UNIVERSITY OF CALIFORNIA

Santa Barbara

Cohort effects and asymmetrical word-level sound change

A Thesis submitted in partial satisfaction of the

requirements for the degree Master of Arts

in Linguistics

by

Christian D. Brendel

Committee in charge:

Professor Eric W. Campbell, Chair

Professor Argyro Katsika

Professor Marianne Mithun

December 2018

The thesis of Christian D. Brendel is approved.

Argyro Katsika

Marianne Mithun

Eric W. Campbell, Committee Chair

December 2018

Cohort effects and asymmetrical word-level sound change

ACKNOWLEDGEMENTS

I would like to thank my committee members—Eric W. Campbell, Argyro Katsika, and Marianne Mithun—as well as the members of prior committees, concrete or virtual, who have never stopped advising—Fermín Moscoso del Prado Martín, Gabriela Pérez Báez, Holly Ryan, Sandy Feinstein, Jeanne Marie Rose, and Neal Woodman. I am also grateful for the logistical advice and language-specific expertise of my colleagues Sandra Auderset, Caroline Crouch, Kevin Schäfer, Nathaniel Sims, Giorgia Troiani, and Karen Tsai.

ABSTRACT

Cohort effects and asymmetrical word-level sound change

by

Christian D. Brendel

The cohort model of lexical retrieval (Marslen-Wilson & Welsh 1978; Marslen-Wilson & Tyler 1980; Marslen-Wilson 1984, 1987), a theory of speech perception not traditionally applied to questions of historical linguistics, offers evidence that the point at which spoken words become unique within their lexicon is cognitively significant. This uniqueness point divides a word into two regions which appear to be processed by different neural machinery. Historical linguistic research suggests that material which maintains meaningful contrasts is more resistant to reduction than material which does not (Blevins 2005; Blevins & Wedel 2009; Wedel, Jackson, & Kaplan 2013). Perhaps, then, the sublexical regions outlined by the cohort model, which differ in their ability to contrast words, are similarly affected by sound change unevenly distributed across the word. To examine the extent to which this difference in processing could relate to a difference in language change, this study uses the concept of the cohort model to bifurcate genetically-related words selected from eight Indo-European languages which feature uniqueness points within their respective lexicons. Through the comparison of phonological distance among these forms, operationalized as the Levenshtein distance (Levenshtein 1965) between phonemic representations, this analysis finds evidence for an asymmetrical distribution of sound change among cognates: within a word, sound change is more likely to occur after the point at which the word is distinct from all other words in the lexicon.

TABLE OF CONTENTS

I. Introduction

The cohort model of recognition (Marslen-Wilson & Welsh 1978; Marslen-Wilson & Tyler 1980; Marslen-Wilson 1984) is a cognitive theory which has shaped the study of speech processing and lexical access. The model describes a process in which lexical recognition of spoken words is driven using live processing of the auditory stream to winnow down the group of words which might correspond to the sounds being heard: an initial segment of speech activates a cohort of words which then are gradually ruled out as input is perceived. Speakers in isolated lexical decision tasks habitually make decisions about what lexical item is being perceived before the word has been fully heard. This effect occurs at what has been labeled the UNIQUENESS POINT, the point at which only one word in the lexicon begins with the sequence of sounds already heard.

Cohort competition—the state in which auditory input has not yet led to the selection of a single lexical item—among phonologically similar words is demonstrated to result in distinctions in speech processing, as discussed below. The cohort model as formulated in its earliest definitions provides a metric—the uniqueness point—which divides words into a UNIQUE REGION and a REMAINDER REGION, which seem to be, in some way, processed differently in the mind. Marslen-Wilson & Welsh (1978) propose that the early recognition of a spoken word allows the cognitive resources devoted to acoustic-phonetic processing to be freed up as soon as possible so that they can instead be devoted to synthesizing the wholistic meaning of a particular message; less attention "need be paid" (p. 61) to the remaining acoustic input of a lexical item. In terms of the early recognition of a spoken word, the remainder region does not contrast or distinguish one cohort competitor from another.

In the context of language change, the extent to which language material serves a necessary, contrastive function affects its ability to resist erosion and reduction over time. Phonemic material responsible for the maintenance of some kind of contrast resists change (in the form of phonemic mergers) compared to material which bears no such functional load (Blevins & Wedel 2009; Wedel, Jackson, & Kaplan 2013). More generally, the effect of predictability on

phonetic reduction has been a topic of recent investigation. It has been proposed that the increasing probability of a lexical item occurring in a local context is associated with the phonetic reduction of that item (Jurafsky et al. 2001), and a similar effect can be seen on sublexical units of meaning which are predictable from the context of use (Blevins 2005). There is a sort of selection pressure constraining sound change on material responsible for distinguishing meaning that does not apply to material that is predictable or redundant.

The cohort model implies that the remainder region of a word is, in a sense, similarly predictable in that, for a given language, after a uniqueness point there can only be one sequence of phonemes between the uniqueness point and the word-final boundary. In other words, given a unique region and a knowledge of the lexicon, the remainder region can be predicted. Perhaps the resistance to sound change described above accordingly applies to the unique region of a word in the cohort model while the remainder region, which serves no contrastive function in early word recognition, is freer to change.

In this thesis, I attempt to find asymmetry in sound change between the two regions delineated by the uniqueness point. I examine cognate pairs from a dataset of eight Indo-European languages representing three major subgroups (Germanic, Romance, and Balto-Slavic), and I select the lexical items featuring a uniqueness point in their respective languages. Through the assessment of Levenshtein distance (Levenshtein 1965)—a measure which quantifies the dissimilarity between two sequences—I calculate degree of phonemic distance between these cognate pairs with respect to these two regions and compare the two distributions of distance—a proxy for accumulated sound changes—in these regions. I aim to show that the region after the uniqueness point features a higher degree of sound change than the region before, suggesting that the distinctions in cognitive processing described in the cohort model reveal a pattern of asymmetrical sound change within the word.

II. Background

A. The cohort model

The original formulations of the cohort model (Marslen-Wilson & Welsh 1978; Marslen-Wilson & Tyler 1980; Marslen-Wilson 1984) propose a process of spoken word recognition in which listeners continually compare what they hear to possible lexical representations that they know. The 'cohort' is the range of words to which a sequence of auditory input could potentially map at any one time, and the size of the cohort is reduced as more auditory input is perceived: with more auditory input, there are increasingly fewer words which the stream of sound could correspond to. The cohort model contends that speakers make early decisions regarding the identity of a word before the entirety of auditory signal has been perceived, allowing words to be identified as soon as there are no other words in a lexicon to which the sequence of sound could possibly map. The point at which this decision can be made is, for spoken words heard in isolation, where "a particular word becomes uniquely distinguishable from any other word in the language beginning with the same sound sequence" (Marslen-Wilson 1984: p. 141). This point is referred to variously as the Optimal Discrimination Point (Marslen-Wilson 1984) or, in more recent formulations of the model which integrate information other than auditory input, the Recognition Point (Marslen-Wilson 1987), but as my study focuses solely on the phonological properties of words, I use the more specific term of UNIQUENESS POINT (Luce 1986; Radeau, Mousty, & Bertelson 1989). This winnowing process is depicted (with orthographic representations) in Figure 1 below:

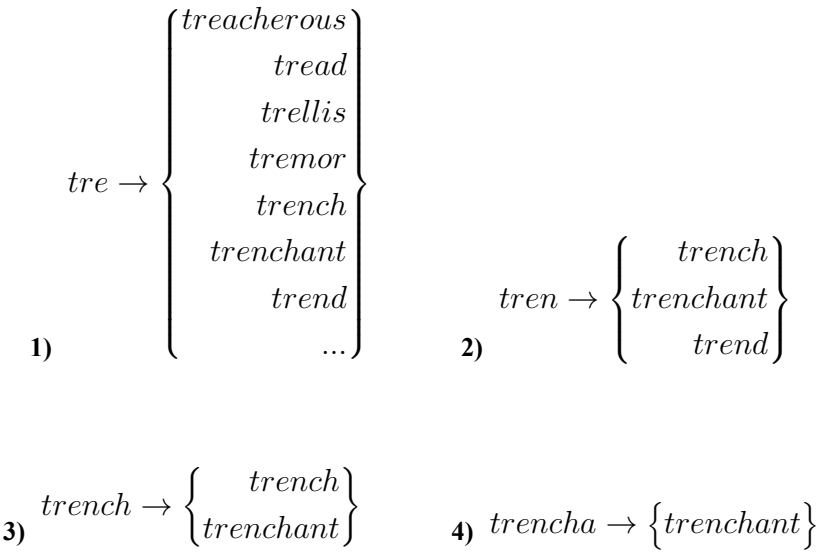


Figure 1: A example of cohort winnowing with progressive auditory input on the left and candidates in English on the right (adapted from Marslen-Wilson 1984). In English, *trenchant* is the only word that could match the sequence beginning with *trencha*. The uniqueness point, then, is immediately after *a*.

The cohort effect is not purely phonological: for example, consonant clusters delay the production of a subsequent vowel and therefore the transitional cues between consonants and vowels, and this phonetic effect has been found to delay the identification of a word (Marslen-Wilson & Tyler 1980). However, regarding lexical candidate selection as contingent on purely acoustic or phonetic properties ignores the human capacity to filter out variable input and signal noise in utterance contexts, and perhaps the operative units in cohort selection should be coded as featural representations which include information about the acoustic-phonetic properties of each phoneme (Marslen-Wilson 1984). Such an approach allows, for example, phonemes like /b/ to activate words with featural near-neighbors like /p/, which differ only in specification of voicing. Additionally, more recent formulations of the cohort model (such as Marslen-Wilson 1987) move away from the idea of an 'all-or-nothing' identification process based on mapping of phonemes to candidates and instead implicate sources of knowledge beyond the auditory stimulus, such as the notion of word frequency, an important character included in most modern formulations and descendants of the cohort model (Klatt 1989; Gaskell & Marslen-Wilson

1997). While this thesis, as an initial application of the cohort model to historical linguistics, uses the concept of uniqueness as formulated in Marslen-Wilson & Welsh (1978), some of the complications which drove the expansion of the cohort model are outlined below.

A challenge for the original 'all-or-nothing' formulation of the cohort model concerns the retroactive identification of an item as a non-word based on data which follows the uniqueness point. In a lexical decision task, if no processing were to occur after the uniqueness point of a word, then listeners would be unable to distinguish non-words like THOUSAIDING /θaʊzaidɪŋ/ from lexical items like THOUSAND /θaʊzənd/, which becomes unique in English at /z/ (Taft & Hambly 1986). Some processing must therefore occur after the resolution of the initial stage of processing involved in selecting an increasingly smaller set of candidates in the cohort model, but the processing of input after the identification of a uniqueness point is not necessarily similar to the method of processing which occurs in advance of the uniqueness point (Marslen-Wilson & Welsh 1978; Cole & Jakimik 1980).

An additional complication is the ability of speakers to identify a non-word input such as GROCODILE /grɒkədaɪəl/ as being a clear misproduction of the valid word CROCODILE /krɒkədaɪəl/. If the cohort model were able to fully explain all lexical identification, words such as GROCODILE would never produce a cohort that includes the intended CROCODILE since the first phoneme /g/ can only produce a cohort of words beginning with /g/. And yet, speakers are able to identify such non-words as likely being misproductions of valid words which share a high degree of similarity after the initial phoneme. Taft & Hambly (1986) conduct several experiments to investigate the degree to which the cohort model can account for the processing of such words. The finding most relevant to this discussion is the delay in lexical decision for non-words like MEP /mɛp/ and MEPSIG /mɛpsɪg/. Both are non-words which diverge from all real English words at the phoneme /p/, and the cohort model would thus predict an equal delay in identification of non-wordhood on average. However, the results of a decision task show that participants take more time to identify the longer MEPSIG as a non-word. One reason for this finding is that the inclusion of additional phonetic information in the stimulus for the

lexical decision task delays the full acceptance of the result of cohort model selection (Marslen-Wilson & Tyler 1980). Additional sounds after a uniqueness point, as discussed above, must be processed in some way to rule out non-words like THOUSAIDING, and perhaps even in non-words this effect can be seen: as long as there is phonetic input, a speaker processes it in some way. Marslen-Wilson (1987) refines his earlier conception of lexical recognition, proposing that bottom-up effects (as seen in the original formulations of the cohort model) are indeed used for early identification of a word in isolation, but that this process competes with a separate but parallel process that evaluates the likelihood of the candidate word to be an appropriate fit in terms of the context.

