QUANTITATIVE INVESTIGATIONS IN HUNGARIAN PHONOTACTICS AND SYLLABLE STRUCTURE

Stephen M. Grimes

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Linguistics,
Indiana University
September 2010

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

_____

(Stuart Davis, Ph.D.)

_____

(Kenneth de Jong, Ph.D.)

_____

(Sandra Kübler, Ph.D.)

_____

(Markus Dickinson, Ph.D.)

January 30, 2009

Stephen Grimes

QUANTITATIVE INVESTIGATIONS IN HUNGARIAN PHONOTACTICS AND SYLLABLE STRUCTURE

This dissertation investigates statistical properties of segment collocation and syllable geometry of the Hungarian language. A corpus and dictionary based approach to studying language phonologies is outlined. In order to conduct research on Hungarian, a phonological lexicon was created by compiling existing dictionaries and corpora and using a system of regular expression rewrite rules (based on letter-to-sound rules) in order to derive pronunciations for each word in the lexicon. The resulting pronunciation dictionary contains not only pronunciations in several transcription systems but also syllable counts and syllable boundaries, corpus frequencies, and vowel and consonant projections for each entry.

The highlight of the dissertation is an investigation into whether the rhyme or body can be posited as an intermediate node in the structure of the Hungarian syllable. While both consonant-vowel and vowel-consonant sequences in Hungarian exhibit both attracting and repelling connections, on average the language behaves neither as a rhyme-type language (such as English) or a body-type language (such as Korean). It is suggested that a more-nuanced description of the Hungarian syllable is required, and an alternative representation is proposed.

The dissertation makes several contributions to research in Hungarian phonology. Results include insights into the statistics of phone-based n-grams for Hungarian, previously unknown phonotactic restrictions, and data on syllable structure with consequences for modeling the structure of the Hungarian syllable. In the cross-linguistic

context, the dissertation inspires related quantitative phonotactic research on disparate languages while simultaneously suggesting several caveats that should be taken into account when performing similar studies. In particular, the difficulties of quantitative studies of the syllable structure of multisyllabic words are addressed.

_____

_____

_____

_____

Table of Contents

1 Approaches to phonotactics

Phonotactics is typically regarded as the branch of phonology that investigates the restrictions in a language on the set of permissible phoneme combinations. The term was introduced to the linguistic literature by Robert Stockwell in the mid-1950's (Hill, 1958:68, Lutz, 1988:221, Menn, 2004:55). While it is not difficult to agree upon the nature of phonotactics, in practice researchers display a great deal of variability in their treatments of and approaches to phonotactics. Goldsmith (1995:3) presents four common interpretations of phonotactics (or conditions on well-formedness) that linguists use in their research:

(1.1) a. A well-formed word is one that is produced by taking an input string created by the morphological component, and applying the phonological rules of the language in the appropriate order.

b. A well-formed word is one that consists of a sequence of well-formed syllables.

c. A well-formed word is one in which all features (or autosegments) are associated to an appropriate skeletal position; all skeletal positions are associated with a syllable; and all syllables are associated with a foot.

d. A well-formed word is one that simultaneously satisfied all the well-formedness conditions of the language (including those given in c.)

Goldsmith further suggests that speakers of a language seem to have some knowledge of phonotactic constraints to the point that this knowledge is sometimes called a phonotactic grammar. Just as with other aspects of linguistic grammar, speakers of a language possess some sort of subconscious phonotactic knowledge. The following quote from Morris Halle (1978: 294) provides a classic argument concerning speakers' phonotactic awareness:

> The native speaker of a language knows a great deal about his language that he was never taught. An example of this untaught knowledge is illustrated in (1.2), where I have listed a number of words chosen from different languages, including English. In order to make this a fair test, the English words in the list are words that are unlikely to be familiar to the general public, including most crossword-puzzle fans:
>
> (1.2) ptak thole hlad plast sram mgla vlas flitch dnom rtut
>
> If one were to ask which of the ten words in this list are to be found in the unabridged Webster's, it is likely that readers of these lines would guess that *thole*, *plast*, and *flitch* are English words, whereas the rest are not English. This evidently gives rise to the question: How does a reader who has never seen any of the words on the list know that some are English and others are not? The answer is that the words judged not English have letter sequences not found in English. This implies that in learning the words of English the normal speaker acquires knowledge about the structure of words. The curious thing about this knowledge is that it is acquired although it is never taught, for English-speaking parents do not normally draw their children's attention to the fact that consonant sequences that begin English words are subject to certain restrictions that exclude words such as *ptak*, *sram*, and *rtut*, but allow *thole*, *flitch*, and *plast*. Nonetheless, in the absence of any overt teaching, speakers somehow acquire this knowledge.

One could object to Halle's blanket assertion without supporting evidence that English native speakers would select *thole*, *plast*, and *flitch*[1] as the English words, although it is likely that most linguists would find Halle's overall argument compelling nonetheless. His claim is that phonotactic knowledge exists and can be tested.

A number of studies demonstrate the existence of this salient phonotactic knowledge (Ohala and Ohala, 1986, Coleman and Pierrehumbert, 1997, Treiman et al., 2000, Bailey and Hahn, 2001, Hay, Pierrehumbert and Beckman, 2003). For example, the study of Coleman and Pierrehumbert asked speakers to rate the acceptability of

---

[1] Although uncommon, *flitch* is in fact an English word.

nonsense words in which part of the nonsense word would be phonotactically ill-formed but the remainder of the word statistically likely to occur. Speakers were seen as attempting to balance their acceptability judgments by allowing the well-formed chunks to offset the negative impact of the statistically unlikely portion; a single ill-formed portion of a word does imply the entire word is ill-formed. Coleman and Pierrehumbert concluded that probabilistic generative grammars constitute a more psychologically-realistic model of phonological competence than competing categorical generative grammars such as Optimality Theory in which a constraint violation is not typically permitted to be moderated or balanced by the well-formed portion of a word.

Further evidence for subconscious phonotactic knowledge can be seen by language-internal patterning as well. In the language Yawelmani Yokuts (Kisseberth, 1973), an examination of potential triconsonantal clusters demonstrates that speakers have knowledge of phonotactic constraints. Triconsonantal clusters are not permitted, and on a phonotactic level speakers respond to this constraint by repairing a potential sequence of three consonants (created by morphological processes) according to rules of the phonotactic grammar.

Having established a rough definition of phonotactics and speakers' subconscious knowledge of it, I will survey approaches to researching phonotactics across disciplines of linguistics in order to contextualize the approach used in this dissertation.

1.1 Phonotactic studies across linguistic sub-disciplines

Despite a simple definition for phonotactics, there is a great variety of approaches to this field of study. This section presents a brief survey of these approaches.

One approach is typological. For example, while it is generally assumed that co-occurrence restrictions are specific to a given language, there is a history of attempting to formalize phonotactic universals across languages by language typologists (cf. Trnka, 1936, Trubetzkoy, 1939, Saporta, 1963, Greenberg, 1978).

Another approach to phonotactics is through the lens of psycholinguistics, a field which has devoted considerable attention to phonotactics. Recent studies suggest that phonotactic knowledge is accessible and independent from lexical knowledge. For example, Gathercole et al. (1999) have investigated the influence of phonotactics on short term memory. For 7- and 8-year-olds participating in a non-word recall task, the researchers found that high phonotactic probability monosyllabic words were recalled with greater precision. Gathercole et al. suggested that it is the frequency of the syllables instead of individual biphones that correlated with accuracy in the task, although it should be noted that the words under examination were only monosyllabic.

The field of psycholinguistics may also investigate the observed facts in (1.3a), which are presumably affected by phonotactics. In turn, explanation is sought (1.3b) in various systems:

(1.3) a. Sources informing phonotactic research in psycholinguistics:
- Production errors
- Perception errors
- Learning biases/restrictions on possible grammars

b. Possible causes:
- Articulatory factors
- Perceptual factors
- Cognitive factors

Turning to the domain of language acquisition, phonotactic knowledge has been proposed as a method of decoding as a way to approach the problem of speech segmentation

(Mattys and Jusczyk, 2001). Evidence from phonotactic probability in young children is informing more traditional approaches to language acquisition (Jusczyk, Luce and Charles-Luce, 1994, Vitevitch et al., 1997, Storkel and Rogers, 2000) and giving possible clues to early-state language representation. For example, in the Jusczyk et al. (1994) paper, a head-turning preference task for 9-month-olds demonstrated sensitivity to the phonotactics of the native language. Infants preferred to listen to monosyllabic words with high phonotactic probability over those with low phonotactic probability. This and similar results inform hypotheses concerning the trajectory of phonotactic representation prior to acquisition of full adult language and suggest that infants may exploit their sensitivity to phonotactic information in learning their native language.

Within computational linguistics and related fields, the study of phonotactics has not been generally at the forefront of research. In fact, such research is not always even termed phonotactics, but there is a history to speak of nonetheless. Phonotactic models comprise the basis of acoustic models of speech recognition. Grapheme-based n-gram models are often used as the basis for automatic language identification systems, as the language-discriminating information is assumed to be encoded in the statistical regularities governing phone sequences.

On the more theoretical side of computational linguistics, there have been yet other approaches to phonotactics. For example, acquisition algorithms have been proposed to learn phonotactic patterns and constraints (Prince and Tesar, 2004, Heinz, 2006, Hayes and Wilson, 2008). Carson-Berndsen et al. (2004) approach phonotactic feature acquisition using finite state automata to represent the plausible phoneme combinations. Additionally, in a small computational study to inform the theory of

English phonotactic constraints, Davis (1989a) conducted an English lexicon-based study of co-occurrence constraints on consonants across an intervening vowel. Specifically, Davis sought to determine whether restrictions on sCVC patterns (where the first and second C are identical) only apply to monosyllabic words or if this restriction is a more general linguistic constraint; this is yet another example of a quantitative approach to phonology, in this case with implications for linguistic theory.

Phonotactics has also been a subject of interest in the field of second language acquisition. For example, Cebrian (2002) investigates the role that phonotactic constraints play in developing L2 phonemic contrasts for second language learners. A related question is how L2 learners and borrowers acquire phoneme combinations that do not appear in their L1.

Finally, phoneticians have sought to provide acoustic and articulatory explanations for the existence of phonotactic constraints (cf. Kawasaki-Fukumori, 1992, Ohala and Kawasaki-Fukumori, 1997). Some consider the set of phonotactic constraints to be a dynamical system and thus sensitive to changes in the lexicon. Phoneticians may also approach this problem from the point of view of co-articulation.

In summary, there is a great diversity in approaches to phonotactics. The remainder of this dissertation is structured as follows. The balance of the first chapter serves as an introduction to the type of approach to phonotactics that will be adopted throughout this dissertation. Specifically, Section 1.2 considers the role that domains – in particular the syllable and the morpheme – play in phonotactics. Section 1.3 discusses gradient and statistically-based phonotactics and addresses whether grammar is

independent from or based solely on these data. A few types of models for capturing statistical phoneme data are discussed.

