robert P., and MINKOVA, Donka. 2001. *English Words: History and Structure.* Cambridge: Cambridge University Press.

TREIMAN, Rebecca, ed. 1997. *Spelling.* Springer Netherlands.

VENEZKY, Richard L. 1999. *The American Way of Spelling: The Structure and Origins of American English Orthography.* Guilford Press.

WELLS, John C. 1982. Accents of English. Cambridge: Cambridge University Press.

WEBSTER, Noah. 1789. Dissertations on the English Language: With Notes, Historical and Critical. Boston: Isaiah Thomas and Company.

WRIGHT, Laura. 2000. *The Development of Standard English, 1300-1800: Theories, Descriptions, Conflicts.* Cambridge: Cambridge University Press.

ZIEGLER, Johannes C., and GOSWAMI, Usha. 2005. *Reading Acquisition, Developmental Dyslexia, and Skilled Reading Across Languages: A Psycholinguistic Grain Size Theory*. Psychological Bulletin 131 (1): 3-29.

Sources that served as an inspiration for this thesis:

https://www.youtube.com/watch?v=TEsqY4MH40s

https://www.spellingsociety.org/history#

List of appendices

Appendix 1

Python code to clean the text

textclean.py

Code > My > textclean.py

```python
import json
import re
import string

def remove_non_alphabetic(text):
    # This will remove any characters that aren't Latin letters, apostrophes or spaces
    return re.sub("[^a-zA-Z' ]", "", text)

def remove_repeated_chars(text):
    # This will remove words with more than 3 same letters in a row
    return re.sub(r'\b\w*([a-zA-Z])\1{3,}\w*\b', '', text)

def process_json_file(file_name):
    with open(file_name, 'r', encoding='utf-8') as json_file:
        data = json.load(json_file)
        with open('cleaned_text.txt', 'w', encoding='utf-8') as out_file:
            for message in data['messages']:  # Assuming messages are stored in a list under 'messages' key
                if 'text' in message:
                    text = message['text']
                    if isinstance(text, str):  # If the 'text' value is string
                        text = remove_non_alphabetic(text)
                        text = remove_repeated_chars(text)
                        out_file.write(text + '\n')  # write cleaned text to file

process_json_file('your_file.json')
```

Appendix 2

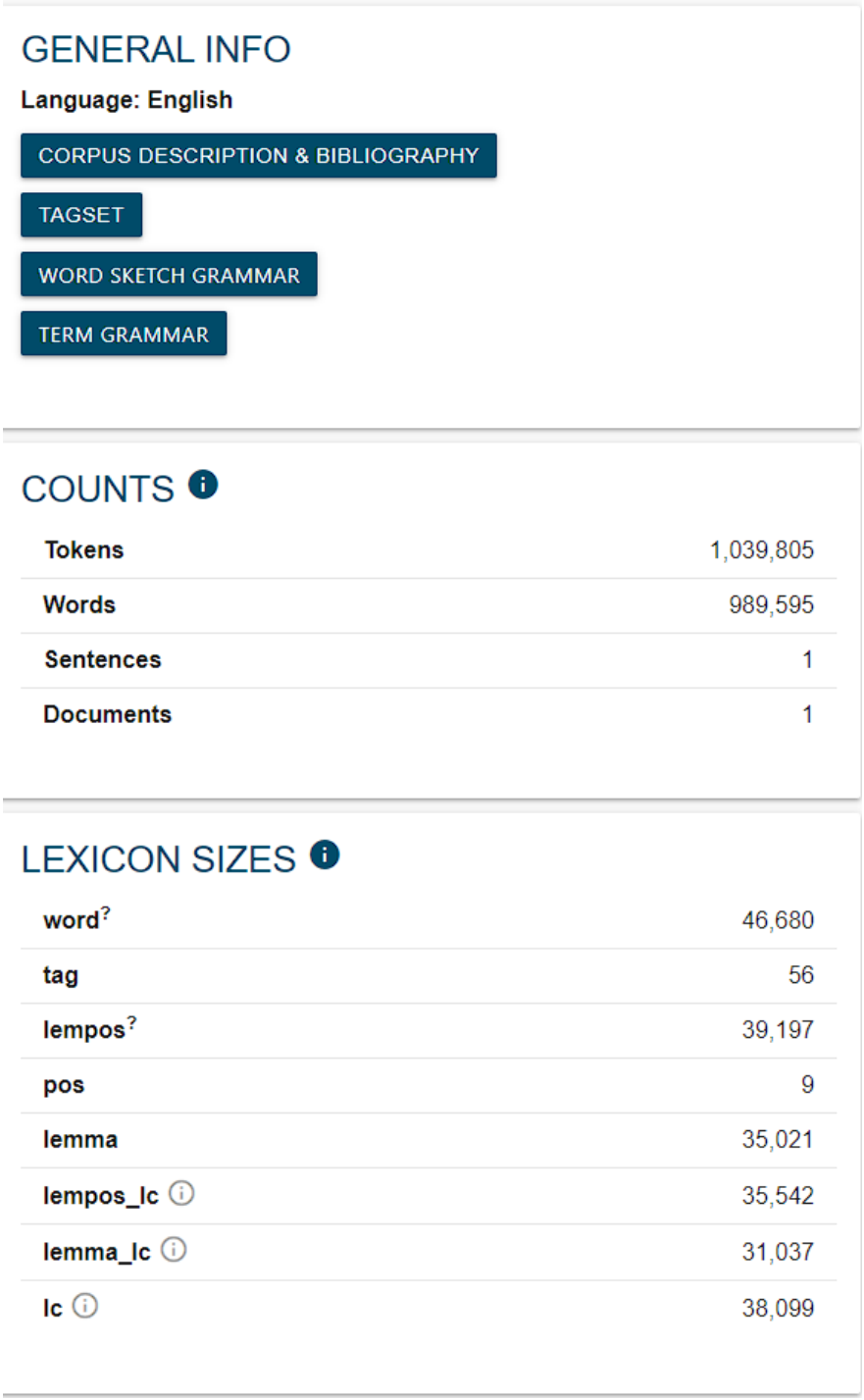
Python code for random word selection

Code > random.py

```python
import random

def random_select_and_write(input_file, output_file, num_lines=10):
    with open(input_file, 'r') as f:
        lines = f.read().splitlines()

    selected_lines = random.sample(lines, num_lines)
    random.shuffle(selected_lines)

    with open(output_file, 'w') as f:
        for i, line in enumerate(selected_lines, start=1):
            f.write(f"{i}. {line}\n")

# Test the function
input_file = 'input.txt'
output_file = 'output.txt'
random_select_and_write(input_file, output_file)

```

Appendix 3

Corpus dashboard on Sketch Engine



GENERAL INFO

Language: English

CORPUS DESCRIPTION & BIBLIOGRAPHY

TAGSET

WORD SKETCH GRAMMAR

TERM GRAMMAR

COUNTS

Tokens	1,039,805
Words	989,595
Sentences	1
Documents	1

LEXICON SIZES

word?	46,680
tag	56
lempos?	39,197
pos	9
lemma	35,021
lempos_lc	35,542
lemma_lc	31,037
lc	38,099