Even if the beginnings of words are most important for lexical recognition, what comes after the uniqueness point is not irrelevant: the ends of words are more important for lexical access than material in the middle of words, and disruptions to the ends of words accordingly affect lexical access more acutely than corruptions of word-medial segments (Hawkins & Cutler 1988). The cohort model has previously been implicated as an explanatory factor for the cross-linguistic typological preference for suffixation over prefixation[1]: if the model is accurate, neural machinery races to select a semantic representation from auditory input as quickly as possible and therefore works most efficiently when content words are initial and grammatical elements are postposed (Hall 1988). The bifurcation of words suggested by the cohort model does not imply that one region is more important than the other, but that these two regions are best suited to semantic and morphological functions, respectively.

The discussion above primarily concerns the access of isolated words in contrived experimental settings. This context is, of course, not the natural environment of connected utterance. However, some evidence suggests that the early recognition of a word applies in, and may even facilitate, the processing of real-time discourse. Marslen-Wilson & Welsh (1978) create a task in which participants listen to prose passages which feature mispronunciations of particular phonemes, and the participants repeat these passages aloud nearly simultaneously. When par-

[1]Languages whose primary affixation strategy is suffixing outnumber those which are primarily prefixing by a ratio of 7:1 (Dryer 2013).

ticipants repeat the passages, they fluently fix highly noticeable mispronunciations that occur in the third syllable of a word which is primed by context[2]. The authors conclude that discourse context and the early recognition of individual words result in a system of lexical recognition optimized "to devote the minimum feasible processing capacity to the detailed interpretation of the incoming acoustic-phonetic input" (p. 61); the earlier that a word can be identified from a stream of phonemes, the sooner that this processing power can be devoted to the continuous input still being received (such as markers of grammatical relation or, indeed, new lexemes). The context of an utterance is not incorporated into the methodology of this thesis, but these findings suggest the cohort effect applies in naturalistic speech, the domain of sound change.

In this study, an approach is taken which appreciates the distinctions in functional parallelism in the word recognition process (Cole & Jakimik 1980; Marslen-Wilson 1987), but follows earlier work in operationalizing the uniqueness point as the point at which a word becomes unique phonologically (Marslen-Wilson & Welsh 1978; Luce 1986; Radeau, Mousty, & Bertelson 1989). This point is used as the basis for dividing words into two regions which seem to be processed differently by speakers, and this study examines whether these distinctions in process result in distinctions in sound change.

B. The cognitive word

1. Problems in working with words

These studies on lexical recognition presume the existence of discrete words in memory, yet the definition of wordhood is not straightforward, despite previous and ongoing research which is predicated on the existence of the word. Haspelmath (2011) outlines the challenge in categorizing words as discrete, strictly segmented units. The practice of segmenting spoken utterances into combinatorial units, he maintains, has been largely influenced by the Western tradition of using visual space to separate chunks of written text; the biases of literacy are im-

[2]A sentence in which the mispronunciation is primed by context is something like *He wanted to smoke a *cigaredge (cigarette)*, as opposed to something like *He wanted to purchase a *cigaredge (cigarette)*.

posed on natural language to the extent that literate speakers of a language with such visual segmentation strategies "have no intuitions that are independent of the writing rules they have learned" (Haspelmath 2011: p. 35). Haspelmath (2011) ultimately concludes that no satisfactory criteria for a cross-linguistic definition of the word exist. While descriptive linguists reject the idea that natural discourse is contingent on institutional prescriptions of grammar and pronunciation, the very idea of a 'word' as the basic unit of syntactic analysis in naturally occurring speech is a perception shaped by this same artificial force.

The prevalent predication of the concept of a word expands beyond reference to syntactic definitions of wordhood. Schiering, Bickel, & Hildebrandt (2010) critique the universal applicability of the Prosodic Hierarchy, which holds the domain of the prosodic word as a crucial nexus of interaction between morphological structure and phonological processes. The authors argue that such a hierarchy does not hold cross-linguistically, and that in particular the domain of prosodic word is better defined on a language-by-language basis when its existence has explanatory power for linguistic structure, as opposed to the assumption that it is the necessary mediator between rhythmic and phrasal phenomena, as the Prosodic Hierarchy situates it. As the supposed interface between the post-lexical and the sublexical—terms which themselves are predicated on the existence of a word—the specification of this critical unit impacts not just the study of phenomena constrained to its purported layer of language structure, but delineates other layers of linguistic analysis that do not ostensibly concern it. Perhaps the most applicable cross-linguistic definition of a word is that it is the unit—be it prosodic, morphological, syntactic, lexical, or otherwise—whose existence is intuitive and fundamental in many of the subdisciplines of linguistics, as well as in the minds of speakers.

The problems inherent in arriving at a cross-linguistic criterion of wordhood are not unique to this thesis. However, this discussion is particularly applicable to both the cohort model and the present study, which too are founded on the assumption of a cognitive basis for the word. The candidate selection process of the cohort model is predicated on the existence of particular chunks of language information which are stored in the mental lexicon, lemmas which coincide

in surface structure with attested wordforms. As the studies below show, even if a satisfying, enveloping pedigree of wordhood cannot be arrived at across all linguistic disciplines, there is nonetheless evidence that some units of language (including monomorphemic units and certain lexicalized polymorphemic units, such as some compounds or items with derivational affixes) are processed differently than inflectional units, and this distinction will be the basis of the word as used in this study.

While I will do nothing as lofty (or impossible) as offering a universal definition of the word, neuro-cognitive research examining morphologically varied forms (described in the next section) can help suggest a cognitive basis for a definition of the word which is useful when discussing lexical recognition and, consequently, for this study. This sort of cognitive word is what composes cohorts in cohort competition.

2. Morphological processes of word formation and cognition

The distinction between inflected and derived forms is of importance when discussing lexical recognition—or storage in the mental lexicon—of morphologically-complex words. For the purpose of this study, it is a practical concern as well: the selection of items for comparison should consist of words which seem to be stored as discrete items in the minds of speakers. Most immediately the question is whether any forms other than simplex words should be selected, and if so, on what basis inflected items should be excluded.

A functionally-based linguistic analysis might contend that one working definition of the difference between a derivational morpheme and an inflectional morpheme is that a derivational morpheme changes core semantic value (e.g. the antonymic relationship between $kind_{ADJ}$ and $[un\text{-}kind_{ADJ}]_{ADJ}$) or lexical class (e.g. from adjective to noun, as in $kind_{ADJ}$ to $[kind_{ADJ}\text{-}ness]_{N}$), whereas an inflectional morpheme yields a change in quantity, degree, tense, aspect, and other such qualities (e.g. $kind_{N}$ and the plural $[kind_{N}\text{-}s]_{N}$). Despite the deceptive similarity of form in the above representations—both types of morpheme appear to be simply glued onto existing roots, so we might wonder if there truly are two phenomena at work—evidence from the

cognitive literature on lexical storage and access of derived and inflected words indeed attests to the differentiation of morphologically-complex forms in terms of fusion and separability (of the sort described in Bickel & Zúñiga 2017). A cline of lexicalization distinguishes (but also blurs the lines between) the binary distinction between inflection and derivation.

Before I begin discussing the evidence, we can see that, in some ways, this distinction is unsurprising: although not a rigorous metric, we can note that derived forms, not inflected forms, are given the status of citation form in dictionaries whereas inflected forms are not, and this longstanding tradition hints at some difference in how speakers conceive of derived versus inflected forms, even if the division of morphemes into these categories is not without ambiguity or questions of cross-linguistic applicability. Another metaphor can be found in the process of caching in computer science, where a balance between finite amounts of memory and finite speeds of calculation must be achieved. If the result of a particular algorithm is frequently-accessed but computationally-expensive or time-consuming to calculate, the result itself can be stored (cached) in memory, negating the need to re-run the calculation each time the result is requested. While the comparison is not perfect (generally, the most frequently-used data is cached, but that is not a claim I am making about language), I will revisit this analogy periodically.

Several experiments have examined the idea that word formation from morphological processes are variously and differently represented in the mind. One pair of studies (Marslen-Wilson & Tyler 1998; Longworth et al. 2005) examine a group of English-speaking patients with nonfluent aphasia who have suffered damage to the frontotemporal system of the left hemisphere of the brain, a region known to underpin the processing of syntax (Marslen-Wilson, Bozic, & Tyler 2014). These patients have difficulty in tasks that require decomposing a morphologically-complex, regularly-inflected form (like *accused*), but exhibit no such impairment in tests involving semantic priming of uninflected words or, more relevant to the current discussion, irregularly inflected forms like *shook*, suggesting that irregular inflected forms can be stored and accessed as whole forms, not generated on-line in the way that regular forms like

accused can be produced.

Evidence from such studies examining patients with brain trauma has been complemented by more recent experiments using functional magnetic resonance imaging (fMRI) of the brain to observe neural activity in participants with normative neurology in response to linguistic stimuli, pointing to a two-part division in the way the brain processes speech (or at least isolated speech). A selection of relevant experiments, which provide evidence for this bifurcation as well as more closely examine the idea of a cognitive basis for wordhood, are described in detail below.

Before discussing this evidence, I present first a conclusion of a formative study by Bozic et al. (2010) which shapes the research protocols of the studies below. Bozic et al. (2010) synthesize and confirm earlier scattered findings about speech comprehension and delineate the two dissociable patterns of neural activity involving speech processing:

1. A BILATERALIZED fronto-temporal system (distributed over both hemispheres of the brain), a generalist mechanism responsible for many nonspecific, non-linguistic cognitive functions, which is implicated in the mapping of sounds to morphologically-simplex words.
2. A LEFT-LATERALIZED network concentrated in the left hemisphere inferior frontal cortex (LIFC), which is engaged by morphosyntactically complex linguistic input and supports the decompositional and combinatorial processes involved in the comprehension of structured utterances.

This second area—the left-lateralized cortex—involves the same region damaged in the case of the aphasia patients referenced above (Marslen-Wilson & Tyler 1998; Longworth et al. 2005) who display difficulty in processing morphosyntax, while their undamaged bilateralized fronto-temporal regions are associated with the cognitive functions—such as decoding phonological information from an auditory stream and connecting these phonemic representations to stored lexical items—which remain unaffected by their injuries.

In the study which proposes this schema regarding the neural loci of speech comprehension, Bozic et al. (2010) examine the differences in processing among categories of words (Ta-

ble 1) among speakers of English. Employing both behavioral and neurological assays, the researchers conducted (i) a gap detection task, where the amount of time required for a participant to identify pauses (thus segmenting words) in a stream of auditory input serves as a diagnostic for the amount of cognitive effort in accessing a particular lexical item, and (ii) fMRI to study the activation of bilateral fronto-temporal-parietal neural regions, the larger area comprising both regions implicated in speech comprehension, during the gap detection task. As a goal of the study was to identify which regions in the fronto-temporal-parietal region were responsible for differing facets of speech processing, the verbal stimuli were selected on the basis of the potential to induce greater processing loads on these separate areas. The wordlist comprised five categories which covaried in terms of morphological and phonological complexity[3].

	Word type	Example	Embedded stem	IRP[4]	Region primarily activated
1	Regular past tense	*prayed*	? (pray)[5]	Y	Left-lateralized
2	Pseudo-past tense	*trade*	Y (tray)	Y	Left-lateralized
3	No stem, IRP	*blend*	N	Y	Left-lateralized
4	Stem only	*claim*	Y (clay)	N	Bilateral frontal
5	Simple	*dream*	N	N	Bilateral middle temporal

Table 1: Word categories surveyed, synthesized from Bozic et al. 2010 (some categories renamed for clarity)

[3]Bozic et al. (2010) regard only the processing of morphosyntax as 'linguistic' and regard the identification of a simplex lexical item (e.g. distinguishing simplex words with partially-overlapping phonemic forms aligned from the word onset, such as English *bar* and *bark*) as 'nonlinguistic'. While the results of their study below demonstrate differences in processing between simplex pairs like *bar*/*barn* and morphologically-complex words like *barred*, I am very critical of the labeling of phonological processing as 'nonlinguistic', which problematically implies that (i) phonemic segmentation does not require linguistic resources (ie, acquisition of the phonological system), and (ii) phonemic representations are not composed of combinatorial units themselves. While the study provides convincing evidence of a distinction between phonological processing and additional morphological analysis, terming the former 'nonlinguistic' is misleading in implying that lexical disambiguation through phonological mechanisms is fundamentally no different from distinguishing the first four notes of Beethoven's *Symphony No. 5* from the opening theme of *Star Wars*.

[4]This codes for the presence of the English Inflectional Rhyme Pattern, defined below this table.