The remaining chapters of the dissertation examine phonotactic phenomena as they relate to Hungarian. In the second chapter, a background of known phonotactic constraints for Hungarian is developed in order to serve as a baseline for which to judge the success of later results in the dissertation. Chapter 3 is a detailed discussion of the efforts involved in creating a lexical resource for examining Hungarian phonotactics from a quantitative point of view. In Chapter 4, this pronunciation lexicon is put to use in testing the question of whether the syllable in Hungarian can be described as having branching substructure such as the intermediate rhyme node proposed for English syllables (based on patterns of segment collocation). Finally, Chapter 5 briefly concludes the dissertation and suggests some of the directions that have been left unexplored for further research.

1.2 Domains of phonotactics

I adopt the view that the word, foot, syllable, and subsyllabic constituents can be relevant domains for phonotactics. This subsection examines the issues and tradeoffs encountered when trying to locate where phonotactic constraints are valid and relevant. It is not clear there exists one uniform domain across languages.

Typically the prosodic word is understood to be the domain for phonotactic phenomena. Enclitics and proclitics are not usually included in this domain. For example, in Hungarian, front-back vowel harmony holds between the stem and suffixes, but verbal proclitics do not trigger or undergo harmony. (This is relevant to phonotactics

to the extent vowel harmony is viewed as a phonotactic constraint – see Section 1.2.1.) Hence it would appear that the domain for vowel harmony and phonotactics in Hungarian is the prosodic word.

Unfortunately, the situation in Hungarian is complicated by the fact that verbal proclitics receive primary word stress. This is a problem because, by definition, the prosodic word should be the domain for word level stress; one would expect vowels in proclitics to harmonize. Vogel (1988) proposes a solution to this dilemma by creating a bifurcation of the prosodic word category into a larger clitic group (the domain for stress) and a subdomain called the phonological word, which includes only stems and suffixes. Hence the issue of a precisely-defined prosodic word category in Hungarian is somewhat obviated by this modified definition.[2]

Returning to the need to refer to domains in phonotactics, the question arises whether it is necessary to appeal to a subdomain below the level of the prosodic (or phonological) word category. Possible subdomains in other languages have been reported to be the syllable, morpheme, and/or the word stem. It may be possible that particular attributes of the domain are relevant – for example, in both German and Dutch the distribution of schwa is prohibited in the initial syllable of lexical morphemes but is

---

[2] The behavior of word compounds and bare accusative noun+verb combinations in Hungarian parallels that of proclitics. Consider the following examples in which primary stress is on the initial syllable of the word/phrase (in examples below acute accents mark long vowels):

sárgaréz	'brass'	(NB. sárga 'yellow', réz 'copper')
kenyeret vág	'slice bread'	(NB. kenyer 'bread', vág 'cut')

Following Vogel 1988, the mixed harmony of the above examples illustrates that the clitic group domain can include word compounds and bare accusative nouns beyond simply proclitics. The domain for harmony is restricted to the phonological word. The unity of the proclitic group is further substantiated by syntactic movement: proclitics and bare accusative nouns are postposed after the verb under focus movement (no such movement exists for word compounds).

allowed in grammatical morphemes (Hall, 1999, Booij, 2000). Part of speech may also be another domain attribute which narrows or expands the range of applicable phonotactic constraints – see Section 2.3.1 for more about domains in Hungarian phonotactics.

1.2.1 Locality in phonotactics

Heinz (2007) partitions the universe of phonotactic patterns into contiguous and non-contiguous patterns. Most studies of phonotactics are traditionally focused on contiguous patterns, and hence this partition naturally draws focus to non-local patterns or restrictions over sequences of non-adjacent sounds. Heinz considers consonant harmony (Hansson, 2001, Rose and Walker, 2004) and consonant dissimilation to be examples of non-local phonotactics but leaves open the question of whether vowel harmony is a non-contiguous phenomenon; the reason for this is that adjacent vowel peaks can appear to be phonetically contiguous despite the presence of intervening consonants. The role of tiers (vowel, consonant, moraic, etc.) in phonotactics is also discussed by Goldsmith and Riggle (2007) and Hayes and Wilson (2008) and is addressed later in this dissertation.

Heinz notes that locality has been considered a key feature of phonological grammars for some time. That is, a structural element may reference the existence of adjacent structural elements but in general may not count further to non-local elements which are not adjacent. For example, it would be odd for a phonological stress rule to say that stress should be placed in the fifth rightmost syllable – counting over a distance of several structural units is atypical of phonological phenomena. McCarthy and Prince (1986:1) corroborate this by remarking that general considerations of locality suggest that

phonological grammars may employ counting techniques, but only locally – "a rule may fix on one specified element and examine a structurally adjacent element and no other".

1.2.2 The role of morphology in phonotactics

Butskhrikidze (2002) remarks that formal approaches to phonotactics tend to be interested in formal units – the syllable, the foot, the onset, the nucleus, and so forth. However, it is suggested that examining the role of morphemes (meaning-bearing units) in phonotactics may be instructive. This approach is suggestive of morphotactics, the study of ordering restrictions on sequences of morphemes (cf. e.g. Sproat, 1992: 83, Beesley and Karttunen, 2000). It has been observed that languages place specific restrictions on certain classes of morphemes. In Dutch native words, for example, a prefix may have at most one syllable, and a suffix may have at most two syllables. Lexical morphemes, however, are not restricted in length (Booij, 1977:22-23). Similarly, there may be a restriction as to the allowable segments in particular morphemes. In Czech, only eight of the twenty-three consonants appear in inflectional suffixes. For these reasons, the approach of Butskhrikidze (2002) is to incorporate morphological constituents into the study of phonotactics. Section 2.3 addresses morpheme-induced phonotactic restrictions in Hungarian.

1.2.3 The syllable and phonotactics

In early generative phonology, there was no role for the syllable. However, it has been pointed out by Kahn (1980) that generative rules formulated as applying in the environment preceding a consonant or word boundary are candidates for rules that could

more generally reference a syllable boundary (which tend to occur before a consonant or end of a word). Haugen states that the best framework for describing the distribution of phonemes is the syllable. "Those who attempt to avoid the syllable in their distributional statements are generally left with unmanageable or awkward masses of material" (Haugen, 1956: 216). Despite this, many researchers have ignored the syllable in phonological descriptions of languages (Chomsky and Halle, 1968, Hyman, 1985, Kaye, Lowenstamm and Vergnaud, 1985) or only appeal to it informally. The following subsections explore arguments for and against referencing the syllable in a study of the phonotactics of a language.

1.2.3.1 Against the syllable

There are phonetic, formal, and lexically-based reasons, amongst others, as to why some linguists either deny or choose not to make use of the syllable in phonotactic descriptions. Harris (1994:45) remarks that "the term [syllable] can be formally taught as a means of labeling some aspect of phonological reality, but it is by no means always obvious exactly what that reality is." Under rapid speech or due to elision, the number of syllables that make up a word can be variable. Typically syllable counting involves vowel/peak counting – a task that is not too difficult. However, as Steriade points out, speakers differ in their responses when asked to identify exact syllable boundaries (Steriade, 1999).

Steriade's reasoning not to appeal to syllables and instead to only examine characteristics of segment strings is related to a formal reason against the syllable which could be termed "redundancy avoidance". In this case, the question of whether the

syllable exists or can be identified by the language user is not the central question, but rather its necessity is of primary importance. By the principle of Occam's razor, one chooses not to posit additional structure in a grammar when existing mechanisms are adequate. Many researchers have noted that "syllable structure can be determined just from the segmental composition of a word" (cf. Spencer, 1996:96). Similarly, some believe that the concept of syllable is dependent upon an appeal to sonority. Hence it is possible to skip the creation of syllable and simply appeal to sonority. For example, the Syllable Contact Law (which limits phoneme sequences across adjacent syllables) and the Sonority Sequencing Principle (which limits phoneme sequences with regard to sonority but not explicitly with regard to syllabicity) are not unique principles but rather similar observations couched in distinct terms – it has been claimed that they can essentially be derived from one another.

A final argument articulated against the syllable is the hypothesis that syllable boundaries are not included in the mental lexicon. This view is dominant in most theoretical frameworks, and this absence of the syllable has some basis (e.g. Levelt, 1992, Roelofs, 1996). The lack of syllable divisions in the lexicon would imply that they are not necessary for phonotactic generalizations over the lexicon.

1.2.3.2 In support of the syllable

Within phonology, classical arguments in support of the syllable were essentially outlined forty years ago (Anderson, 1969, Fudge, 1969, Hooper, 1972, Vennemann, 1978). According to Fudge, the syllable acts as a domain for prosodic processes and serves as a location for organizing and expressing constraints on possible segment sequences (Fudge, 1969). The syllable cannot function as independent from the word due to the principle of

exhaustiveness – a unit of a given level is exhaustively contained in the superordinate unit of which it is part (Nespor and Vogel, 1986:7). The principle of exhaustiveness, also appearing as the Strict Layer Hypothesis (Selkirk, 1984) and the Prosodic Licensing Principle (Ito, 1989) implies all syllables are part of the prosodic word. This fact alone, however, does not entail the necessity of the syllable in phonotactic descriptions.

There is some debate as to how much of general phonology should be accounted for within a phonotactic theory. Depending on the theoretical framework, phonotactics could be viewed as interacting with all of the following phonological principles: the Sonority Sequencing Principle, the Obligatory Contour Principle, the Syllable Contact Law, and the Balancing Principle. These principles have typically been used in conjunction with syllable boundaries to describe restrictions on phoneme co-occurrence.

Despite the views for and against the syllable and the debate over what value it adds, it is the view of this author that the use of the syllable in a phonotactic description is necessary if and only if there does not exist a syllable string parsing algorithm for the language. That is, if syllable boundaries can be determined using only the phone string of a word, then the information contained in the phoneme string with no syllable boundaries is equivalent to the information contained in a syllable-segmented phone string. Such an algorithm exists for Hungarian (see Section 2.4), and hence it is argued that the final outcome of the syllable debate does not bear crucially on the phonotactic description of the language.

1.3 Statistical and gradient phonotactics

The study of phonotactics has traditionally involved a strict division of phoneme sequences into allowable and impermissible sets. A string or phoneme cluster is

categorically judged as either grammatical or ungrammatical with no intermediate categorization possible. However, for phoneme sequences which do not appear, one can draw a distinction between accidental lexical gaps and systematic lexical gaps (Halle, 1962). In an accidental lexical gap, a phoneme sequence is unattested but without principled reason. In this dissertation, a somewhat generalized approach to phonotactics is adopted that attempts to measure the likelihood of segment sequences based on their observed distributions in the lexicon and corpora. This is in accordance with emerging research trends in the field of computational phonology. There exists a relationship between frequency and grammaticality in phonology, and many early models of phonotactic grammaticality even incorporate gradient phonotactics in one form or another (Greenberg and Jenkins, 1964, Chomsky and Halle, 1968, Clements and Keyser, 1983).