[5]Bozic et al. (2010) regard the relationship between embedded stem and stimulus form among these words as fundamentally different from the other stimulus types because inflected forms are not thought to have independent lexical representation in the brain, but instead are produced as the result of on-line processes. See the later summary of Bozic et al. (2013)

Phonological complexity here refers to the competition between forms in a cohort. Simplex words which partly contain another word (marked by EMBEDDED STEM in Table 1) were selected on the rationale that these embedded stems could misleadingly trigger a sound-to-lexical process (presumably handled by bilateral machinery) due to the cohort effect (e.g. initially selecting *core* while perceiving the auditory signal of *court*). Morphological processing was targeted by forms which were either a well-formed of an English verb (like *prayed*) or a simplex noun which features the English Inflectional Rhyme Pattern (IRP), defined as a word-final coda such that:

$$\begin{bmatrix} +\text{coronal} \\ \alpha\text{voice} \end{bmatrix} / \begin{bmatrix} \alpha\text{voice} \end{bmatrix}\_\_\_\_\#$$

Figure 2: English Inflectional Rhyme Pattern

This specification captures commonly occurring regular English inflectional suffixes (the nominal plural suffix *-s* and the past suffix *-ed*) and is purported to be a phonological cue to morphological structure (Post et al. 2008); its presence signals to the processing machinery that the sequence just perceived may well be a suffix and not part of the stem. The final category of words consists of simplex forms which have neither embedded stem nor IRP.

The reaction time of participants in the gap detection task was significantly shorter for simplex (Category 5) words than those with an IRP and/or embedded stem. Most importantly for the present review, the fMRI results showed a strong dissociation in activated region between items which were predicted to be analyzed morphologically and those which were expected not to induce decomposition: items with IRP, regardless of whether or not they are actually suffixed forms, were heavily left-lateralized. The auditory signal can induce an attempt at morphological decomposition even if the lexical item identified as the result of cohort model selection is itself not able to accept the candidate affix.

Although this study shows convincing evidence that there is a strong cognitive basis for the division of morphosyntactic processes and simple lexical retrieval, the question presently most

relevant to this thesis is, of course, whether there is a similar neuropsychological distinction between derivational morphology and inflectional morphology. For the purpose of data selection in this thesis, it is vital to determine how derived forms are accessed and if they are processed more similarly to regular inflected forms or lexical items stored as whole entries in memory. If a listener must morphologically decompose derived forms before completing a lexical decision task, derivationally-complex forms would not be suitable for inclusion in an analysis relying on the cohort model of selection. Previous research relying on lexical decision tasks provide differing answers to this question. Highly productive derivational morphemes like *-ness* cause priming effects for other words featuring the same derivational morpheme (Marslen-Wilson, Hare, & Older 1993) (i.e. a word like *toughness* primes a semantically distant word like *darkness*), suggesting that on some level decomposition of derivationally-complex words occurs. However, Ford, Davis, & Marslen-Wilson (2010) show that these effects are limited only to derivational morphemes of high synchronic productivity; by contrast, morphemes which are discrete but much less productive in modern English (such as *-ic* as in *mythic* and *-th* as in *warmth*) are less likely to undergo decomposition. The evidence from these lexical decision-based studies perhaps illustrates grammaticalization in motion, with more grammaticalized, less transparent forms more likely to be stored as whole chunks in memory.

In an attempt to clarify the uncertain status of the representation of derivationally-complex words by using instead the lens of neurobiology, Bozic et al. (2013) study human subjects with normative neurology to examine the degree, if any, that derivationally-complex words activate the left-hemispheric system (the same region of the brain responsible for the aphasia patients' difficulties processing inflected words in Marslen-Wilson, Hare, & Older 1993 and Longworth et al. 2005). If derived forms are decomposed into constituent morphemes as a listener perceives speech, the left-hemispheric system—the center of syntactic processing—should be engaged, but if these derived forms are stored in memory—cached—as whole lexical items, the left hemisphere frontotemporal system should not show signs of selective activation. Participants were exposed to a diverse set of words both simplex and complex, including pairs comprised

of stem and derived form (such as *brave* and *bravely*) and words whose derivational affixes varied in productivity. This set also included derived words where the semantic relationship between the derived form and the stem is opaque (like *arch* and *archer*) and words that could potentially be analyzed as beginning with a stem (*scan* and *scandal*) but which have no [recognizable] suffix. An fMRI brain scan of participants showed that the task did not selectively activate the left-hemispheric system, regardless of the productivity of the derivational affixes or the transparency in relationship between derived form and stem. This result suggests that derivationally-complex words are processed similarly to simplex words and in a very different manner to inflected forms.

All of the studies described above were conducted on English, which has a relatively small set of inflectional morpheme types and where the burden of decomposing words is less frequent and perhaps less demanding compared to languages where few tokens are bare stems (Marslen-Wilson, Bozic, & Tyler 2014). Consequently, these findings might not hold true concerning languages for which the contrast between 'inflected form' and 'uninflected form' is not as clear as in English[6]. Szlachta et al. (2012) investigate the locus of processing of nouns in Polish, a language which features a greater typical density of inflectional morphemes on an average token and whose case-marked nominal paradigm stands starkly opposed to the lack of case-marking on open-class nouns in English.

(1) An example of three case-marked forms in Polish

a.	*dom-∅*	b.	*dom-u*	c.	*dom-owi*
	house-NOM		house-GEN		house-DAT
	'house'		'[of the] house'		'to the house'

(Szlachta et al. 2012)

[6]Contrast nouns in English, whose singular forms in usage largely are coordinate with their stems (e.g. *dog*, *shoe*), with Icelandic, where inflectional affixes are obligatory on most nouns (e.g. *hund-ur* 'dog-NOM', where the stem *hund-* is only realized bare in production in the accusative *hund* dog.ACC.)

The examples in (1) illustrate three of the seven grammatical cases in Polish, which also features three genders and a distinction in animacy parasitic on these systems (Stone 1990). The zero-marked nominative form[7] of *dom* 'house' is but one option in a paradigm that speakers constantly navigate. Szlachta et al. (2012) contend that the productivity and breadth of this system renders few words morphologically simple. The researchers aim to examine for Polish the applicability of Bozic et al. (2010) regarding the strong difference in neural activation between inflected and non-inflected forms in English. Activation patterns for stimuli of increased perceptual complexity (ie, embedded stems) were associated with strong activation of the bilateral frontal region, as in Bozic et al. (2010), but when comparing overtly-marked forms like *dom-u* 'house-GEN' with forms like *dom* 'house', no difference in activation pattern was seen between the two. However, when compared with the baseline level of neural activation, both categories were left-lateralized and bilateral frontally activated. The overtly marked and the zero-marked forms of the nouns (and verbs tested, as well) pattern together, suggesting that zero-marked forms in Polish are indeed zero-marked: cognitively, all Polish nouns are treated as inflectionally complex, whether they have overt inflection or not. Given the morphological character of the highly productive Polish inflectional system, the results of the fMRI study are within expectation—inflection is primarily left-lateralized in both English and Polish, but the inflectional machinery is invoked for processing almost every noun and verb in Polish, unlike English. The categories of bare stem vs. inflected vs. derived varies substantially, even among genetically-related European languages, and the realities of the morphological system have consequences for the conclusions we make about the neural frameworks underlying it.

As broad conclusions about morphological representation in the brain are dependent on the nature of the languages the owners of those brains speak, so too do the choices made in selecting viable wordforms affect the footing and claims of the present study. In addition to the

[7] The study assumes that the nominative form should be considered *dom-∅* and not *dom*—in other words, the assumption is that the nominative form is not default, but instead is the result of a rich inflectional system for which the nominative case is marked with a null. The findings of Szlachta et al. 2012 suggest neurological reality in the theoretical proposal that this zero-morpheme 'exists'.

immediate concern of the diachronic implications of the cohort model in particular, this thesis (or this thesis author) is more broadly concerned with adopting and adapting the insights from cognitive and neurolinguistic theory and examining the extent to which they can be applied, either literally or through metaphor, to the comparative analysis of language change—namely, (i) examining the extent to which, if any, the duality in processing auditory data is reflected by or fossilized in relative differences in sound change operating within the domain of the word, and (ii) applying the conclusions of the cohort model to bifurcate words at their points of intralexicon phonological uniqueness, a point at which the cohort model predicts a shift in processing strategy from a fast, greedy operation that winnows down candidate words as input is perceived to the secondary process, however it operates, which invalidates earlier selections if subsequent phonological material results in a non-word. The recent fMRI neurological studies and the traditional behavioral experiments conducted in the formulation of the cohort model both advocate for a two-part segmentation. If a knowledge of contemporary cognition has relevance for investigating language change historically, the conclusions of the recent fMRI research might suggest that, due to the similarity in storage and comprehension of highly lexicalized derived forms, there should be no significant difference in sound change between simplex forms and derived forms, but perhaps inflection displays a different pattern due to its ever-on-line, ever-productive nature.

It seems, then, that there is a spectrum of cacheability of content words, ranging from words with inflectional markers (which must be processed on-line in the left-hemispheric system, the least cacheable) to non-inflected words, including both non-complex words and derived words, which evidently cease to be analyzed as morphologically complex in terms of lexical access and seem to be stored in the same manner as non-complex words.

Due to the wealth of evidence suggesting the cognitive affinity between simplex and derived items, this study includes derived items where available but excludes inflected forms (outside of any inflectional data present in highly lexicalized citation forms, like *sitzen* for German ‘to sit’, which features the infinitive marker *-en*) due to the significant differences in processing

associated with them. The inclusion of derived forms increases the pool of candidates which can be analyzed in this study. As the most recent neuropsychological evidence suggests that lexical access functions similarly between derived forms and other morphologically-simplex or compound words in experiments involving the cohort effect, their inclusion only enrichens the study. Criticisms that can be applied to the nature of examining such an effect on isolated words without any surrounding context may be warranted, but these concerns apply equally to an analysis conducted with or without derived words.

C. Hypotheses

The cohort model offers a protocol: if the lexicon is known, all words with a uniqueness point can be bifurcated, and the cohort model proposes that the processing mechanisms for the phonological material on either side of this word, in day-to-day perception, are different strategies. This study is grounded in both (i) the conclusions of the cognitive literature above suggesting that these two regions are subject to different processing strategies, and (ii) the implication from the historical linguistics literature (Blevins 2005; Blevins & Wedel 2009; Wedel, Jackson, & Kaplan 2013) that material which is necessary for distinguishing meaning might resist some forms of sound change more strongly than material not involved in maintaining these distinctions. The primary hypothesis in this thesis is that for cognate pairs which each feature a uniqueness point (at least one of which is before the final boundary), the degree of sound change is greater in the remainder region than in the unique region. If this one-tailed hypothesis—that the remainder region changes more than the unique region—is supported by my analysis, then there is evidence that either or both of these factors could play a role in sound change.

If this hypothesis must be rejected, then a more conservative two-tailed hypothesis could be formulated: that there is a significant difference in sound change between the unique region and the remainder region, regardless of which region features more or less change. Accepting this secondary hypothesis would still give support to the idea that the distinct mechanisms in speech processing are factors in asymmetrical sound change (i above), but would not provide

support for the idea (ii above) that material which bears greater functional load is subject to lower rates of sound change.

III. Methodology

This analysis applies the cohort model to the dissemination of sound change within a particular language family. I investigate a set of cognate words—selected on the basis of the likely domain of the cohort candidate (lemmas and derived forms, as described above)—and divide them into regions delineated by the uniqueness point (whose existence has been asserted by formulations of the cohort theory). The quantification of difference between these subword regions is taken to be a diagnostic for sound change undergone between etymon and reflex and could reveal patterns of asymmetries in the distribution of sublexical sound change.

Eight languages from the Indo-European family were selected for this analysis. These languages along with the lexicographic resources utilized are listed in Table 2.

Language	Family	Resource	Citation
Dutch	Germanic	CELEX2	Baayen, Piepenbrock, & Gulikers 1995
English	Germanic	CELEX2	
German	Germanic	CELEX2	
Serbian	Balto-Slavic	vwr	Kostić 1999
Russian	Balto-Slavic	CMU Sphinx	*CMU Sphinx* n.d.
Italian	Italic (Romance)	CMU Sphinx	
Spanish	Italic (Romance)	CMU Sphinx	
French	Italic (Romance)	Lexique 3	New, Pallier, & Ferrand 2005

Table 2: Lexicons analyzed

The selection of the Indo-European language family as a whole is a point which is discussed in some detail below. In regards to the choice of these particular eight languages of all the extant Indo-European languages, the composition of this set was designed to balance multiple concerns, namely (i) the desire to incorporate data from diverse subfamilies to examine the extent to which the cohort effect might affect change, while simultaneously (ii) not forcing any one language to be an exemplar for the entirety of its most immediate taxon. In other words, the incorporation of some measure of diversity across the Indo-European family is important to evaluate the potential universal applicability of this methodology, while the selection of more than one member per subfamily protects against any over-representation of the idiosyncrasies of

any one language. Furthermore, (iii) the eight languages selected had what were deemed (at the beginning of this study) to be digitized lexical resources which lent themselves to cross-linguistic, automated phonological comparison. In reality, this last point is true to varying degrees. The dictionaries represented are sourced from four different projects and vary in degree of analytic depth (in terms of establishment of relationships between the various headwords, particularly between roots and lemmas) as well as hand-curation. The idiosyncrasies of particular dictionaries are described as appropriate below.