1.3.1 Gradient phonotactic grammaticality

Partitioning words into well-formed and non well-formed categories underestimates the phonotactic knowledge that speakers possess (Coleman and Pierrehumbert, 1997, Frisch, Pierrehumbert and Broe, 2004, Heinz, 2007). In the view of these researchers, speakers are said to be capable of making finer, gradient distinctions, and lexical items can vary in their well-formedness depending on the phoneme combinations they contain. In a recent paper, Coetzee and Pater (Coetzee and Pater, 2008) propose a grammatical theory of gradient phonotactics stated in terms of weighted constraints in the sense of Harmonic Grammar (cf. Smolensky and Legendre, 2006). Other modifications to Optimality Theory and Harmonic Grammar to reflect lexical statistics include Hammond's

Probabilistic Optimality Theory (Hammond, 2004) and the Gradual Learning Algorithm (Boersma and Hayes, 2001).

One reason to admit gradience to the grammar is the observation that novel words, or so-called "wug" forms, demonstrate probabilistic acceptability that depends on their component phoneme combinations. For example, Albright (2006) notes that speakers of English judge nonce words such as *stin* to be rather good, *smy* to be marginal, and *bzarshk* to be unacceptable. Hayes and Wilson (2008) discuss a number of other reasons to believe a gradient model is useful.

Anttila (2008) distinguishes two types of gradient grammars. The first arises from degrees of acceptability according to a grammar. Often phonological grammars are formalized (or later modified) so as to predict relative likelihoods of segment combinations based on markedness considerations. In the model of Boersma and Hayes (2001), a continuous measure is used as the basis of a categorized well-formed/ill-formed distinction. Categorical phonotactic gradience can be imposed by using a boundary below which no forms are acceptable; above this boundary, forms are judged grammatical with possibly differing degrees of fitness.

For this first type of gradience, Anttila proposes a modification of Optimality Theory in which the complexity of the grammar is inversely correlated with phonotactic grammaticality. According to this proposal, the more ranking information a phonotactic structure requires in order to surface faithfully, the less well-formed it is. The following is Anttila's statement of his Complexity Hypothesis:

(1.4) The Complexity Hypothesis: The probability of an (input, output) mapping is inversely correlated with its grammatical complexity. (Anttila, 2008)

The apparent problem with the Complexity Hypothesis is that it is not likely a testable theory – unless the complexity of the grammar is based on other, independent considerations, constraints could be added or deleted by virtue of post hoc reasoning.

The second type of gradient phonotactic grammaticality that Anttila speaks of is lexical, and it is based on lexical statistics. A novel word would derive "support" from existing words depending on the number of its lexical neighbors, defined traditionally in terms of string edit distance or some other similarity metric. Anttila goes on to say that "the best approach seems to be to develop explicit theories of both types [of gradient grammaticality] and try to figure out what kind of division of labor is empirically justified", an idea also proposed by Coetzee (2008). The view is contrasted with that of Hay et al., who say that "phonological grammar is a simple projection of lexical statistics" (2003:59) – the implication here is that because grammar is derived from lexical statistics, there can be no type of gradience other than statistical. Coetzee claims that there is partial independence of usage frequency and grammar:

(1.5) Independence of grammar and frequency (Coetzee, 2008)
- (a) Language users have linguistic knowledge about structures with which they have no experience.
- (b) Successful grammar learning requires some prior linguistic knowledge – knowledge that does not depend on experience.
- (c) Not all results of speech processing experiments can be explained by reference to usage frequencies.

Lexical statistics constitute a concrete set of data that are easily analyzed and not dependent on a particular theory. It is this second type of gradient phonotactic grammaticality distinguished by Anttila that is examined throughout this dissertation.

1.3.2 Defining phonotactic probability

Coleman and Pierrehumbert (1997) asked research subjects to rank nonsense forms on a scale of well-formedness from 1 to 7. They found that subjects' well-formedness judgments were correlated with the neighborhood density of the word and with the frequency of the component phones. Bailey and Hahn (2001) also demonstrated that phonotactic probability and neighborhood density play a role in speaker judgments of phonotactic well-formedness. This section examines exactly how linguists such as Coleman and Pierrehumbert (1997) and Bailey and Hahn (2001) approach calculating phonotactic probability and the different possibilities that exist for this calculation.

In the sections below, it is assumed that the term *probability* denotes the relative frequency of occurrence of some element (uniphone, biphone, etc.). This is determined by the count frequency of that element divided by the count frequency of all elements in the universe under consideration. Unseen events, such as novel combinations of phones, are assumed to have zero probability until evidence is attested to the contrary. Such evidence could come from a larger corpus, but while novel words might be attested in ever larger corpora, I do not expect the phonotactic probabilities to change much with the addition of new forms.

1.3.2.1 Uniphone model

A uniphone, or segment frequency model, assumes that the probability of a word is the product of the probability of the *n* component phones, which are denoted p_1, p_2, … p_n.

(1.6) $P(p_1\, p_2 \ldots p_n) = \prod_{i=1}^{n} P(p_i)$

In (1.6), the probability of each p_i is its relative frequency in some dictionary or corpus; this could be a type or token frequency. Below in (1.7) the formula is "unpacked" to show how a probability calculation for the famous nonce word *blik* would work.

(1.7) $P(\text{blɪk}) = \prod_{i=1}^{4} P(p_i) = P(\text{b}) \times P(\text{l}) \times P(\text{ɪ}) \times P(\text{k})$

The largest drawback for the uniphone model in assigning a probability estimate is that the model does not distinguish between different orders of phones (because multiplication is commutative, and the order of the multiplicands does not affect the outcome probability). For example, the uniphone model assigns identical probabilities to the words *tap* and *pta* – a less than desirable result!

1.3.2.2 Biphone model, or uniphone model with mutual information

The biphone model is also referred to as the uniphone model with mutual information; transitional probabilities between phones are now included in this model.

(1.8) $P(p_1\, p_2 \ldots p_n) = \prod_{i=1}^{n-1} P(p_i p_{i+1})$

Once again, to make clear the notation in (1.8), the probability of a word is now assumed to be the product of each biphone pair comprising it, as shown in (1.9).

(1.9) $P(\text{blɪk}) = \prod_{i=1}^{3} P(p_i p_{i+1}) = P(\text{bl}) \times P(\text{lɪ}) \times P(\text{ɪk})$

It is also possible and somewhat standard to add an additional symbol to the phone alphabet – the word boundary symbol '#'. This is appended to both the beginning and end of each word's phone string, and thus this symbol typically becomes the most frequent symbol in the phone alphabet.

(1.10) $P(\#blɪk\#) = \prod_{i=1}^{5} P(p_i) = P(\#b) \times P(bl) \times P(lɪ) \times P(ɪk) \times P(k\#)$

Triphone models are also defined analogously. An implementation of the biphone model by John Goldsmith is available for download under the title Phonological Complexity Calculator[3]. This software takes as input a word list and assigns complexity values for word phonotactics based on an entropy measure discussed in the next section.

1.3.2.3 Log probabilities and measures of entropy

For other researchers calculating phonotactic probability, it has been common to use negative log probabilities. The reason for doing this has not been motivated very well. The log of a number between zero and one is a negative number, and hence the negative log is a positive number. The question is then primarily why to use the logarithm at all.

The answer to this question lies in the distribution of probabilities in a corpus (cf. Zipf, 1935). Zipf's Law states that the majority of words in a corpus occur infrequently, while a small number are used quite often. Logarithms are essentially exponents, and when comparing large numbers such as frequencies, the logarithm transformation allows for easier comparison of relative frequencies by comparing the magnitudes of numbers.

---

[3] Available at http://hum.uchicago.edu/~jagoldsm/PhonologicalComplexity/

The logarithm magnifies infrequent items while shrinking the relative importance of outliers; this allows for a more straightforward grouping of frequencies into classes more interpretable to humans.

The other reason why logarithms are employed is computational. Programming languages only store decimal points to a certain precision. Given that probabilities are small numbers that need to extend to several decimal points, multiplying probabilities can quickly result in lost accuracy. As the multiplication of frequencies is equivalent to the addition of logarithms, addition is the preferred alternative because addition is capable of maintaining the precision in the programming language's representation of the number.

1.3.2.4 Hybrid phonotactic probability models

It is useful to examine two additional phonotactic probability models that have gained attention recently. The first, which I will refer to as the Syllable Constituent Model (Hammond, 2004), assigns a phonotactic probability by breaking forms up into traditional prosodic units – syllables, onsets, rhymes – and then calculating the frequency of those units over a corpus. The expected probability of a nonce form is calculated by multiplying together the frequencies of its sequential parts. For example, the frequency score of a nonsense form like [blɪk] is calculated by determining the frequency of its onset and the frequency of its rhyme and multiplying them together:

(1.11) $P(blɪk) = P(bl) \times P(ɪk)$

There is a certain appeal of this model because it has a linguistic sophistication (in its use of syllables) that is absent from the models described above. This is also its weakness –

the Syllable Constituent Model requires parsing a word into constituent syllables and furthermore into onsets and rhymes. Many of the incorrect parses of *blik* that could have been considered were not included in the calculation in (1.11):

(1.12) Syllable parses for *blik* not considered (only monosyllables)
P(blɪk) = P( ) × P(blɪk)
P(blɪk) = P(b) × P(lɪk)
P(blɪk) = P(blɪ) × P(k)
P(blɪk) = P(blɪk) × P( )

In the Syllable Constituent Model, an algorithm for parsing into onset and rhyme is assumed to exist. Hence the process of assigning phonotactic probability potentially suffers from a lack of robustness. For a non-word such as *ikbl*, it is unclear whether a default syllable parse could be assigned.

Another measure of phonotactic probability that has illustrative value here is the Phonotactic Probability Calculator for English (Vitevitch and Luce, 2004). It uses two measures to estimate phonotactic probability: (a) positional segment frequency (how often a particular segment occurs in a certain position in a word) and (b) biphone frequency as discussed above. Both estimates of frequencies were derived from 20,000 words in the Merriam-Webster Pocket Dictionary of 1964; the frequencies were compiled by Kuĉera and Francis (1967).

According to Vitevitch and Luce, positional segment frequency is formulated as follows:

> Positional segment frequency was calculated by searching the computer readable transcriptions for all of the words in the dictionary (regardless of word length) that contained a given segment in a given position. The log (base 10) values of the frequencies with which those words occurred in English (based on Kuĉera and Francis, 1967) were summed together and then divided by the total log (base 10)

> frequency of all the words in the dictionary that have a segment in that position to provide an estimate of probability.