A. Rationale of choosing Indo-European

This project was conducted using Indo-European data for four primary reasons:

1. The availability of accessible and digitized phonological data
2. The availability of accessible and digitized resources establishing etymological relationships among specific words
3. The lack of lexical tone in the languages identified in Table 2
4. The (relative) clarity in segmentation of words[8]

Firstly, while the application of the cohort model is not limited to Indo-European languages, the ready availability of lexical databases featuring phonological transcriptions of all items is necessary to identify the set of unique subwords in a language's vocabulary—in other words, the more complete the lexicon, the more accurately the uniqueness points of words can be identified, so languages which feature relatively complete lexicographic resources provide a higher number of datapoints on which the analysis can be performed. While ideally this analysis would feature a much larger number of languages which sampled all branches of the Indo-European language family, these eight were selected as a starting point due to the accessibility and completeness of the data.

[8]Of course the definition of wordhood is a fraught topic, as I discussed above, but since this analysis (and as many of the analyses involving lexical access described above) relies on prescriptive sources (ie, dictionaries) for information, I refer here to the relative unambiguity of headwords in dictionaries of major European languages.

Secondly, the present study relies heavily on the establishment of cognate sets featuring unique subwords across languages, and the longstanding attention of historical linguists to Indo-European, coupled with more recent efforts to digitize these resources, facilitates the mass comparison of cognates across the family. The Indo-European Lexical Cognacy Database (IELex) project (Dunn 2012, based on cognate decisions from Dyen, Kruskal, & Black 1992 (the cognate database used for this thesis) and Ringe, Warnow, & Taylor 2002) provides 5,013 sets of cognates across 163 Indo-European languages, facilitating the automated comparison of cognates across genetically-related languages. For the purpose of this study, the most accessible lexicographic resources were selected, but optimally the majority of the languages included in the IELex database could be included in the survey (although for many of these languages, digitized, open-access sources do not presently exist).

Similar lexical databases of high quality indeed already exist for other language families, such as the Sino-Tibetan Etymological Dictionary and Thesaurus (STEDT) project (Bruhn et al. 2015). The existence of these sources provides an avenue for replication of this study, including the ability to investigate the effect of lexical tone, an additional character relevant for some languages which is not examined in this paper due to its absence (in general) from the Indo-European dataset. This concern is the third reason for the selection of Indo-European data: the parsing of lexical tone adds an additional layer of complexity to lexical access. For instance, in Cantonese, lexical tone has a high functional load but has been seen to induce greater error in lexical decision tasks between tonal minimal pairs than in segmental minimal pairs (Cutler & Chen 1997), and in Mandarin tone has been shown both to prime segmentally-identical pairs in certain contexts (Lee 2007) as well as to induce competition (that is, increased decision time) among segmentally-dissimilar candidates which share lexical tone (Poss, Hung, & Will 2008). Additionally, the locus of tone processing in lexical tasks may exist in different cognitive machinery: in an event-related potential (ERP) study, Malins & Joanisse (2012) find that Mandarin speakers display more activity in the left hemisphere when processing tonal minimal pairs (such as *huá* ‘flower’ and *huâ* ‘painting’) than when processing words that comprise a

cohort based on early segmental similarity (such as *huá* 'flower' and *huí* 'gray')[9]. The selection of Indo-European data enables these concerns regarding suprasegmentals to be reserved for future study and allows this initial examination of cohort-induced effects in diachrony to be tested on data which raises comparatively fewer questions regarding the synchronic processing of the candidates themselves.

The final consideration is perhaps the most important of the four on a theoretical level: as discussed above, cross-linguistically and even within a single language, the definition of *word* varies greatly. The assumption that a word written in a dictionary is a unit which visually represents an identical unit (of a spoken word) stored in the mental lexicon provides some separation from the issue of defining the morphosyntactic or phonological word in natural language. The cost of this abstraction is that we study an artificial conception of 'the word' which is defined by a written standard, and moreover a chiefly Western standard. Accordingly, the decisions in data selection in the present study were made with the intention of remaining as close as possible to the linguistic sources of the empirical evidence which underpins the cohort model. The challenge of working with data from languages whose literary traditions diverge more greatly from the European languages on which the cohort model is based increasingly necessitates further questions of wordhood which will not be addressed in this study.

The present study focuses exclusively on Indo-European data, but the question of the generalizability of this model to non-Indo-European languages remains a topic of significant importance. While studies on isolating Sino-Tibetan languages are not uncommon, the applicability of the cohort model to highly synthetic languages remains an open question. If the cohort model holds for these languages (as supported through research on these speakers' processing of lexical access), the present diachronic methodology could be applied to non-Indo-European languages in future projects.

[9]It is uncertain whether these findings are limited to languages like Chinese with relatively small segment inventories and whose lexical tones bear high functional load in distinguishing near-homophones.

B. Cognate identification

None of the eight dictionaries provide etymological information or glosses in any of the other languages represented. Consequently, these relationships were determined through a comparative Indo-European database initially compiled by Isidor Dyen (Dyen, Kruskal, & Black 1992) on punch cards in 1970 and now digitized. The Dyen list (as I will refer to it) is a compilation of 200 word senses documented for 84 Indo-European language varieties[10] (to various degrees of completion per variety). In total, there are 21,602 wordforms and 3,308 distinct relationships among cognate groups with four levels of confidence judgments. These reflexes are from the Swadesh list of core vocabulary as formulated by Swadesh (1952), a selection of words purported to be resistant to lexical replacement and borrowing over time.

The degree of lexical replacement is a concern of the present study only inasmuch as the lack of lexical replacement in a finite table of cognates improves the number of eligible extant cognate sets to investigate—the impact of the cohort effect, if seen at all, should apply to equally to core vocabulary as well as to any other vocabulary for which a broad set of cognates can be found. While the use of the 200-word Swadesh list might allow for a broad net of mass lexical comparison across all 84 languages (for which purpose it was used in Dyen, Kruskal, & Black 1992), the Dyen list contains only orthographic representations of reflexes, representations which often deviate from the standard orthography of a language due to their encoding in Latin (ASCII) characters only[11]. This fact makes it suitable for lexicostatistics which operate

[10] Technically, there are 95 varieties in the list, but 11 of these varieties are identical or similar to other varieties in the dataset, and are noted as originating from Fodor 1961 (a source I have not been able to track down). The words these duplicate varieties comprise are also either near variants of or identical to the corresponding language from Dyen's original punched form. All 95 were included in the search for cognates between the individual language dictionaries. Of the data analyzed for this study, only two—Serbian and Dutch—had duplicate language variety entries. Since the Dyen list was only used for the establishment of mapping between cognate forms across real words that were already present in their respective lexical databases, and due also to the fact that all duplicate cognate pairs were eliminated from the final dataframe for this thesis, the redundancy in the Dyen list did not impact the analysis.

[11] For example, languages whose traditional orthography is not in Latin characters are Romanized. Russian in particular is represented in an esoteric way that slightly deviates from any contemporary standardized system of Russian Romanization I have seen, disregarding palatal series such that *арти́кль* 'article (grammatical)' is represented as ARTIKL, where the character L is used to represent both Russian sequences <ль> (a palatalized lateral alveolar approximant) and <л> (the non-palatalized lateral alveolar approximant).

	Words in lexicon	in Dyen	linked in both
Dutch	120,876	233	183
English	39,632	200	136
French	43,586	201	164
German	50,682	207	161
Italian	9,293	228	168
Russian	42,980	289	164
Serbian	25,121	272	162
Spanish	22,558	216	159

Table 3: Words present in the various lexicons, in the Dyen list, and the number of items from the lexicon which were successfully linked to cognate groups

only on the domain of lexeme (and no lower), but for studies (such as the present thesis) that examine matters of sound, it cannot function as a standalone resource: its primary value is in as a relational table for joining input from different sources which contain language-specific information.

Consequently, the orthographic representations gathered from the lexical resources of each language were standardized into the encoding used in the Swadesh list, a 'lossy' format which reduces the information content of the item. The orthographic form for each word entry in a lexicon was converted into the esoteric format used in the Dyen list, and if a matching form was found in the Dyen list, this word (including its phonological representation from the lexical resource) was included in the study[12]. The number of words from the individual language dictionaries which were identified as a member of at least one cognate set in the Dyen list is as listed the final column of Table 3.

Manual verification of the cognacy status was necessary: for each of the 200 meanings in the Swadesh list, the Dyen list peculiarly in some cases provides several extant words for that concept in a given language, one of which is assured to be at the highest level of cognacy judgment (this level is specified for each cognate group as a whole), but which of these words is the intended member of that group is not transparent. For example, in (2), sample entries

[12]The parser I created to load the plaintext Dyen list file into a hierarchical representation will be available as an open-source Python package 'DyenParser' for use by other linguists working with the Dyen list

from three cognate subgroups are listed:

(2) Sample of raw rows for the meaning NECK

```
GERMAN HALS, NACKEN
FRISIAN HALS, NEKKE
...
FLEMISH HALS
...
NEW ENGLISH[13] NECK
```

The items inside each group are cognate with the other entries in that group, and the Dyen list further specifies that these three groups are confidently cognate with each other. While German *Nacken* 'nape' is cognate with English *neck*, German *Hals* 'neck' is not cognate with English *neck*. There is no indication in the structure of the Dyen list which of the two items provided in the German and Frisian rows is the basis for the cognacy judgment with the related groups—by default, parsing the Dyen list programmatically will result in a cognate relationship established between German *Hals* 'neck' and English *neck*, which is a false positive. To combat this, for the eight languages in my dataset, all of the extracted cognate pairs whose corresponding entries in the Dyen list featured multiple extant words were manually checked to verify that the correct item was identified as the reflex. In total, 78 such false-positives were excluded from the final data. The full table of pairwise cognacy is provided in Table 4.

	Dutch	English	French	German	Italian	Russian	Serbian
English	141						
French	53	53					
German	191	117	48				
Italian	53	54	172	49			
Russian	50	51	48	43	46		
Serbian	56	50	48	44	48	197	
Spanish	57	52	158	48	183	46	48

Table 4: Cognate lemmas in each lexicon identified via the Dyen wordlist which were included in the initial dataset ($n = 2,204$)

[13] As opposed to Old English.

Two transformations of linguistic consequence were applied to the representation of words in the Dyen list for some of the languages investigated in this study based on my familiarity with the morphological systems of these languages. There are likely other transformations which could be applied based on a more specialized knowledge of the morphological systems of the surveyed languages, but I attempted to prioritize data fidelity with these two notable exceptions which were applied to improve the number of forms successfully matched with corresponding entries in the dictionaries, as in many cases the form chosen to represent a cognate in the Dyen list was not the citation form chosen to represent the same lemma in the lexicons.

1. English verbs are stored as infinitives of the form TO GO. The sequence *to* was removed because verbal lemmas are represented with the bare infinitive in the CELEX2 English dictionary (but inflectional infinitive suffixes, such as German *sitz-en* (sit-INF, 'to sit'), were not adjusted since these forms are listed as the citation form in the other dictionaries)[14].