Based on the description, one must assume that words of length *n* contain positions number sequentially 1 to *n*.

The use of a positional calculator for phonemes is awkward; it disregards many important phonological generalizations. Estimating the probability of a phoneme based on its distance from word-initial position would be analogous to guessing the part-of-speech of a word based on its distance from the beginning of the sentence. Just as the latter ignores syntax and phrasal structure, the former is ignorant of syllable structure and sonority constraints. While not all syllables appear in all positions of the word, it is largely syllable position and sequencing constraints that are likely to determine whether a word is well-formed. As words have many different lengths, position number would only be a useful cue in first or second position. However, the use of a word-initial string delimiter such as '#' in biphones or triphones could also capture such generalizations. Word-internal and word-final constraints could not easily be generalized due to the indeterminate number of segments intervening from the initial position to the position under question. Position number may have been a useful tool in the context of CVC words, but its use cannot be generalized. Another reason to doubt that segment position is relevant beyond the initial position of the word is the comment against counting noted by Kenstowicz – "the well-established generalization that linguistic rules do not count beyond two" (Kenstowicz, 1994:597).

1.3.2.5 Average phonotactic probability

Average phonotactic probability is an attempt to abstract away from word length in attempting to assign a well-formedness judgment. By averaging the component probabilities, this proposal (due to Goldsmith) is a way to compensate for the fact that longer words have reduced probability simply due to their length. Many people share the intuition that a word such as *bye* should have the same phonotactic probability as its reduplicated form *bye-bye*. In the calculation of phonotactic complexity, one iteratively adds the negative log probabilities of the n-phones contained in the word and divides by the number of n-phones. The example given here is using biphones.

$$\text{(1.14)}\quad \text{APP}(p_1 p_2 \ldots p_n) = \frac{\sum_{i=1}^{n} -\ln(P(p_i p_{i+1}))}{n}$$

Average phonotactic probability seems useful. However, a warning – the contribution of word length to token probability and phonotactic probability seems to not be well understood. While word length is indirectly proportional to probability, the effect of word length on phonotactic probability is unclear. Ultimately the form of phonotactic probability adopted should be the measure that most closely correlates with word-likeness judgments of native speaker informants.

1.4 Summary

The purpose of this introductory chapter has been to examine the theoretical presuppositions and frameworks relating to recent work in quantitative phonotactics. An understanding of developments in the field provides essential background and perspective

for the present research. As should now be apparent, this has been a relatively active area of work within phonology within the past several years.

In this chapter we have seen that several issues conspire to make it difficult to design a universal metric for phonotactic well-formedness. The proper definition of a domain, such as the prosodic word, is relevant. Internal hierarchy of words including syllables and their subunits is also relevant. Word length may also play a role, and additionally neighborhood density in the lexicon is also relevant (although it was not addressed in this discussion).

In the next chapter, the focus turns towards an examination of the phonotactics of the language of research for this dissertation, Hungarian.

2 Hungarian phonotactics and lexical statistics

The purpose of this chapter is to provide a rudimentary introduction to the Hungarian sound system and to survey the known phonotactic constraints of the language. This background is necessary in order to provide a baseline for evaluating the phonotactic results obtained in later chapters of the dissertation. This chapter also contrasts with Chapter 1 by instantiating the theoretical discussion of phonotactics with details specific to Hungarian. Finally, the last section of this chapter explores phone frequency data and statistical phone collocations of the Hungarian lexicon.

Hungarian is traditionally regarded as a member of the Ugric branch of the Finno-Ugric language family. The language has a complex morphological system with over eighteen case suffixes, and it is often cited as being a so-called agglutinative language. There are approximately fourteen million speakers of Hungarian, ten million of whom reside in Hungary. Thus Hungarian is among the seventy largest languages of the world (Grimes, 1996). Its closest linguistic relatives are the Ob-Ugric languages Khanty (Ostyak) and Mansi (Vogul).

Hungarian has received a great deal of attention from linguists both within Hungary and abroad. Within the domain of syntax, the language is typically described as a Subject-Verb-Object language[4], although the case system allows relatively free word order restricted primarily by topic and focus considerations. Preverbal focus and related word order issues are the most widely studied aspect of Hungarian syntax. Within phonology, vowel harmony has received the most attention and will be addressed in

---

[4] More precisely the most neutral word order is SVO in cases where the object is definite and SOV in cases where the object is indefinite.

greater detail later in this dissertation. The rich morphology of Hungarian also poses several interesting linguistic issues – there have been reported to be anywhere from sixteen to twenty-eight grammatical case markers.

Unless otherwise mentioned, the style of speech described throughout this work is the dialect known as Educated Colloquial Hungarian (ECH) – the dominant variety spoken in Budapest. ECH contrasts somewhat with the variety of Hungarian known as Standard Literary Hungarian (SLH). SLH is said to have been rooted in varieties spoken in eastern Hungary and Transylvania.

2.1 Hungarian segment inventory

2.1.1 Segment length

Hungarian makes a binary length distinction for all segments – both vowels and consonants alike. The only exception to this generalization is that there is no long counterpart to [h]. In some environments or for certain allophones, segmental length contrasts are not possible; for example, geminates do not appear in consonant clusters – only in intervocalic position and word finally.

Minimal pairs differing in segment length can be found for most vowels and consonants. Short segments are generally more frequent than their long counterparts for both vowels and consonants. Long segments can either be specified lexically, or they can alternatively be created by one of a few morphological processes. Most cases of productive gemination (lengthening) occur with the instrumental and translative case suffixes that lengthen a final singleton consonant of a noun stem. The primary morphophonological alternation involving length in the vowel system is the so-called

Low Vowel Lengthening process (cf. Vago, 1980). In Low Vowel Lengthening, a low, short vowel is invariably long before a suffix. This happens whether the morphological base form is simplex or complex – the lengthened vowel can be part of the original stem or a suffix itself. Note, however, that both long vowels and geminate consonants may be found in monomorphemic lexical forms and need not be created by morphophonological processes.

In Hungarian orthography, long vowels are represented using an acute accent over the vowel (or two accents over the vowel in the case of ö and ü). Long consonants, on the other hand, are written as a sequence of two identical consonants. Hence there is a sort of bifurcation in the way length is treated by the writing system. There are both historical and linguistic reasons for this. In the phonology literature, much discussion has focused on what the language-internal representation of geminates should look like. So-called "fake" geminates are geminates comprised of a sequence of identical consonants (Hayes, 1986:326-327). Following Hayes' definition, a "true" geminate cannot be split by epenthesis; similarly, there do not exist phonological processes which only act on a single half of the geminate. Polgárdi (2005) describes the fake geminates appearing in Hungarian as those arising through concatenation in which identical segments appear at the concatenation boundary. An example of this is when a stem ending in [b] is suffixed with the [b]-initial suffix -ben, as in the word *habban* 'foam-INESSIVE'. The proper representation of geminate consonants (one root node versus two) has long been an issue in Hungarian phonology (cf. Obendorfer, 1975, Vago, 1992, Grimes, 2005). I return to the issue when deciding on a representation for consonants and vowels in the pronunciation dictionary in Chapter 3.

2.1.2 Consonants

The range of phonological segments found in Hungarian is similar to segments found in western European languages. However, Hungarian is distinctive for its series of palatal consonants. An International Phonetic Alphabet (IPA) chart for Hungarian consonants is reproduced here in (2.1). Several segments appear in voiced and voiceless pairs, and sounds requiring two symbols in their representations below are regarded as affricates.

(2.1) Consonant inventory of Hungarian

	Bilabial	Labiodental	Alveolar	Postalveolar	Palatal	Velar	Glottal
Plosive	p b		t d			k g	
Nasal	m		n		ɲ		
Trill			r				
Fricative		f v	s z	ʃ ʒ			h
Affricate			ts dz	tʃ dʒ	cç ɟʝ		
Approximant			l		j		

The reader may consult

Appendix A at the end of the dissertation to obtain the most common orthographic representation of the IPA sounds.

All instances of geminate consonants occur in intervocalic position or at the end of the word; they do not appear initially or as part of consonant clusters. Underlying geminates that would appear in a cluster with another consonant due to morphology will surface as singletons due to a length reduction process applying to consonant clusters – this reduction of GC or CG to CC is likely related to restrictions on maximal syllable weight as well as geminate articulation.

Triconsonantal clusters in Hungarian are generally limited to foreign borrowings and typically involve a sibilant. Geminates which are underlying are outnumbered by derived geminates which occur at morpheme boundaries through gemination processes or concatenation of identical sounds. Overall it can be said that geminate consonants have a relatively small functional load[5] in Hungarian (Obendorfer, 1975); this observation will be borne out by frequency data presented in Section 2.5.1.1.

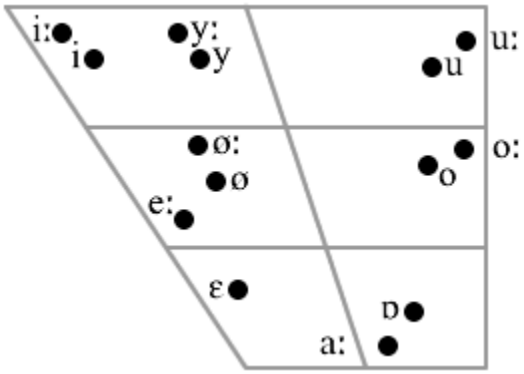
2.1.3 Vowels

There are fourteen vowels in Hungarian – seven short-long pairs. In the orthography, long vowels are indicated by an acute accent over the vowel. An excellent summary of the vowel inventory is presented by Siptár (1994).

(2.2) Hungarian vowel space

---

[5] The functional load of a particular segment or feature is the relative importance that the segment or feature has in making distinctions (minimal pairs) in the language. Charles Hockett (1955) proposed an interpretation of functional load in information theoretic terms.



(from Szende, 1994)

For further information on the vowel inventory, see (Siptár, 1994). The above diagram reflects the fact that long vowels are articulated more to the periphery of the vowel space. Despite the articulation discrepancies between long and short vowels, grammars of Hungarian and even phonologists generally idealize the vowels as only differing in length; quality differences can be explained by target undershoot, whereby the short and long vowels share the same articulation target but there is insufficient time for an articulator to reach its full targeted position in the case of a short vowel.

Target undershoot, however, does not explain the differences in quality between short-long pairs of the two "low" vowels. Although the height feature for the pair of vowels *e* [ɛ] and *é* [e] cannot be disputed to be *phonetically* mid (with the long vowel higher and more peripheral than the short vowel), these vowels seem to have a dual status in the *phonological* system. Specifically, the *e/é* pair belongs to both mid- and low-vowel paradigmatic alternations. Two Hungarian locative suffixes have three allomorphs containing mid vowels (-hez/-höz/-hoz 'towards' and -en/-ön/-on/-n 'on'). The quality of the vowel of the suffix is determined by vowel harmony according to the backness and roundedness specifications of the vowels. The existence of such a vowel alternation instantiates a class of vowels called "mid" in Hungarian.

By contrast, other locative suffixes (e.g. the inessive -ban/-ben) or the dative case suffix (-nak/-nek) have only two allomorphs according to whether the stem takes front or back vowel harmony agreement. The rounding harmony observed for front vowels does not apply in this case because the front vowel is specified as [+low], and there are no front rounded vowels with a [+low] specification on which to realize this harmony. (See Kornai 1991 for an interesting and detailed discussion of what phonological features are necessary and sufficient to create the proper natural classes for Hungarian vowels.) In any case, I assume that the existence of suffix pairs using [e] and [a] constitutes evidence that these vowels form a natural class. I follow a somewhat standard practice by referring to this class of vowels as low vowels despite their actual phonetic values.

2.2 Phonotactic Constraints

Several phonotactic generalizations for Hungarian have already been identified by other linguists. This section is a comprehensive survey of several known constraints.

2.2.1 Vowel phonotactics

The distribution of vowels in Hungarian appears to exhibit more restrictions than found for the consonant category. The constraints are presented sequentially below with brief commentary following each.

(2.3) [a:] is prohibited word-finally in major lexical categories in underived contexts.[6]
(Kenesei, Vago and Fenyvesi, 1998: 409)

[6] The only exceptions of which I am aware are *burzsoá* 'bourgeois' and *hajrá* 'rah! rah!'. The former is a loanword and the latter should be considered an ideophone.

A *derived* context where long final [a:] is permitted to appear is the translative case, where consonant-final nouns lengthen their final consonant and add a harmonic vowel (either [a:] or [e:], according to vowel harmony). For example, the word *bolond* 'fool/foolish' becomes *bolonddá* in the translative case, such as in the sentence *bolonddá tett engem* 'he made a fool out of me'. However, [e:] is the harmonic equivalent of [a:] for the translative case, and yet the constraint in (2.3) only references [a:]. It is unclear whether this phonotactic observation is the result of a systematic lexical gap or is simply a fact about Hungarian phonology.

(2.4) In polysyllabic words, a word-final /i/ is uniformly short.
(Kenesei et al., 1998: 409)

By contrast, word-final /i/ is uniformly long in monosyllabic words. In fact, all final vowels in open, monosyllabic words are long, a fact stated in the following constraint:

(2.5) In monosyllabic words, word-final short vowels are not permitted in major lexical categories (i.e. content words).

I have noted before that the only monosyllabic content words ending in a short vowel are *fa* 'tree' and *ma* 'today'. These words both contain the same low vowel, and it appears that low vowels are blocked from lengthening in this environment (e.g. Grimes, 2007).

(2.6) The phonemes [o] and [ö] are prohibited at the end of any morpheme.

This fact in (2.6) is often stated in a positive manner. That is, it can also be stated as "all instances of mid, round vowels are long in morpheme-final position". However, I am most familiar with this constraint being stated for word-final position, not morpheme-

final position. In fact, however, as Hungarian has no suffixes ending in these vowels, any vowel which is morpheme-final is also likely word-final in some derived or underived context.

(2.7) High vowels in the initial syllable are long.

I have demonstrated elsewhere (Grimes, 2005) that the functional load of length in high vowels is low. That is, length is not very contrastive within the high vowels. In fact, length contrast is not possible in some positions. In the initial syllable, the distinction in high vowel length is absent and all initial high vowels are long; this coincides with primary stress in the initial syllable. The table in (2.8) compares the traditional pronunciation found in Standard Literary Hungarian (SLH) with the speech found in ECH, the unmarked dialect of Budapest. The cases in (2.8a) show that high vowels exhibit length alternation in polysyllabic words, while in monosyllabic words in (2.8b) no alternation is possible.

(2.8)

Written Form	SLH	ECH	Gloss
a. fiú	[fiu:]	[fiu]	'boy'
tetű	[tetü:]	[tetü]	'louse'
házból	[ha:zbo:l]	[ha:zbol]	'from the house'
hegyről	[hed^{y}rö:l]	[hed^{y}röl]	'down the hill'
Szomorú	[somoru:]	[somoru]	'sad'
b. fű	[fü:]	[fü:]	'grass'
sí	[si:]	[si:]	'ski'

The comparison of dialects shows that the constraint in (2.7) is indeed active, as otherwise the final vowels in (2.8b) would be permitted to be pronounced short optionally in ECH. I now consider one final phonotactic rule concerning vowels:

(2.9) The vowels of a word must agree (harmonize) according to their specification of the backness feature.

Vowel harmony is not universally treated as a fact about phonotactics because phonotactics traditionally refers to adjacent (series of) segments. Nonetheless, to the extent that it constrains the possible variety of words in the language, I have included it in this list.

2.2.2 Consonant phonotactics

Hungarian grammar places no restrictions on either tautosyllabic (same syllable) onset-coda combinations when there is an intervening vowel or on consonants appearing non-consecutively in successive syllables. Restrictions on consonants instead pertain only to adjacent consonants (clusters). In word-initial position, a small, restricted set of onset clusters is permitted – these onset clusters always appear in words of foreign origin. Some analyses claim that complex onsets are not allowed at all in Hungarian phonology (cf. Siptár and Törkenczy, 2000). Nonetheless, assuming rising sonority of the consonants towards the syllable peak, some CC and CCC clusters are found word-initially. The quote in (2.10) regards word-internal consonant clusters:

(2.10) In Hungarian there are no phonotactic restrictions that constrain which consonants can be juxtaposed in a cluster $C_\alpha C_\beta$ when C_α is the last consonant of the first half of a compound word and C_β is the first consonant of the second half of the compound. The restrictions one may find are purely accidental or non-phonological. (Siptár and Törkenczy, 2000: 5)

In the footnote Siptár and Törkenczy go on to note that the few, non-accidental regularities that can be found are due to postlexical assimilations such as voice

assimilation and nasal place assimilation. Again, it is a matter of definition whether to treat postlexical assimilation as a phonotactic phenomenon; Siptár and Törkenczy apparently do not regard this as a matter of phonotactics.

Not all segments are found in all positions. The statements in (2.11a) and (2.11b) examine the segments not found word-initially and word-finally.

(2.11a) Word-initial segment restrictions (Kenesei et al., 1998: 386)
i. Word-initial /ty/ appears only in two instances: the noun *tyúk* 'hen' and the interjection *tyük* 'phew'.
ii. Word-initial /dz/ does not occur
iii. Word-initial /dzs/ is found in only approximately twenty loanwords.
iv. Geminates do not appear in word-initial position.

(2.11b) Word-final segment restrictions (Kenesei et al., 1998: 386)
i. All consonants are admissible word-finally except /h/.

The restriction on word-final /h/ is in fact the result of a general prohibition of /h/ in syllable-final position. The final /h/ is either deleted or is realized as its allophonic counterpart, a voiceless velar fricative [x]; the two options are mutually exclusive and depend on the lexical item and not the surrounding phonological environment. Hence words behave similarly to either *cseh* or *doh*, as shown in (2.12).

(2.12)

cseh type			*doh* type		
cseh	[če]	'Czech'	doh	[dox]	'musty smell'
cseh-től	[četö:l]	'Czech' (abl.)	doh-tól	[doxto:l]	'musty smell'(abl.)
csehes	[čeheš]	'Czech-like'	doh-os	[dohoš]	'musty'

(from Siptár and Törkenczy, 2000: 274)

This concludes the known constraints on consonant-consonant interactions.

2.2.3 Vowel-consonant interactions

There are relatively few restrictions on vowel-consonant co-occurrence. A unique restriction that is not immediately obvious even to the specialist is given in (13). The conditions for its application are rare, and hence there exist only perhaps a dozen relevant lexical forms that form the basis for the generalization.

(2.13) The nasal-obstruent clusters [mp] and [mb] can only be preceded by rounded vowels. (Kenesei et al., 1998: 419)

There are not many examples of such words, but the list includes *különbség*[7] 'difference', *gömb* 'sphere', *tömb* 'block', *gomb* 'button', *comb* 'thigh', *domb* 'hill', *lump* 'carouser' and *krumpli* 'potato'. This constraint in (2.13) is a tautosyllabic restriction; words such as *ember* 'man' show the generalization doesn't apply across syllable boundaries. The other vowel-consonant restriction appears in (2.14):

(2.14) A geminate may not follow a long vowel in monomorphemic, monosyllabic words.

The only potential exceptions to this rule are *áll* 'stand', *váll* 'shoulder', *száll* 'fly', and *épp* 'just'. The spelling here is misleading, however; in each of these cases, the geminate is pronounced short in normal speech, reinforcing the weight of the claim.

Finally, some vowel-consonant phonotactic restrictions are not typically stated purely as constraints on possible segment strings but instead reference syllable boundaries. The geminate example above in (2.14) is one such constraint, but another appears in (2.15):

---

[7] Orthographic /n/ is [m] in the consonant cluster in this word.

(2.15) In underived monosyllables ending in VVCC, VV can only be [e:] or [a:] and CC cannot be a geminate.

Here VV represents a long vowel and not a vowel sequence. (There are no diphthongs in Hungarian.) The constraint in (2.15) is another example of the exceptional behavior of low vowels.

Finally, there is one additional constraint that is worth mentioning, although it is somewhat unclear if this is a phonotactic constraint:

(2.16) The mirror principle of sonority holds for Hungarian.

In (2.16) the mirror is a metaphor for the vowel nucleus of the syllable – the possible segment combinations on one side of the vowel are a reflection (reverse ordering) of the possibilities on the other side of the vowel. Kornai (1990) makes this notion much more specific:

a. If PQ is a possible syllable onset (P, Q arbitrary consonants), then QP is not.
b. If PQ is a possible onset, then QP is a possible coda, and conversely, if RS is a possible coda, then SR is a possible onset.
c. If PQ is a possible coda, then QP is not.

Kornai's discussion is both intriguing and presents the case for the mirror principle in much greater detail than is possible here.

2.3 Phonotactic domains in Hungarian

Chapter 1 included a discussion concerning the proper phonological domain at which phonotactic constraints apply or are interpreted. The domain of application of

phonotactic rules is a relevant question for Hungarian; consider the following statement from Siptár and Törkenczy:

> Analytic morphological domain boundaries are opaque to phonotactic constraints; phonotactic constraints do not apply across them. Hungarian does not impose phonotactic restrictions on two consonants in a cluster occurring across a word boundary (for instance in a compound word). (Siptár and Törkenczy, 2000:5)

An example given is the intervocalic consonant sequence /kp/, which is only found in compound words such as *kerékpár* 'bicycle'. (This compound is composed of *kerék* 'wheel' and *pár* 'pair'.) The /kp/ cluster is otherwise not countenanced in the language. Inflectional and derivational suffixes also create otherwise absent consonant sequences.