[14]Perhaps in a refinement of this study, the citation forms in lexicons studied could be stripped of *all* inflectional data to isolate stems like *sitz-* for German 'sit'. This is achievable for the CELEX2 Germanic dictionaries, which include such stems, but was not attempted for the present study because the citation forms in dictionaries are being used as a proxy for the presumedly cached storage of particular lexemes in the minds of real speakers. It is of course inaccurate to assume that headwords in a dictionary are a perfect representation of the cached storage of lexicalized forms in a speaker's mind, but this abstraction is an operational necessity for the purpose of this study. The choices made in data standardization are balanced between a need for consistency between forms in these languages—comparing apples to apples, so to speak—and the degree of egregiousness, in whatever sense, entailed in leaving a citation form as-is for the purpose of eventual comparison of phonological distance. In these three cases I have made subjective decisions about this balance: for example, I have decided that the infinitive marker in English *to sit* would should not be included—if *to* were included, in my analysis German /zɪtsən/ would be compared with something like English /təsɪt/, which seems like a comparison between unequals. While a similar concern applies to the comparison of English /sɪt/ with German /zɪtsən/ instead of the stem /zɪts/, I find that the balance of concerns is weighted differently: the desire for standardization is outweighed by the fact that this form is nonetheless present in the German dictionary and compared to the English *to* is more closely attached to the content word: compare *to boldly go*, where intervening material can occur and *to* functions as a proclitic, while no such intervention can occur between the verb and German *-en*. In other words, when there is an inflectional affix present in the citation form for a word in a dictionary, I assume there is 'a good reason for it'—perhaps these intuitions in lexicography suggests that such forms are lexicalized and therefore perhaps engage less of the compositional machinery of speech perception. However, this arbitrary decision should be compared with experimental evidence of the inverse assumption.

2. The separable reflexive marker on the end of Spanish verbs of the form *sentarse*[15] (sit-INF-REFL, 'to sit oneself [down]') was removed since the non-reflexive form of these verbs are still utilized in Spanish.

C. Selection and conversion of lexical resources

Once cognate identification was complete, the phonemic representations of the relevant words in the eight dictionaries were selected. The question of representation is intrinsic, and future iterations of this analysis should utilize a matrix of distinctive features (as in Marslen-Wilson 1987) rather than presuming all phonemes are equally 'different' from each other. The present methodology, however, does not yet incorporate this featural representation. Given, then, the choice between phonetic forms and phonemic forms, the use of phonetic forms was not possible for the resources selected (none of these resources contain any information about allophony). Additionally, phonemic representation provides a layer of abstraction that should reduce noise in the data and allow this study to focus on the most meaningful distinctions, allowing future work to incorporate an awareness of the fine phonetic distinctions known to affect cohort competition (e.g. Marslen-Wilson & Tyler 1980).

Two other important factors, syllabification and lexical stress, were also excluded from this study, even though word-level stress is important in the segmentation of words in continuous speech (Cutler 1989). Stress is transcribed discretely only in three of the eight resources (CELEX2), and syllabification only in four (CELEX2 and Lexique 3). There is immense potential explanatory value in including these features in this analysis: syllable prominence could perhaps explain any delay in uniqueness point. However, even though stress is not treated in this study as an independent character, some effects of stress are implicitly present in the data.

[15] This form illustrates an issue with the simplicity of the cognate judgments in the Dyen list. *Sentarse* is listed as cognate with French *asseoir* 'to sit'. Nichols (2014) regards this exact pair ultimately as non-cognate: despite sharing the root, the French form features a stem alternation and a prefix, material which is not cognate with the Spanish material. Ideally, all words in this data should reflect entirely cognate material. I did not attempt to apply this constraint to all 21,602 items on the Dyen list, but I adopt this viewpoint in the expansion of the Germanic dataset with derived forms (described below).

For example, when stripped of stress marking, CELEX2 English encodes *advertise* as /advə-taɪz/ but *advertisement* as /ədvɜtɪzmənt/[16]. While stress *per se* is not depicted, the difference between initial and post-initial stress is realized in a difference in vowel quality. Although this study does not enable the examination of stress as a discrete predictor, the effects of stress nonetheless are present in the data.

For each language, phonological representations were converted into an intermediary, arbitrary transcription system where one phoneme is represented by exactly one Unicode character. These choices have significant implications for comparison of Levenshtein distance (Levenshtein 1965), a measure of distance between two strings of text which I will describe in greater detail below. Additionally, coding these representations involves theoretical decisions motivated by the study of the phonological systems of the languages surveyed. Russian palatalized consonants like <ЛЬ> can be, and generally are, transcribed with two characters, as in IPA /lʲ/ (where the superscript /ʲ/ is meant to modify the preceding character, but nonetheless is still a separate character on a computer). However, since in Russian phonology the palatalized series of consonants comprises independent phonemes contrasting with a non-palatalized series of phonemes (Comrie 1990), for the purposes of machine comparison based on discrete changes in a string of phonemes, the phoneme written /lʲ/ should be transcribed as one character that is not composed of any other characters (λ might be a convenient choice due to the use of this character in IPA /ʎ/, but the character itself is arbitrary as long as it is different from other phonemes in the inventory). This constraint is necessary for the one-to-one comparison of phonemes among related languages[17]. These conversion processes were tailored for each lexicon.

The eight lexical resources are disparate in nature and intended purpose.

[16]In CELEX2, the 'primary' phonemic representation listed for a given English word, which were the forms selected in this study, are British English variants.

[17]Although the potential of using weighted Levenshtein distances which look at difference between the feature specification of phonemes (Sanders & Chin 2009; Schaefer 2016) is a valuable direction for future research: such an analysis would allow for incorporation of the knowledge that /ʎ/ is more similar to /l/ than to /k/, for instance.

1. CMU Sphinx (Italian, Russian, Spanish)

The CMU Sphinx (*CMU Sphinx* n.d.) dictionaries are provided as part of a text-to-speech synthesizer, and accordingly lack any detail other than orthographic form with corresponding pronunciation, and the data is not manually transcribed, likely leading to an artificial overabundance of regularity. For Russian, the wordlist consisted of inflected wordforms, so data not present in another wordlist of Russian lemmas (Zaliznjak 1977) were excluded, and the orthographic forms in Cyrillic were converted to an approximation of the representation of Russian words in the Dyen list[18]. For Spanish, the dictionary represented Peninsular (Castilian) Spanish. The CMU Sphinx phonemic data is stored in a style esoteric to each dictionary, but generally is more broadly transcribed than the manually-transcribed resources below, meaning that this data is potentially over-simplified compared to the narrower transcription in the other dictionaries.

2. vwr (Serbian)

The `vwr` (visual word recognition) package (Keuleers 2013) for the programming language R contains a wordlist for Serbian sorted for frequency of occurrence (Kostić 1999). To remove inflected forms, this wordlist was filtered for lemmas through the use of the srWaC web corpus of Serbian (Ljubešić & Klubička 2016), a resource which includes citation forms for tokens. Both sources contain only orthographic representations in Latin characters, so an automated conversion process was performed to yield the necessary phonemic representations. This process likely leads to the overestimation of regularity between orthographic form and phonemic form as with the CMU Sphinx data described above.

3. Lexique 3 (French)

The Lexique 3 (New, Pallier, & Ferrand 2005) dictionary contains French wordforms and their corresponding lemmas. The rows where the wordform was identical to the lemma were

[18]This conversion is not perfect, and there are unmatched Russian cognates in the Dyen list that likely are included in the CMU Sphinx Russian dictionary, but require manual evaluation to determine if they refer to the same word.

selected. Some degree of stemming is included in the resource, but no inclusion of derived forms (see the CELEX2 discussion below) was attempted since the morphological breakdowns in this resource were inconsistent (for example, *impossible* is decomposed into two morphemes *im-possible*, rather than the maximal breakdown of three morphemes *im-poss-ible*, and I am uncertain why). The phonemic representations are stored in SAMPA, a rich format for directly representing IPA characters.

4. CELEX2 (English, Dutch, German)

CELEX2 (Baayen, Piepenbrock, & Gulikers 1995) contains the highest degree of granularity in most relevant axes for this study. Phonemic representations are transcribed manually (particularly necessary for the irregularity of English orthographic-to-phonemic correspondence), parts-of-speech are represented, and morphological breakdowns allow for citation forms to be decomposed into content words and bound morphemes. The phonemic transcriptions are stored by default in the DISC format, where one phoneme equals exactly one Unicode character.

The analysis of phonological distance in this study does not utilize this awareness of morpheme boundaries. However, the presence of these distinctions allowed for the inclusion of additional derived forms from the CELEX2 dictionaries. Data from every other dictionary were limited to the lexemes which exactly matched the forms given in the Dyen list—for example, Italian *cane* 'dog' and French *chien* 'dog' were found cognate, but *canile* 'kennel' and French *chenil* 'kennel' would not be included in the dataset although they are built with the same cognate material (both root and derivational affix). This is a limitation of the other five dictionaries in this project, since ideally such forms should be included but ultimately were not, as such a task would require one of the following: (i) manual identification of all derived forms of cognate roots between each language (which was impractical given the scope of this study), (ii) the selection of different lexical databases (better available sources were not identified), or (iii) an additional algorithm to programmatically recommend other likely cognate forms. The third option was avoided due to the concern that the effects of selecting data based on segmental

similarity would overlap with the main analysis, which also is fundamentally based on segmental similarity. In other words, such a filter could bias the data towards forms which have greater similarity, leading the analysis to overestimate the average similarity among words.

For the CELEX2 data, a compromise was possible. Morphologically complex forms consisting of no more than one content morpheme were selected when that content morpheme was a member of a cognate set. Then, in a pairwise fashion, forms in other Germanic languages which (i) were built on that root, (ii) were marked with the same part-of-speech, and (iii) featured the same number of total morphemes were identified as potentially fit for comparison (e.g. words like German *mutterlos* 'motherless' and English *motherless* 'motherless'). Since the resultant set of candidates was of manageable size, these proposed analogous and cognate forms could be manually verified, leading to the inclusion of additional words as in Table 5, all of whose component morphemes were deemed to be cognate (a requirement of true cognacy per Nichols 2014). The alignment of morphemes was only a tool to assist in the process of manual selection of additional derived forms of cognates for inclusion in the dataset and was entirely absent in the algorithms used to calculate distance.

With the addition of this additional data, the full number of cognate pairs selected for analysis is listed in Table 6, with most cognates found between German and Dutch. Cognate identification was strongest within members of the same subfamily, as shown in Table 7.

Language Pair		Derived	Stems	Total
Dutch	English	37	141	178
English	German	42	117	159
German	Dutch	313	191	504
Total		392	449	841

Table 5: Cognate pairs for Germanic languages in initial dataset, including discovery of derived forms of entirely cognate material

	Dutch	English	French	German	Italian	Russian	Serbian
English	178						
French	53	53					
German	504	159	48				
Italian	53	54	172	49			
Russian	50	51	48	43	46		
Serbian	56	50	48	44	48	197	
Spanish	57	52	158	48	183	46	48

Table 6: All cognate pairs identified ($n = 2,596$)

	Germanic	Romance	Balto-Slavic[19]
Germanic	841		
Romance	467	513	
Balto-Slavic	294	284	197

Table 7: Cognate pairs by family

[19]Since only two languages are represented in the Balto-Slavic data (Russian and Serbian), there is a lower raw number of cognates here. Proportionate to the number of languages compared between each family, most Balto-Slavic cognates were paired with other Balto-Slavic cognates.

D. Generation of uniqueness points

Once the data from all of these dictionaries was converted to the appropriate phonemic representations featuring a one-to-one mapping between character and phoneme, the points of uniqueness were calculated within each lexicon. For each word in a lexicon, the phonemic representation was iterated over, phoneme-by-phoneme, beginning with the first phoneme. This subset of phonemic material was then compared with the initial phoneme in all other words in the lexicon of that language. If any other word contained this phoneme, the algorithm expanded the search to include the next phoneme in the target word and again searched through the rest of the lexicon to determine if any words started with this sequence. This process was repeated until either (a) the word boundary was reached, meaning that the word featured no sublexical region which was unique in the lexicon[20], or (b) no other words in the lexicon contained the sequence. If the latter occurred, the point at which the word became unique in the lexicon was stored for the analysis described below.

In this way, words were divided into two regions described at the outset of this paper: the subword before the uniqueness point (the UNIQUE REGION), which the cohort model proposes is the most necessary for early lexical access, and the subword after the uniqueness point (the REMAINDER REGION), which is processed with different mechanisms and is not involved in the initial decision of candidate selection for lexical access. It is this second region which I predict will feature a greater degree of sound change compared to the first region.

E. Comparison of forms

To operationalize the difference between strings of phonemes between the members of each cognate pair, the diagnostic of Levenshtein distance (Levenshtein 1965) was used. Levenshtein distance is a measure of distance which compares the number of insertions, deletions, and alterations of characters necessary to make one string of text into another. As such, it quantifies

[20]These words are either homophones or words whose phonemic representation is initially-embedded in other items in the lexicon, like *bar* /bɑɹ/ is embedded in *barn* /bɑɹn/ in General American English.

the distance between two sequences as the minimal number of differences between discrete characters. See (3) for example values computed for simple orthographic forms. In linguistics, the measure has been variously used, for example, in the quantification of genetic relationship (Wichmann et al. 2010), pronunciation differences (Wieling et al. 2014), and phonological distance (Sanders & Chin 2009).