Based on this information, the syllable itself is not an adequate domain for phonotactic rules – for a prosodic word to be phonotactically well-formed it is not sufficient for it to be composed simply of well-formed syllables; there are active cross-syllabic constraints. One syllabic constraint that apparently *does not* hold is the constraint that there be a fall in sonority between segments that span a syllable boundary – the Syllable Contact Law is inoperative in Hungarian (Siptár and Törkenczy, 2000:131). It may be the case, however, that distinct domains are relevant for each phonotactic constraint.

It is less clear what to make of the issue of analytic morphological boundaries. To study the phonotactics of Hungarian while ignoring morphological boundaries could miss many important generalizations. Indeed, otherwise impermissible consonant sequences appearing together could be used as a cue to morphological boundaries. However, this dissertation takes an unsupervised approach to examining Hungarian phonotactics. Morphological boundaries are not known *a priori*, and hence the prosodic word is rather

treated as a string of phones without internal structure. This allows for what could be described as a theory-neutral examination of the phonotactics. That is, this dissertation will seek to confirm or deny known phonotactic constraints based primarily on a data-oriented exploration of the distributions of phones in Hungarian; particular domains and morpheme boundaries are in general not assumed to be known.

2.3.1 Phonotactics of lexical subcategories

Certain lexical subcategories in Hungarian have distinct phonotactics from the language as a whole. This has been noted for certain lexical categories in other languages (Chomsky and Halle, 1968, Ito and Mester, 1995, Hall, 1999). For Hungarian, Törkenczy (2006) notes that the phonotactics of Hungarian verbs is more restricted than the phonotactics of the Hungarian language as a whole; fewer segment combinations are considered well-formed in the verbal system than in the phonotactics of non-verbs (see also Törkenczy, 2001, Trón and Rebrus, 2001, Rebrus and Trón, 2005). Further, it is possible that Hungarian place names may constitute a distinct phonotactic subgrammar (Rebrus and Trón, 2002). In addition to unique segment sequences, place names and proper names also have non-standard spellings. The complications they cause for creating a pronunciation dictionary are discussed in the next chapter (see Section 3.3.2). For a more general discussion on the phenomenon of particular parts of speech having special phonotactics, see (Kelly, 1991, Smith, 2001).

2.4 Syllables and syllabification in Hungarian

It is generally assumed that syllabification takes place during the course of derivation from underlying lexical representation to surface phonetic form. Depending upon morphological processes such as affixation, a single stem may be syllabified in multiple ways. However, the syllabification process is generally rule-based.

2.4.1 Syllabification algorithm

Syllabification in Hungarian is only potentially ambiguous when two or more consonants appear between vowels. The tension between where a syllable boundary is found can be described as being subject to two constraints – no complex onsets are permitted, but yet the onset should contain as many consonants as possible. In Optimality Theory, this tension is would typically expressed using the two opposing constraints in (2.17).

(2.17) *COMPLEXONSET: Syllable onsets may not contain more than one segment.
NOCODA: Syllables may not contain coda consonants.

The NOCODA constraint has the effect of ensuring the onset is not empty. These constraints appear here for illustrative purposes, but it is more convenient below to discuss syllabification in terms of rules.

The syllable parsing algorithm used to construct the syllable divisions found in the pronunciation dictionary is discussed in more detail in Chapter 3. The syllable parsing decision tree appears below in (2.18a-c). In (2.18a), note there are no diphthongs in standard Hungarian, and hence adjacent vowels are always parsed into distinct syllables. Note also that in the case of identical sequential vowels, two short vowels do not combine to create a single long vowel.

(2.18a) Case 1: VV (No intervening consonants)

Result: V.V

A syllable boundary intervenes, as diphthongs are not possible (cf. Kenesei et al., 1998: 414-5). For certain vowels, a consonant is also inserted here to interrupt hiatus.

(2.18b) Case 2: VCV (One intervening consonant)
Result: V.CV

This syllabification, as opposed to a VC.V syllabification, is to be expected according to near-universal cross-linguistic preferences for the basic CV syllable type.

(2.18c) Case 3: VC+CV (Two or more intervening consonants)
Result: Syllabified VC+.CV disallowing complex onsets.

Examples of (2.18c) rarely appear within the same morpheme aside from a few irregular monomorphemic examples. Instead, internal CCC clusters generally only occur when spanning the boundary of some analytic domain.

Given a maximum of one consonant in the onset under Case 3, one might be surprised or question whether the principle of Onset Maximization (cf. Selkirk, 1982b) is active in Hungarian. Onset Maximization is when syllable boundaries appear to have been selected such that as many consonants as licensed phonotactically appear in the onset of the following syllable. At first glance, it appears that Hungarian does not respect Onset Maximization if the language does not prefer complex onsets over complex codas. One confounding case occurs when a potential CC onset sequence is rising in sonority; here onset syllabification of the CC is possible but not required. For example, the word apró ‘small’ may be syllabified as [ap.ró] or [a.pró].

Furthermore, the examination of underived VVCCV sequences can be instructive. If the first vowel is long, recall that the vowel should undergo shortening as described in (2.14) (a constraint often referred to as *VVCC). The failure of the vowel to reduce may be the result of an onset maximization of syllables in rarified cases. For example, in the underived words *csúzli* 'slingshot' and *ródli* 'sledge', the vowel retains full length. This implies that the medial consonant cluster forms a complex onset of rising sonority. Alternatively, this can be viewed as a variant formulation of the Syllable Contact Law – a preference for a sonority drop across syllable boundaries. Syllabification is again summarized in (2.19).

(2.19) Syllabification rules – C1 here indicates a sequence of one or more consonants

Input structure	VV	VCV	VC1CV
Resulting syllable boundary	V.V	V.CV	VC1.CV

2.4.2 Exceptional syllabifications

There are some cases where morphological identity or etymology plays a role in creating syllabifications contradicting the algorithm described above. Consider the following compound word: *űrállomás* 'space station'. The compound is comprised of *űr* 'space' and *állomás* 'station'. One would expect the [r] to be re-syllabified into the onset of the second syllable according to (2.19). However, it remains in the primary syllable to reflect its identity as an independent word. This is just as described in Section 2.3 – while an analytic[8] boundary is a barrier to phonotactic constraints and syllabification, a

[8] An analytic or isolating language such as Chinese is a language in which words are composed of single morphemes. Synthetic languages fuse morphemes together to create words. Hence morpheme boundaries are typically synthetic and word boundaries are typically analytic; internal boundaries between compound words can possess properties of both types of boundaries.

synthetic one is transparent to syllabification/phonotactic interaction. In Hungarian it appears that compound word boundaries are analytic boundaries, at least with respect to syllabification.

2.5 Segment frequency in Hungarian

While the previous sections in this chapter have largely comprised reviews of the literature and a compilation of accepted phonotactic facts of Hungarian, this section introduces Hungarian segment frequency data that to my knowledge have received little or no attention. The result is an illustration of basic facts on segment frequencies in Hungarian.

2.5.1 Uniphone segment frequency

The table in (2.20) below presents segment frequency in Hungarian based on a dictionary wordlist (Kornai, 1986). The frequencies are *type* frequencies – each word is counted only once. Hence the token frequency of the word does not affect the segment frequency counts here. Type frequency is used in this instance because phonologists such as Bybee (2001) have argued that it is relative type frequency that serves as a reference for phonological and paradigmatic generalizations. Throughout I report both type and token frequencies for comparison.

A phone may appear multiple times in a word, and each instance of a phone is counted. Note that lemma frequency – that is, the number of derivations and inflections of a given stem – could in theory contribute to more frequent and productive stems having their component phones counted multiple times, allowing a backdoor for word token frequencies to affect type frequency counts. However, this effect is diminished because the wordlist is based on a dictionary and not on a corpus – while a dictionary can be expected to contain multiple derivations of a stem, it generally only contains a single inflection for a given noun or verb.[9]

Despite the inventory of consonants being much larger than the inventory of vowels (twenty-five consonants versus seven vowels), vowels constitute approximately 40% of all segments based on the table in (2.20) listing 106,523 vowel occurrences and 157,997 instances of consonants. Similarly, six of the twelve most frequent segments are vowels.

---

[9] In the case of Hungarian, the dictionary entry is the unmarked third person singular for verbs and singular nominative for nouns. By using these bare stem forms without inflectional suffixes, this prevents frequent morphological markers from introducing frequency biases.

(2.20) Segment frequency organized by frequency

á	11595	b	4864
e	24021	p	4000
l	18355	h	3895
a	18204	f	3801
t	18196	ö	3766
r	14171	ó	3700
k	12845	u	2972
o	11721	ő	2807
s	11517	ny	2600
n	9203	cs	2190
i	9040	ú	1996
é	9005	c	1976
m	8343	gy	1880
g	7699	ü	1326
d	7459	ű	1280
sz	7122	zs	781
v	5797	ty	403
z	5658	dzs	62
j	5120		
í	5090		

The following subsections examine this data closer by examining frequencies of particular segment features.

2.5.1.1 The prominence of length contrast in Hungarian

Here I attempt to derive information from uniphone frequency statistics to infer the role that segment length plays in the language. Tables in (2.21) and (2.22) list data for the frequency of vowels and consonants. It is clear here that short vowels are more frequent than long vowels.

(2.21) Comparison of short and long vowel uniphone frequency

Basic Vowel	Frequency	Long Vowel	Frequency	Percent basic
a	36187	á	22981	61.2%
e	47590	é	17910	72.7%
i	17915	í	10128	63.9%
o	23361	ó	7391	76.0%
u	5920	ú	3941	60.0%
ö	7268	ő	5604	56.5%
ü	2623	ű	2546	50.7%

(2.22) Comparison of short and long consonant uniphone frequency

Basic Segment	Frequency	Geminate	Frequency	Percent basic
b	9358	bb	167	98.2%
c	3642	cc	274	93.0%
cs	4065	ccs	228	94.7%
d	14794	dd	28	99.8%
dz	122	ddz	1	99.2%
f	7253	ff	32	99.6%
g	15247	gg	87	99.4%
gy	3663	ggy	60	98.4%
h	7499	hh	0	100.0%
j	10009	jj	141	98.6%
k	24926	kk	318	98.7%
l	35348	ll	1163	96.8%
m	16144	mm	94	99.4%
n	18195	nn	113	99.4%
ny	5045	nny	128	97.5%
p	7753	pp	130	98.4%
r	27953	rr	256	99.1%
s	22572	ss	352	98.5%
sz	13330	ssz	628	95.5%
t	35228	tt	910	97.5%
ty	757	tty	48	94.0%
v	11154	vv	0	100.0%
z	11199	zz	83	99.3%
zs	1549	zzs	0	100.0%

The frequency of length in vowels is relatively uniform – short vowels make up between half to three-quarters of all vowel types. Because consonants are singletons (i.e. basic or non-geminate) in more than 95% of all cases (as seen in 2.22), I conclude that the

functional load of length in vowels is likely to be much higher than in consonants. The lack of importance of length in consonants was noted earlier by Obendorfer (1975). It is nonetheless intriguing to find how closely the set of vowels and the set of consonants cluster in terms of their ratios of short to long segments.

From (2.22) it can also be observed that some segments do not have long counterparts. While geminate [h] is not possible, for [v] and [Z] it is simply the case that these geminates failed to appear in the sample. They exist as derived geminates but do not appear in lexical entries. One caveat to the geminate frequency table in (2.22) is that geminates are somewhat underrepresented because the frequency list is based on dictionary forms and not a corpus-based wordlist; in a normal text geminates will appear with higher frequency as case markers.

In general there are tradeoffs when deciding whether to use a corpus or dictionary for this research. In the above case we have seen an example where using a corpus may have given better coverage of the language, and this is the greatest advantage of using a corpus – it more closely represents the natural language. However, confounding factors are introduced when using this raw data that can be eliminated when using an analyzed dictionary. For example, the dictionary contains only one entry for each lemma; however, in a corpus, because frequent words are more likely to appear multiple times with several different derivations and inflections, counting type frequency is difficult because one ends up counting several different instantiations of the same lemma. This is undesirable because it appears to over-count the effect of what we wish to consider as a single set of phonotactic constraints applying to a given word.

2.5.2 Biphone segment frequency

Whereas a uniphone model for phonotactics only predicts segment frequency, examining the biphone frequency list yields more insight into phonotactic relationships.

Given a phone inventory of 39 segments, there are $39^2 = 1521$ possible biphones. Of this number, 1053 appear at least once, or 69.2% of all possible biphones. The table in (2.23) gives the twenty most frequent biphones. No strictly vowel (VV) or consonant (CC) sequences appear in this list.

(2.23) Most frequent biphones in Hungarian

Biphone	Frequency
el	4169
le	2945
er	2692
te	2542
et	2499
ás	2496
at	2380
al	2213
me	2212
en	2211
ta	2093
la	1999
or	1854
ál	1763
ik	1744
ár	1712
ko	1667
re	1627
tá	1611
ol	1596

The most frequent biphones [el] and [le] are both verbal prefixes meaning ‘away’ and ‘down’. However, their high frequency may be simply due to the high uniphone frequencies of the component phones – [e] and [l] are the second and third most frequent

uniphones. At the other end of this spectrum, there are rare biphone sequences; 49 biphones only appear once. In addition, there are 468 possible combinations of two segments which never appear.

2.5.3 Triphone segment frequency

Given a phone inventory of 39 segments, there are $39^3 = 59{,}319$ possible triphones. Of this number, 11,708 actually occur in the dictionary, or 19.7%. Hence the triphone space is less dense than the biphone space, a fact certainly to be expected. There are 2737 triphones that only occur once. The table in (2.24) lists the twenty most frequent triphones.

(2.24) Twenty most frequent triphones in Hungarian

Triphone	Frequency
ele	879
fel	669
tás	647
meg	639
len	615
ere	608
dik	584
ség	575
tel	556
mek	520
let	511
ter	499
ság	494
lás	476
ala	457
ete	431
tés	426
tal	419
lan	403
tat	383

It can again be noted that all of the most common triphones are CVC or VCV sequences, with CVC sequences being four times as frequent as VCV sequences. No consonant clusters or vowel sequences appear near the top of the list. The first triphone containing a CC subsequence is [All][10] 'with a type frequency of 381; this is a verb meaning 'to stand'. The first triphone with a CC subsequence that is not a geminate is [eSt] at frequency 280. Also in the list of most frequent triphones, [sEg] is notable as a common suffix that attaches to an adjective or noun to form an abstract noun, similar to the function of *-ness* in English. The suffix [dik] is a common verb ending on verbs with reciprocal meaning. Meanwhile, [fel] is a verbal prefix meaning 'up', and [meg] is also a verbal prefix used to mark the perfective aspect; its devoiced counterpart [mek] also appears high on the list.

2.6 Summary

This concludes Chapter 2, which has introduced the most widely known phonotactic constraints of Hungarian. These constraints are a benchmark of known patterns. It is hoped that generalizations concerning segment distributions in the literature may be brought out as part of investigations in later chapters. In the meantime, Chapter 3 departs from phonological theory to report on the steps taken in order to create a pronunciation dictionary for Hungarian for use in Chapter 4.

---

[10] Chapter 3 contains details on the phonetic alphabet used in the dictionary. Specifically, Chapter 3 contains details explaining why geminates are treated as CC sequences.

3 The creation of a pronunciation dictionary

This chapter describes the process of creating a pronunciation dictionary for Hungarian for the purpose of aiding in linguistic research on Hungarian phonology and phonotactics. I created the pronunciation dictionary by transforming orthographic forms to pronunciation representations by utilizing systematic deviations between Hungarian orthography and pronunciation. A collection of rules or algorithms used to generate pronunciations can be referred to as *letter-to-phoneme* (L2P) or *letter-to-sound* rules.[11] Letter-to-phoneme and syllabification algorithms together comprise fundamental problems in the domain of applied computational phonology for text-to-speech and related tasks. Standard L2P techniques may include using letter chunking, phoneme classifiers, sequence-based models (Bisani and Ney, 2002), or hybrid approaches (van den Bosch and Canisius, 2006). L2P is often a necessary prerequisite to text-to-speech or pronunciation modeling for speech recognition, but this dissertation demonstrates that there are a number of additional applications.

The present work is similar to work done by Olaszy, who extracted pronunciations from text using similar methods for use in a text-to-speech application (Olaszy, 2003, Olaszy and Kálmán, 2005). The rule-based creation of such a dictionary can be expected to be reasonably accurate due to the similarity of Hungarian orthography to actual pronunciation. This chapter includes discussion of goals and requirements for creating a Hungarian pronunciation dictionary, and each phonological change creating a mismatch between orthography and pronunciation is highlighted. Following the discussion on the

[11] In the Hungarian context the term *letter* might best be avoided because the word *betű* 'letter' in Hungarian does not refer to an individual grapheme but to one or more graphemes used to represent a single speech sound.

creation of the dictionary, I discuss possible future enhancements. Strategies for evaluating the quality of the dictionary are also discussed in this chapter. Finally, potential applications to linguistic research – aside from those being addressed in the present dissertation – are addressed at the end of the chapter.

3.1 Introduction to the task

While students of the English language quickly learn that English spelling is by no means regular or consistent, many Hungarians believe that the Hungarian alphabet is completely phonetic. Here, a phonetic alphabet refers to the existence of a one-to-one mapping between symbol and sound. It can quite easily be demonstrated by counter-example that Hungarian orthography is not phonetic, and in fact several types of orthographic-pronunciation discrepancies exist. Consider as an example the word /szabadság/ [sabač:a:g][12] 'freedom, liberty', in which no fewer than four orthographic-pronunciation discrepancies can be identified with the written form of this word:

(3.1) a. The sequence /sz/ is a digraph corresponding to the sound [s] while /s/ alone would correspond to [š].
b. A general process of voicing assimilation applying between two consonants requires the [dš] to be pronounced [tš].
c. The [tš] consonant cluster subsequently undergoes affrication (or coalescence) and is pronounced [č:].[13]
d. The acute accent on the vowel /á/ indicates vowel length – compare /a/ [ɔ] and /á/ [ɑ:]. The issue is not so much an orthographic-pronunciation discrepancy as a

[12] Here I adopt the practice of enclosing graphemes with /forward slashes/ and pronunciations using [square brackets] based on analogy with the widespread practice of using these grouping symbols for /underlying forms/ and [surface forms] in phonology. This usage does not imply that the orthographic form is a phonemic form, but rather that it is an input to a derivation.

[13] The use of č here is equivalent to IPA [tʃ], but the choice of one symbol over two symbols is not meant to indicative of the status of segmental status of affricates in Hungarian. For recent work on this topic, see Pycha, A. (2007). Phonetic vs. phonological lengthening in affricates. *Proceedings of the 16th International Conference on the Phonetic Sciences*, 1757-1760.

problem of character encoding – the dictionary must be able to be shared across multiple computing platforms using symbols universally understood by different systems. Also, a decision is necessary as to whether to represent long segments of the language with a unique symbol or using a doubled version of the segment's short counterpart.

Fortunately for both the Hungarian language learner as well as for the creator of a pronunciation dictionary, the above discrepancies are fairly representative of the types of systematic deviations of the writing system from speech. In fact, the majority of the sound-symbol discrepancies in Hungarian are regular. Hence one is able to develop a system of replacement rules which rewrite the grapheme strings into a phonemic transcription that disambiguates pronunciation. Fortunately in Hungarian there are no known homographs (distinct words spelled identically but pronounced differently).

Exceptional pronunciations of course do occur. Any exceptional word is one in which the deviation between orthography and pronunciation is not sufficiently systematic. These non-systematic cases cannot be handled by a rewrite rule and are instead listed as exceptions. The list of exceptions can be thought of as a lexicon, whereas the rewrite rules comprise a sort of grammar that bears some resemblance to the actual phonological grammar of Hungarian. In fact, the orthographic-pronunciation rewrite grammar would almost constitute a proper subset of the phonological grammar if not for the fact that it also handles discrepancies that are not phonological in nature. More discussion of implementation details can be found in Section 3.5.

An advantage of using this automated technique for generating word pronunciations lies in its generalizability to unseen words, also known as out-of-vocabulary (OOV) words. The set of words for which pronunciations can be generated is in principle unbounded just as is the lexicon of a natural language. The ability to

generalize to unseen words is crucial – new words and novel word combinations are continually appearing. For applications such as speech recognition, the ability for a pronunciation dictionary to provide statistically probably pronunciation conjectures for previously unseen words is crucial.

3.1.1 Pronunciation dictionaries in other languages

The term *pronunciation dictionary* is at times used interchangeably with *phonological lexicon*, although a phonological lexicon typically includes pronunciation information in addition to richer lexical data such as frequency, stress, syllabification, and so forth. The present work actually constitutes a phonological lexicon, but I direct most of the attention here towards investigating the proper generation of pronunciations for lexical items.

Several pronunciation dictionaries exist for English, including the Hoosier Mental Lexicon[14] (Nusbaum, Pisoni and Davis, 1984), the Carnegie Mellon Pronouncing Dictionary (CMU, 1993), PRONLEX (distributed by the Linguistic Data Consortium), and the CELEX2 database from the CELEX organization (the Dutch Center for Lexical Information) (see Celex, 1993, Baayen, Piepenbrock and Gulikers, 1996). PRONLEX (also known as COMLEX English Pronouncing Lexicon) was designed for speech recognition and contains 90,694 word forms from the Wall Street Journal and Switchboard corpora (LDC, 1995). CELEX2 contains the Oxford Advanced Learner's Dictionary and Longman Dictionary of Contemporary English; CELEX2 is based on British English pronunciation, while the other three pronunciation dictionaries above are based upon American English speech.