(3a) $$\text{lev}(hat, hat) = 0$$

(3b) $$\text{lev}(cat, hat) = 1$$

(3c) $$\text{lev}(at, hat) = 1$$

(3d) $$\text{lev}(at, hats) = 2$$

For this analysis, Levenshtein distance was applied to measure the difference between strings of phonemic representations. For each cognate pair whose members both featured a uniqueness point (at least one of which was non-final[21]), I calculated the following Levenshtein distances:

1. the unique region of the first word u_1 and the unique region of the second word u_2, divided by the average length of the unique regions (4a)
2. the remaining region of the first word r_1 and the remaining region of the second word r_2, divided by the average length of the remaining regions (4b)
3. the total of these two Levenshtein distances, representing the total between the cognate pair, divided by the average length of the words entirely w (4c)

[21] This constraint is to account for the cases in which two cognates both feature a uniqueness point at the end of the word, meaning the remainder region is mere silence for both words. For example, Spanish *oreja*|# 'ear' and English *ear*|# were both found to be unique (marked with |) at the final word boundary (#), so this pair was excluded since there is nothing to compare in the remainder region. However, Spanish *oreja*|# 'ear' and Italian *orrec*|*chio* 'ear' were found to be cognate as well. Since the Italian form features a uniqueness point before the final word boundary, this pair was included. (Note that English *ear* was found unique despite the existence of words like *earring* since CELEX2, as it documents British English pronunciations, distinguishes the /ɪ/ in *earring* from theorized underlying linking rhotics in words like *ear* that are non-rhotic in isolation.)

(4a) $$dist_u = \frac{\text{lev}(u_1, u_2)}{\left(\frac{\text{len}_{u_1} + \text{len}_{u_2}}{2}\right)}$$

(4b) $$dist_r = \frac{\text{lev}(r_1, r_2)}{\left(\frac{\text{len}_{r_1} + \text{len}_{r_2}}{2}\right)}$$

(4c) $$dist_p = \frac{\text{lev}(u_1, u_2) + \text{lev}(r_1, r_2)}{\left(\frac{\text{len}_{w_1} + \text{len}_{w_2}}{2}\right)}$$

Note that the length of each of the regions in each computation of the Levenshtein distance are not necessarily the same: the uniqueness point for a word relies on where it becomes unambiguous in the rest of that language's lexicon, and this point is different even among cognates. Thus, each of these two Levenshtein distances was divided by the average length of the relevant region in the pair of cognates as reflected in each formulae in (4). Dividing by average length was a technique for normalization accounting for the length of each region and of the word in general.

The reader may wonder why (4c) is as written instead of as simply $\text{lev}(w_1, w_2)$. The sum of the Levenshtein distances for each of the two regions in the cognate pair is not necessarily equal to the Levenshstein distances for the word (5).

(5) $$\text{lev}(u_1, u_2) + \text{lev}(r_1, r_2) \neq \text{lev}(w_1, w_2)$$

For instance, consider the pair of cognates German *Schlechtigkeit* and Dutch *slechtigheid* 'badness'. Their phonemic representations are (`SlExtIxkWt`) and (`slExt@xhKt`) respectively[22]. The raw Levenshtein distance between these two strings is:

(6) $$\text{lev}(\texttt{SlExtIxkWt}, \texttt{slExt@xhKt}) = 4$$

[22]Recall that these are encoded in the one-to-one scheme I have used for this study, in which the form chosen to represent a particular phoneme is mostly arbitrary: I am not suggesting German has a labialized back vowel or that Dutch has lost a nucleus in the ultimate syllable of this word. The importance is that phoneme (`W`), whatever it is, is distinct from phoneme (`K`) in Dutch (these happen to correspond to /ai/ and /ɛɪ/ respectively, according to the CELEX2 dictionary).

However, when the uniqueness point is identified for each word with respect to the lexicon it belongs to, these strings are bifurcated (characters that are not shared are underlined):

(7)

u	r
SlExtI	xkWt
slExt@x	hKt

The Levenshtein distances are calculated as in (4), such that lev$(u_1, u_2) = 3$ and lev$(r_1, r_2) = 3$, the sum of which is greater than the value obtained from comparing the entire word lev(w_1, w_2). Thus in my methodology, there are six total sound changes in this cognate pair, three in the unique region and three in the remainder. Once these raw values are divided by the lengths of their respective regions, we have $dist_u = 0.4615$, $dist_r = 0.8571$, and $dist_p = 0.6$, showing that that the words in this cognate pair are relatively more dissimilar to each other in the remainder region than in the unique region ($dist_r > dist_u$).

Levenshtein distance applied to phonemic representations of cognate material can function as a diagnostic for sound change. If the phonemic representations of homologous regions in two cognates are dissimilar, historical change must have applied. Thus $dist_u$ corresponds to change localized to the initial region of uniqueness of a word, $dist_r$ to change in the region after the uniqueness point, and $dist_p$ to the change between the two words entirely.

IV. Results

The primary hypothesis, as a reminder, is that across all sampled cognate pairs featuring uniqueness points (at least one of which is non-word-final), $dist_r$ will be greater than $dist_u$; in other words, that the degree of sound change in the region after the uniqueness point of the word will be greater than the degree of sound change in the region before. The null hypothesis is that there is no difference in degree of sound change between these regions.

As pictured in Table 8, 10.1310% of identified cognate pairs consisted of words which each had a uniqueness point in its language (at least one of them before the final boundary).

	Dutch	English	French	German	Italian	Russian	Serbian
English	14						
French	0	0					
German	139	15	2				
Italian	1	2	17	3			
Russian	0	0	1	1	5		
Serbian	0	0	5	1	10	20	
Spanish	2	2	5	0	16	1	1

Table 8: Cognate pairs whose members feature uniqueness points in their lexicons (at least one non-word-finally) ($n = 263$)

For these cognate pairs, the difference in distributions between the two relevant regions divided by the uniqueness point was normally distributed[23], and samples were regarded as paired due to the fact that the regions being compared came from the same words. Consequently, the primary hypothesis ($dist_r > dist_p$) was evaluated using a one-tailed t-test for paired samples. This test shows that the Levenshtein distance in the region after the uniqueness point is greater than that of the region before the uniqueness point ($t = 16.6145$, $p < .0001$) and that distance in the remainder region is greater than in the word overall ($t = 16.2395$, $p < .0001$). The unique region of the word does not show this relationship with the word overall ($t = -14.4926$, $p < .0001$). Thus the null hypothesis can be rejected[24].

[23] Per D'Agostino & Pearson 1973.

[24] Due to the evidence supporting the primary hypothesis, the secondary two-tailed hypothesis mentioned in Background is not evaluated.

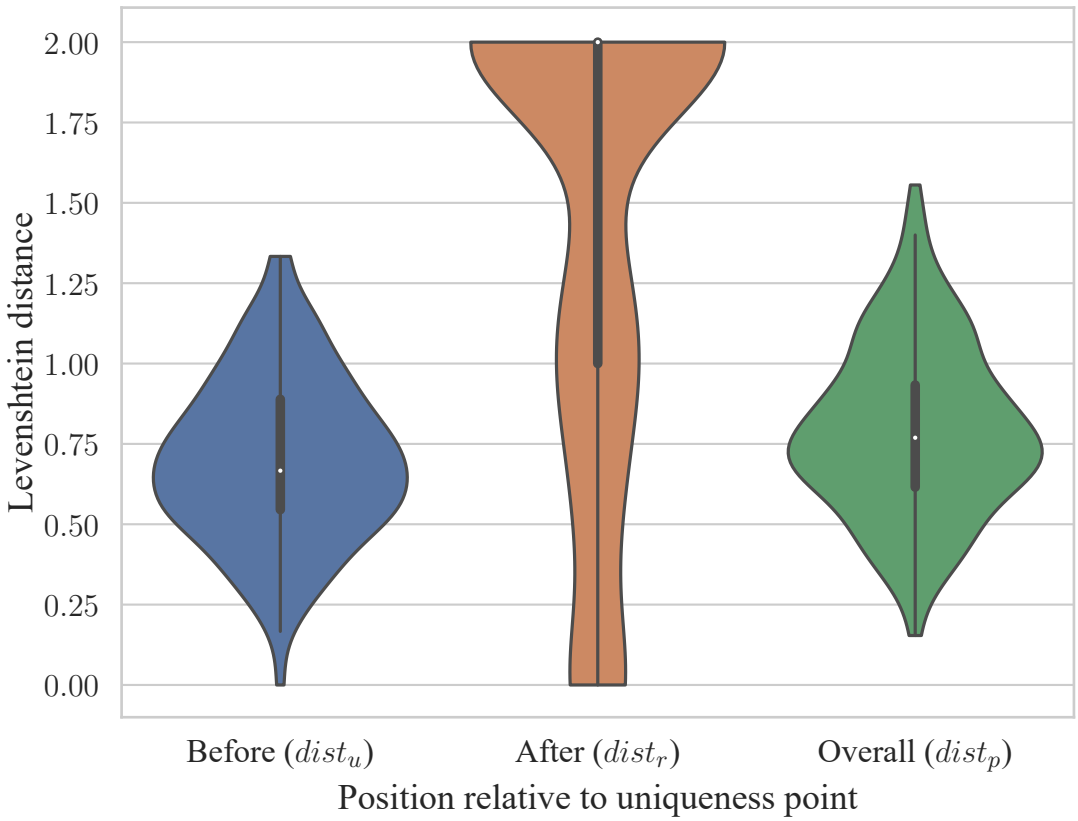


Figure 3: Distributions of the regions of the cognate pairs based on Levenshtein distance over average region length

$dist_r > dist_u$: $t = 16.6145$, $p < .0001$
$dist_r > dist_p$: $t = 16.2395$, $p < .0001$
$dist_u$: $mean = .6980$, $sd = .2564$
$dist_r$: $mean = 1.4403$, $sd = .7258$
$dist_p$: $mean = .7899$, $sd = .2664$

The distributions of this data are visualized in the violin plot in Figure 3, where the three shapes represent the distribution of Levenshtein distances of (i) the unique region (pre-uniqueness point), (ii) the remainder region (post-uniqueness point), and (iii) the word (as sum of the two regions). The width of each shape corresponds to the estimated probability that a Levenshtein distance in the dataset will have the corresponding y value, with the height of each shape representing the range of distances observed in each region. The distributions of $dist_u$ and $dist_p$ are more centralized while the $dist_r$ shows a great variety of distances across its greater range.

The hourglass-shaped increase in observations approaching the maximum 2.00 of $dist_r$ represents the fact that one of the words in the cognate set features no remainder region (ie, words that are unique only at the word-final boundary). There are many words in the sample which feature such a pattern—for example, German *mutterlos* 'motherless', which becomes unique at <o> /o/ and English *motherless*, which becomes unique at <ss> /s/—but due to the normalization mechanism of average region length, these words are perhaps penalized too harshly. *Motherless* has no remainder region since it becomes unique at the word boundary, so the length of remainder region is coded as 0. The raw Levenshtein distance of 1 for the *mutterlos*/*motherless* pair is divided by the average length of the remainder region, yielding $\text{avg}(0, 1) = .5$. This value is converted into a score of $\frac{1}{.5} = 2$. However, a paired region with the same raw Levenshtein distance receives a much lower transformed score of distance if both words feature non-silent remainder segments. For example, the (distant) cognate pair Spanish *cuando* 'when' and Dutch *wanneer* 'when' become unique at <d> /ð/ and <ee> /e/, respectively. Their remainder regions, then, are *o* and *r*, a raw Levenshtein distance of 1 which is transformed into 1 since both remainder regions are non-zero.

The average-based transformation, then, penalizes deletion quite more harshly than modification (when sequences are merely modified instead of deleted or added to, the lengths stay the same). Such a penalty might not be indefensible—perhaps deletion could be considered quantifiably 'more' of a sound change than lenition—but I do not intend to play favorites with types of sound change in this study, and this methodology would need substantial modification to assign weight to particular types of sound change.

That said, many pairs do feature zero-length remainder regions, so the clustering around this point is worth disentangling from this transformation issue. To control for this unintended effect of average-based transformation, a more traditional normalized measure of Levenshtein distance (Petroni & Serva 2009; Wieling et al. 2014) was next calculated, using division by the greater of the two lengths (max length) as the transformation. This measure mitigates the effect of word length when comparing Levenshtein distances across cognate pairs to each other,

which was the motivation for my initial inclusion of normalization by average length of paired region, but does not penalize deletion in the same way as my measure. Ultimately, the choice of normalization did not affect the conclusion: the results of these measures are also significant in showing a slightly smaller effect of $dist_r > dist_u$ ($t = 7.4724$, $p < .0001$) and $dist_r > dist_p$ ($t = 7.0731$, $p < .0001$), and Figure 4 still demonstrates a centralization around high Levenshtein distances, but does not weight the zero-length remainder regions as significantly, resulting in a more even distribution overall.

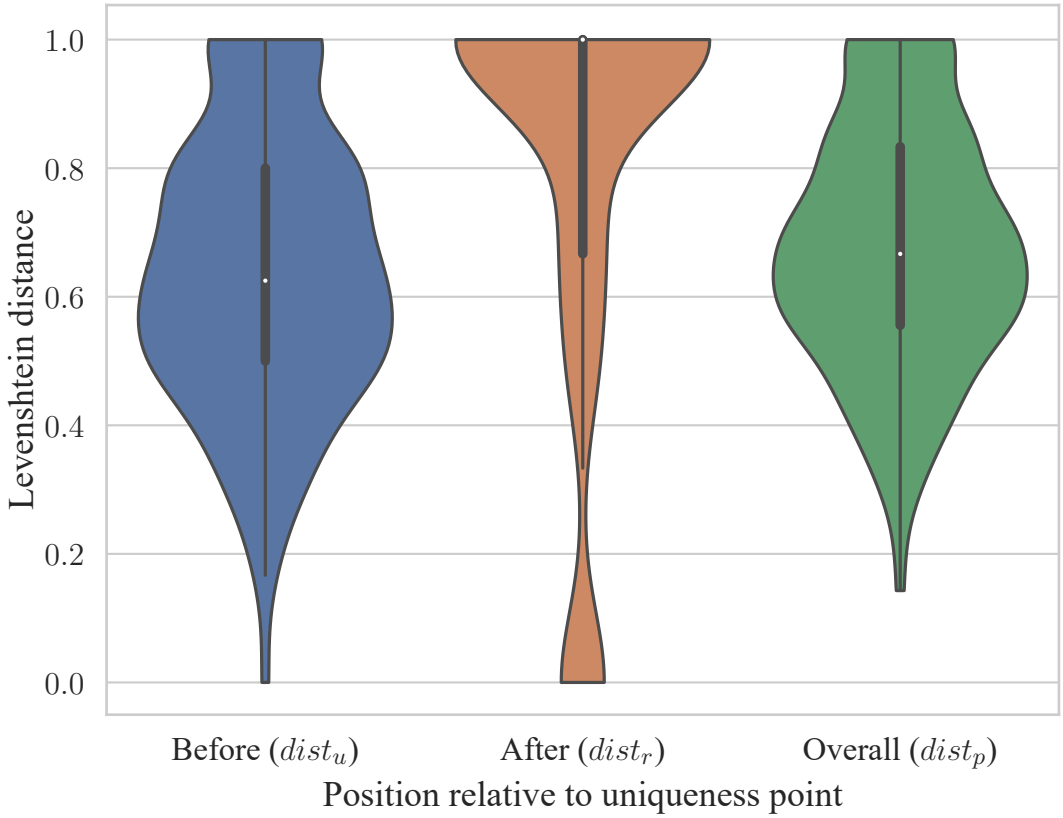


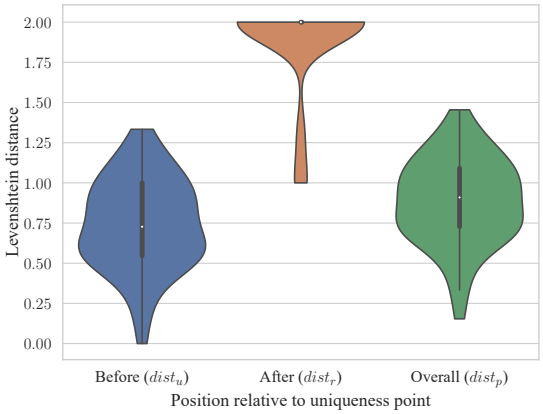
Figure 4: Distributions of the regions of the cognate pairs based on Levenshtein distance over max region length

$dist_r > dist_u$: $t = 7.4724$, $p < .0001$
$dist_r > dist_p$: $t = 7.0731$, $p < .0001$
$dist_u$: $mean = .6383$, $sd = .2156$
$dist_r$: $mean = .8093$, $sd = .3348$
$dist_p$: $mean = .6744$, $sd = .1916$

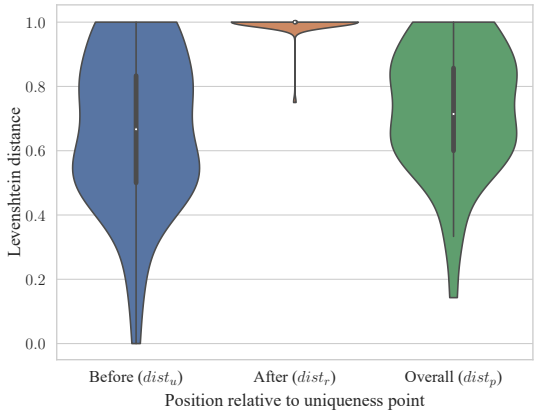
However, the median for $dist_r$ still lies at the maximum of the range, suggesting ultimately that when sound change occurs in a remainder region, it is generally catastrophic sound change.

These statistical tests were applied to subsets of the data to examine if the effects seen were indeed across all eight languages or if they were due to the over-representation of cognates from the Germanic dataset. The CMU Sphinx and vwr dictionaries of Italian, Spanish, Russian, and Serbian (I call this set ISRS) form a group based on the fact that the phonemic transcriptions for these languages were automatically generated. While the Germanic dataset features rich, manual transcription and includes additional derived words of fully cognate material, the ISRS group is automatically transcribed at a broad level and features only the lemmas included in the Dyen list.

The same t-tests as above were conducted on the ISRS data. The effects were in the same direction and even stronger. The distributions are shown in Figure 5 below. While the conclusions are the same, the distribution of remainder region is much more restricted in range (compare Figure 5 to Figures 3 & 4 above).



a) by average region length
$dist_r > dist_u$: $t = 20.1082$, $p < .0001$
$dist_r > dist_p$: $t = 9.5341$, $p < .0001$
$dist_u$: $mean = 0.7412$, $sd = 0.2834$
$dist_r$: $mean = 1.8843$, $sd = 0.3044$
$dist_p$: $mean = 0.8854$, $sd = 0.2690$

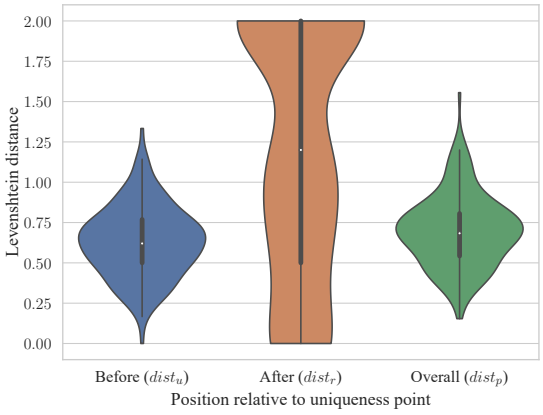


b) by max region length
$dist_r > dist_u$: $t = 9.5341$, $p < .0001$
$dist_r > dist_p$: $t = 9.4606$, $p < .0001$
$dist_u$: $mean = 0.6702$, $sd = 0.2393$
$dist_r$: $mean = 0.9953$, $sd = 0.0343$
$dist_p$: $mean = 0.7323$, $sd = 0.1964$

Figure 5: Distributions for Italian, Spanish, Russian, and Serbian (ISRS)

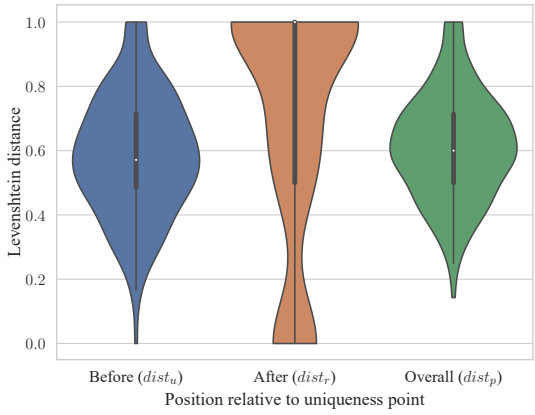
This unusual shape is a byproduct of data paucity: the average length of the remainder region is only .7736 phonemes for the ISRS languages compared to 1.3065 for the Germanic languages. This difference is a direct consequence of the inability to establish cognacy between words derived from the lemmas provided in the Dyen list for any languages but English, Dutch, and German. The inclusion of derived words for the Germanic data results in far more information for study after the uniqueness point, while the uniqueness points in the ISRS data are typically followed by at most one phoneme before the final word boundary. This finding, then, confirms the value of including derived words where possible, as discussed at length in the Methodology section above.

A concern from the ISRS distributions is that perhaps the very high average Levenshtein distance in the remainder region is responsible for this effect in the full dataset. Therefore, the Germanic data was examined in isolation as well.



a) by average region length

$dist_r > dist_u$: $t = 9.3211$, $p < .0001$
$dist_r > dist_p$: $t = 9.1532$, $p < .0001$
$dist_u$: $mean = .6305$, $sd = .2215$
$dist_r$: $mean = 1.2049$, $sd = .7820$
$dist_p$: $mean = .6996$, $sd = .2257$



b) by max region length

$dist_r > dist_u$: $t = 3.6000$, $p < .001$
$dist_r > dist_p$: $t = 3.2800$, $p < .001$
$dist_u$: $mean = .5856$, $sd = .1915$
$dist_r$: $mean = .7029$, $sd = .3795$
$dist_p$: $mean = .6136$, $sd = .1673$

Figure 6: Distributions for Germanic languages

The tests run solely on the Germanic dataset (Figure 6) support the conclusions drawn from the full dataset, although the effect and p-values for the values normalized by max region length are somewhat lower (Figure 6b) than in the other tests. The hypothesis $dist_r > dist_u$ seems to

hold at multiple levels of grouping, including the entire eight language set. Since these three dictionaries consisted of more accurate manually-transcribed phonemic representations of each word, the results for the Germanic subset of data suggest that the effects seen for the larger dataset are not a byproduct of the automated orthographic-to-phonemic conversion performed for the ISRS dictionaries nor of the relative lack of information contained in the ISRS remainder regions.

V. Conclusion

The results of this study show that the uniqueness point can divide cognates into two regions which feature unequal degrees of sound change, and that the region which features the greater degree of sound change is not involved in cohort competition for lexical access. This finding suggests that for the sampled Indo-European data, sound change varies asymmetrically within the domain of the lexical item, and that the greatest degree of change is localized towards the end of the word. This distinction is consistent with research suggesting that phonemic material with low functional load (Blevins & Wedel 2009; Wedel, Jackson, & Kaplan 2013) and predictable units of sublexical meaning (Blevins 2005) are more likely to undergo certain types of language change: it appears, too, that the parts of words not necessary for early recognition in lexical access are implicated in this propensity for change, although I do not quantify the extent to which the sound changes localized to the remainder region are specifically reductions or mergers.

While this study provides solid evidence of the existence of a highly significant asymmetry in the degree of sound change with respect to the position within the word, some limitations must be resolved before the more specific claim about the uniqueness point as a significant predictor can be evaluated. The suggested outcomes now need to be scrutinized to confirm that incorporation of the uniqueness point is a necessary addition to the model and that the conclusions shown cannot simply be explained by looking at whether these cognates on average tend to increase in distance as they get longer. I cannot say from the present results that the cohort model is responsible for this pattern. Artificially jittering the uniqueness point one segment to the right and one segment to the left resulted still in a significant effect, but this effect was both diminished and less significant, and disappeared entirely once the point was moved even further away, suggesting that there is some importance in the location of the uniqueness point. A more in depth analysis could investigate the clusters of sound change within a word, entirely agnostic to the theoretical construct of the uniqueness point, and the results of this naive analysis could be compared with the present study to see if the boundaries of any emergent clusters align with

the uniqueness point of the word.

The dataset assembled here allows such research to be conducted. Future work on this dataset would involve the refinement of the existing cognate decisions and the retrieval of derived cognates for all languages, not just the Germanic languages whose data format facilitated these decisions. Additionally, the incorporation of prosodic information (stress and syllable), as well as an evaluation of the phonological inventories and their historical relationships to their counterparts could provide a greater degree of granularity, as could the introduction of featural Levenshtein distances. Another dimension to include in future modeling is the frequency of the words surveyed, a factor theorized to influence both the diffusion of sound changes across a lexicon (Pierrehumbert 2001; Bybee 2007; Hay et al. 2015) and the scalar activation of candidates in a cohort, an element not part of the earliest formulations of the cohort model but present in essentially all subsequent iterations of the model (Taft & Hambly 1986; Marslen-Wilson 1987). Incorporation of this measure in particular would serve both the cognitive and historical-comparative functions of this study.

The asymmetry of sound change within the word observed in these eight Indo-European languages is not by chance and does not seem to be constrained to any one subfamily. Ideally, this study should be reproduced on larger and more diverse datasets, within Indo-European and without, to describe the cross-linguistic extent of the effect. However, the present study provides little in the way of evidence for similar processes outside of the Indo-European language family, especially for languages which are highly polysynthetic. Much as how the research on cognitive processing of English speech could not *a priori* describe the neurological processes involved in processing morphologically-complex forms in Polish, where no noun seems to be uninflected cognitively (Szlachta et al. 2012), these findings from Indo-European cannot be generalized outside of this dataset, and applications of this model to other data must be preceded by synchronic study of the representation of words within the minds of speakers of the relevant languages.

References

Baayen, R. Harald, Richard Piepenbrock, & Léon Gulikers (eds.). 1995. *CELEX2 LDC96L14*. Philadelphia: Linguistic Data Consortium.

Bickel, Balthasar & Fernando Zúñiga. 2017. The 'word' in polysynthetic languages: phonological and syntactic challenges. In *The Oxford Handbook of Polysynthesis*, Michael Fortescue, Marianne Mithun, & Nicholas Evans (eds.), pp. 158–185. Oxford: Oxford University Press.

Blevins, Juliette. 2005. The role of phonological predictability in sound change: privileged reduction in Oceanic reduplicated substrings. *Oceanic Linguistics* 44(2). 517–526.

Blevins, Juliette & Andrew Wedel. 2009. Inhibited sound change: an evolutionary approach to lexical competition. *Diachronica* 26(2). 143–183.

Bozic, Mirjana, Lorraine K. Tyler, David T. Ives, Billi Randall, & William D. Marslen-Wilson. 2010. Bihemispheric foundations for human speech comprehension. In *Proceedings of the National Academy of Sciences*, pp. 17439–17444.

Bozic, Mirjana, Lorraine K. Tyler, Li Su, Cai Wingfield, & William D. Marslen-Wilson. 2013. Neurobiological systems for lexical representation and analysis in English. *Journal of Cognitive Neuroscience* 25(10). 1678–1691.

Bruhn, Daniel, John Lowe, David Mortensen, & Dominic Yu. 2015. *Sino-Tibetan Etymological Dictionary and Thesaurus Database Software (STEDT)*.

Bybee, Joan L. 2007. *Frequency of Use and the Organization of Language*. Oxford: Oxford University Press.

CMU Sphinx (n.d.). URL `https://cmusphinx.github.io/`.

Cole, Ronald A. & Jola Jakimik. 1980. A model of speech perception. In *Perception and Production of Fluent Speech*, Ronald A. Cole (ed.), pp. 133–164. Hillsdale, NJ: Lawrence Erlbaum.

Comrie, Bernard. 1990. Russian. In *The World's Major Languages*, Bernard Comrie (ed.), pp. 329–347. Oxford: Oxford University Press.

Cutler, Anne. 1989. Auditory lexical access: where do we start? In *Lexical Representation and Process*, William D. Marslen-Wilson (ed.), pp. 342–356. Cambridge: MIT Press.

Cutler, Anne & Hsuan-Chih Chen. 1997. Lexical tone in Cantonese spoken word processing. *Perception and Psychophysics* 59. 165–179.

D'Agostino, Ralph & E. S. Pearson. 1973. Tests for departure from normality. Empirical results for the distributions of b_2 and $\sqrt{b_1}$. *Biometrika* 60(3). 613–622.

Dryer, Matthew S. 2013. *Prefixing vs. Suffixing in Inflectional Morphology*. Ed. by Matthew S. Dryer & Martin Haspelmath. URL `https://wals.info/chapter/26`.

Dunn, Michael. 2012. *Indo-European lexical cognacy database (IELex)*. Max Planck Institute for Psycholinguistics. URL `http://ielex.mpi.nl`.

Dyen, Isidore, Joseph Kruskal, & Paul Black. 1992. An Indoeuropean classification: a lexicostatistical experiment. *Transactions of the American Philosophical Society* 82(5). 1–132.

Ford, Michael A., Matt H. Davis, & William D. Marslen-Wilson. 2010. Derivational morphology and base morpheme frequency. *Journal of Memory and Language* 63. 117–130.

Gaskell, M. Gareth & William D. Marslen-Wilson. 1997. Integrating form and meaning: a distributed model of speech perception. *Language and Cognitive Processes* 12(5-6). 613–656.

Hall, Christopher. 1988. Integrating diachronic and processing principles in explaining the suffixing preference. In *Explaining Language Universals*, John A. Hawkins (ed.), pp. 321–349. Oxford: Blackwell.

Haspelmath, Martin. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* 45(1). 31–80.

Hawkins, John A. & Anne Cutler. 1988. Psycholinguistic factors in morphological asymmetry. In *Explaining Language Universals*, John A. Hawkins (ed.), pp. 280–317. Oxford: Blackwell.

Hay, Jennifer B., Janet B. Pierrehumbert, Abby J. Walker, & Patrick LaShell. 2015. Tracking word frequency effects through 130 years of sound change. *Cognition* 139. 83–91.

Jurafsky, Daniel, Alan Bell, Michelle Gregory, & William D. Raymond. 2001. Probabilistic relations between words: evidence from reduction in lexical production. In *Typological studies in language, Vol. 45. Frequency and the Emergence of Linguistic Structure*, Joan Bybee & Paul Hopper (eds.), pp. 229–254. Amsterdam: John Benjamins.

Keuleers, Emmanuel. 2013. *Useful functions for visual word recognition research*. URL `https://cran.r-project.org/web/packages/vwr/vwr.pdf`.

Klatt, Dennis H. 1989. Review of selected models of speech perception. In *Lexical Representation and Process*, William D. Marslen-Wilson (ed.). Cambridge: MIT Press.

Kostić, Đorđe. 1999. *Frequency Dictionary of Contemporary Serbian Language (Frekvencijski rečnik savremenog srpskog jezika)*. Yugoslavia: University of Belgrade, Institute for Experimental Phonetics, Speech Pathology, & Laboratory for Experimental Psychology.

Lee, Chao-Yang. 2007. Does Horse activate Mother? Processing lexical tone in form priming. *Language and Speech* 50(1). 101–123.

Levenshtein, Vladimir I. 1965. Binary codes capable of correcting deletions, insertions and reversals (Двоичные коды с исправлением выпадений, вставок и замещений символов). *Proceedings of the USSR Academy of Sciences (Доклады Академий Наук СССР)* 163(4). 845–848.

Ljubešić, Nikola & Filip Klubička. 2016. *Serbian web corpus srWaC 1.1*. Slovenian language resource repository CLARIN.SI. URL `http://hdl.handle.net/11356/1063`.

Longworth, Catherine E., William D. Marslen-Wilson, Billi Randall, & Lorraine K. Tyler. 2005. Getting to the meaning of regular past tense: evidence from neuropsychology. *Journal of Cognitive Neuroscience* 17(7). 1087–1097.

Luce, Paul. 1986. A computational analysis of uniqueness points in auditory word recognition. *Perception & Psychophysics* 39(3). 155–158.

Malins, Jeffrey & Marc Joanisse. 2012. Setting the tone: An ERP investigation of the influences of phonological similarity on spoken word recognition in Mandarin Chinese. *Neuropsychologia* 50(8). 2032–2043.

Marslen-Wilson, William D. 1984. Function and process in spoken word recognition. In *Attention & Performance X.* H. Bouma & D. Bouwhuis (eds.), pp. 125–150. Lawrence Erlbaum Associates.

— 1987. Functional parallelism in spoken word-recognition. *Cognition* 25(1). 71–102.

Marslen-Wilson, William D., Mirjana Bozic, & Lorraine K. Tyler. 2014. Morphological systems in their neurobiological contexts. In *The Cognitive Neurosciences*, Michael S. Gazzaniga & George R. Mangun (eds.). 5th edn. Cambridge, MA: MIT Press.

Marslen-Wilson, William D., Mary Hare, & Lianne Older. 1993. Inflectional morphology and phonological regularity in the English mental lexicon. In *Proceedings of the 15th Annual Meeting of the Cognitive Science Society*. Princeton, NJ: Erlbaum.

Marslen-Wilson, William D. & Lorraine K. Tyler. 1980. The temporal structure of spoken language understanding. *Cognition* 1. 1–71.

— 1998. Rules, representations, and the English past tense. *Trends in Cognitive Science* (2). 428–435.

Marslen-Wilson, William D. & Alan Welsh. 1978. Processing interactions and lexical access during word-recognition in continuous speech. *Cognitive Psychology* 10. 29–63.

New, Boris, Christophe Pallier, & L. Ferrand (eds.). 2005. *Lexique 3*. URL `http://www.lexique.org/`.

Nichols, Johanna. 2014. Derivational paradigms in diachrony and comparison. In *Paradigm Change: in the Transeurasian Languages and Beyond*, Martine Robbeets & Walter Bisang (eds.), ch. 3. (Studies in Language Companion Series 161). Amsterdam: John Benjamins.

Petroni, Filippo & Maurizio Serva. 2009. Automated words stability and languages phylogeny. *CoRR* abs/0911.3292.

Pierrehumbert, Janet B. 2001. Exemplar dynamics: word frequency, lenition, and contrast. In *Frequency Effects and the Emergence of Lexical Structure*, Joan L. Bybee & Paul Hopper (eds.), pp. 137–157. Amsterdam: John Benjamins.

Poss, Nick, Tsun-Hui Hung, & Udo Will. 2008. The effects of tonal information on lexical activation in Mandarin. *Proceedings of the 20th North American Conference on Chinese Linguistics* 1. 205–211.

Post, Brechtyje, William D. Marslen-Wilson, Billi Randall, & Lorraine K. Tyler. 2008. The processing of English regular inflections: phonological cues to morphological structure. *Cognition* 109 (1). 1–17.

Radeau, Monique, Philippe Mousty, & Paul Bertelson. 1989. The effect of the uniqueness point in spoken-word recognition. *Psychological Research* (51). 123–128.

Ringe, Don, Tandy Warnow, & Ann Taylor. 2002. Indo-European and computational cladistics. *Transactions of the Philological Society* 100(1). 59–129.

Sanders, Nathan C. & Steven B. Chin. 2009. Phonological distance measures. *Journal of Quantitative Linguistics* 16(1). 96–114.

Schaefer, Kevin Jay. 2016. *Machine learning in language reconstruction: A-Star models of sound change*. University of California, Santa Barbara MA thesis.

Schiering, René, Balthasar Bickel, & Kristine A. Hildebrandt. 2010. The prosodic word is not universal, but emergent. *Journal of Linguistics* 46(3). 657–709.

Stone, Gerald. 1990. Polish. In *The World's Major Languages*, Bernard Comrie (ed.), pp. 348–366. Oxford: Oxford University Press.

Swadesh, Morris. 1952. Lexicostatistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society* 96. 452–463.

Szlachta, Zanna, Mirjana Bozic, Aleksandra Jelowicka, & William D. Marslen-Wilson. 2012. Neurocognitive dimensions of lexical complexity in Polish. *Brain Language* 121(3). 219–225.

Taft, Marcus & Gail Hambly. 1986. Exploring the cohort model of spoken word recognition. *Cognition* 22(3). 259–282.

Wedel, Andrew, Scott Jackson, & Abby Kaplan. 2013. Functional load and the lexicon: evidence that syntactic category and frequency relationships in minimal lemma pairs predict the loss of phoneme contrasts in language change. *Language and Speech* 56(3). 395–417.

Wichmann, Søren, Eric W. Holman, Dik Bakker, & Cecil H. Brown. 2010. Evaluating linguistic distance measures. *Physica A: Statistical Mechanics and its Applications* 389(17). 3632–3639.

Wieling, Martijn, John Nerbonne, Jelke Bloem, Charlotte Gooskens, Wilbert Heeringa, & R. Harald Baayen. 2014. A cognitively grounded measure of pronunciation distance. *PLoS ONE* 9(1). e75734.

Zaliznjak, Andrey. 1977. *A Grammatical Dictionary of the Russian Language*. Ed. by Andrei Usachev & Sergei Starostin. URL `http://www.smo.uhi.ac.uk/~oduibhin/russian/rwordlist.htm`.