---

[14] The Hoosier Mental Lexicon contains pronunciations from Webster's English Dictionary.

For languages other than English, CELEX2 also contains lexicons for German and Dutch. ELRA (the European Language Resources Association) distributes phonetic lexicons based on Spanish and Catalan. The standard pronunciation dictionary for French is BRULEX (Content, Mousty and Radeau, 1990). The LDC also distributes pronunciation dictionaries for Egyptian Arabic, German, Japanese, Korean, Mandarin, and Spanish. Furthermore, proprietary pronunciation dictionaries developed by language technology companies also exist for English and for any language for which there has been work done on speech recognition. Unfortunately, non-proprietary, freely available pronunciation dictionaries available for use in linguistic research are relatively limited. Of the lexicons listed above, only the CMU pronouncing dictionary is freely available, and most lexicons cost upwards of several thousands of dollars for usage rights.

There is also a distinct lack of pronunciation dictionaries and phonological lexicons available for studying non Indo-European languages. As a consequence, much of the recent research on the phonological structure of the lexicon is based almost exclusively on English. The development of a pronunciation dictionary for Hungarian, a Finno-Ugric language, offers an opportunity to study a lexicon that is not derived from the Indo-European word stock. Hungarian is particularly worthy of study because it is a so-called agglutinative language with a relatively high morpheme-to-word ratio, meaning that the majority of words appearing in a given corpus likely consist of two or more morphemes[15]. Additionally, Hungarian also has a more complex verbal inflection system

[15] If the lexicon of forms under investigation is derived from a dictionary (as opposed to from a corpus), we expect that the agglutinative characteristics would not be as rich as could be found in an examination of fully-inflected forms from a corpus.

than Germanic or Romance language families – the language families for which pronunciation dictionaries are available to date.

Nevertheless, Hungarian cannot rightfully be considered a *resource-light* language – languages for which corpora and other computational linguistics tools do not exist or are not readily available. Rather, several computational tools are already available for Hungarian (e.g. Kornai, 1986, Váradi, 2002, Halácsy et al., 2004). In fact it is beyond the present scope to survey all of the available tools, which have focused on strategies for morphological analysis and part-of-speech tagging. However, because of existing resources, the present research and dictionary creation is in some sense collaborative and certainly made more feasible by building on the previous work of several others.

It must be stressed that due to the relatively close relationship between orthography and pronunciation, assigning pronunciations to each written form is relatively straightforward. The degree to which a language's orthography is regular has been termed *orthographic depth* or *orthographic transparency* (Sproat, 2000, Neef, Neijt and Sproat, 2002). The extent to which Hungarian has shallow orthographic depth will be illustrated in detail in Section 3.4, where the aspects of Hungarian phonology not already reflected in the writing system are discussed in detail. Such a process would be extensible to other languages for which there are predictable relationships between orthography and pronunciation.

3.1.2 Limitations of this pronunciation dictionary of Hungarian

The pronunciation dictionaries described in the preceding section are quite useful in linguistic research and technological linguistic applications. However, there are limitations to their intended use. Most pronunciation dictionaries are not intended to be used as definitive guides to the pronunciation of the language. This is accomplished by publications directed to the public at large; for Hungarian this includes *The Hungarian Pronunciation Dictionary* (Fekete, 1995), which aids L2 Hungarian speakers or L1 speakers living outside Hungary in acquiring correct pronunciation. Another resource, *Pronunciation dictionary: The correct pronunciation of foreign names and words* (Tótfalusi, 2006), helps native speakers of Hungarian pronounce foreign words and names. These resources are generally insufficient for the research linguist because they are often incomplete (including only difficult-to-pronounce words), or in other cases these dictionaries list multiple pronunciation variants for each word (distinguishing between pronunciations in careful and rapid speech contexts, for example). The reader is left to determine which pronunciation is preferred or whether the pronunciation variation occurs across dialects or speakers. More specifically, these resources are also inadequate for determining the correct vowel length of certain high vowels.[16] In short, no existing resource served the present purposes.

[16] A personal motivation for undertaking this work as an L2 Hungarian speaker was to sort out length vacillation in instances where the orthography is inconsistent.

3.2 Design requirements for the Hungarian pronunciation dictionary

The pronunciation dictionary of Hungarian under consideration here was inspired by the Hoosier Mental Lexicon (herein HML) developed in the Psychology Department at Indiana University (Nusbaum et al., 1984). In many ways, the HML served as a guide for developing requirements concerning formatting and content, and the body of research based on the HML encouraged me to undertake this project in order to encourage more comparative work on Hungarian. For approximately 20,000 English words, the HML lists both written forms and broad phonetic transcriptions in a phonetic alphabet. The HML also includes additional data such as the length of the phonetic form (raw segment count), its consonant-vowel structural makeup, the corpus frequency of the word, familiarity ratings, and other additional information.

3.2.1 Hungarian corpora and word lists

In developing a pronunciation dictionary for Hungarian, my initial input was a word list of orthographic Hungarian developed at the Research Institute for Linguistics in Budapest during the 1980's (Kornai, 1986). This dictionary contains approximately 67,000 entries. It is my intention to later extend this work to be based on the entire lexicon of the Hungarian National Corpus (Váradi, 2002), which includes a more comprehensive and extensive lexicon with 2,950,000 unique entries, as well as part-of-speech labels for most words. Without the permission of the Hungarian Academy of Sciences, I would not be able to make a pronunciation dictionary based on the Hungarian National Corpus as widely available as is possible with the freely downloadable Kornai corpus.

As I noted earlier, Hungarian enjoys relatively rich computational linguistic resources. While I am using a dictionary-based word list as the basis for the pronunciation dictionary, word frequencies are based on corpus data. Integrating corpus data and dictionary data is not always trivial (see Souter, 1993 for an example concerning an English corpus). Because the frequency distribution of words in a corpus is well-known to follow Zipf's Law, even large corpora will occasionally fail to provide the adequate coverage that dictionaries provide, and there ideally should be methods for estimating probabilities of low frequency words.

For Hungarian, the average word frequency is smaller than compared with other languages. Larger corpora are typically required to provide a comparable amount of coverage. Put another way, the tail of Zipf's curve is rather long for Hungarian. According to Oravecz and Dienes (2002), a quarter-million word corpus of Hungarian has some 50,000 distinct words. The same size corpus of English has only 19,000 distinct words.

The large vocabulary size makes many computational linguistic tasks more difficult in Hungarian. For example, Oravecz and Dienes found a significant degradation in their HMM (Hidden Markov Model) tagging performance using HMM-based POS tagging tool TnT. They achieved only 67.07% performance on unknown words compared with around 84-85% for unknown words with a comparable amount of English training data. This performance degradation was suggested to be due in part to the proliferation of morphological forms in Hungarian, and it must be noted that Oravecz and Dienes were not, strictly speaking, using TnT for the purpose intended.

When Oravecz and Dienes used TnT in conjunction with a morphological analyzer, they were able to achieve part-of-speech tagging accuracy above 95%. An additional reason that Hungarian tagging is difficult is due to the relatively free word order causing transitional probabilities between words to become less reliable as compared to languages with more rigid syntax.

In order to include a given word in the pronunciation dictionary, I required that it appear in each of three resources. The table in (3.2) is intended to illustrate the degree of (non-)overlap of word types in the three resources that primarily contributed to the pronunciation dictionary. Here the resources are a digitized dictionary database (Kornai, 1986), the Hungarian National Corpus (Váradi, 2002), herein HNC, and a web-based corpus called simply Webkorpusz (Halácsy et al., 2004) that arose out of the WordSword project and whose aim was to create a tokenized corpus of Hungarian larger than any existing corpus. The intersection of words contained in two or more of these resources is necessarily smaller than any the number of words found in of the component resources.

(3.2) Relative sizes of electronic resources by number of words (types)

Resource	Words types
Kornai	67,397
HNC	108,159
Webkorpusz[17]	7,200,000

[17] The original Webkorpusz, based on a crawl of the.hu domain, contained some 19.1 million word types over a 1.486 billion word corpus. Texts appearing multiple times and files containing no useable text were filtered out. The size figure cited here for Webkorpusz represents the most error-free kernel of the corpus: the so-called 4% corpus. This was created by only accepting documents in which 96% or more of the words contained in it were accepted by a spellchecker. The authors report that 4% is the average number of spell check errors in a regular document, and attempting to require fewer spelling errors would "not increase the quality of the remaining text but would eliminate all pages that do not adhere to a strict spelling norm."

Combined Resources[18]	Word types
Kornai & HNC	33,883
Kornai & Webkorpusz	59,726
Kornai & HNC & Webkorpusz	33,660

My pronunciation dictionary includes frequency information from both the Webkorpusz and HNC, but not from the Kornai resource, which as a dictionary has no frequency data. Part-of-speech information is available from the HNC, and when combining words appearing in all three sources, the size of the pronunciation dictionary is approximately 33,660 words. By requiring that words appear in each of three distinct resources, the inclusion of spurious or obsolete words in the pronunciation dictionary is minimized. Here I placed primary importance on the quality of words included in the dictionary at the expense of widespread coverage.

3.2.2 Contents of the Hungarian phonological lexicon

The focus of the present work is pronunciation, but in addition to pronunciation, the dictionary also includes the following information for each word:

(3.3) (a) Orthographical form (from Kornai, 1986)
(b) Pronunciation (present work)
(c) CV tier representation (present work[19])
(d) Syllable structure (present work, described in Section 2.4)
(e) Frequency counts (from Halácsy et al., 2004)

[18] By combined resource, I refer to the collection of words appearing in each of the individual sources. Hence the combined resource is actually equal to or smaller than the smallest of the sources.
[19] See also Péter Szigetvári's research and his resources available on his personal webpage for work on the CV syllable structure of the Hungarian lexicon.

This report focuses primarily on the relationship between the items in (3.3a) and (3.3b) – the creation of a pronunciation from each orthographic form. We may distinguish a *phonological lexicon* from a pronunciation dictionary; information including CV tier, syllable structure, and frequency count data would extend the pronunciation dictionary into being a phonological lexicon. A screenshot of the first page of the expanded lexicon is included at the end of the dissertation as Appendix B.

3.2.3 Dialects and Idiolects

For certain words, more than one pronunciation is possible, and this variation can be observed across dialects, registers, or individual speakers. Variation is ignored by making a decision to select only one pronunciation for each written form. When possible, the most frequent variant is chosen. It was my goal to select a standard, phonological transcription for each pronunciation. Phonological alternations found in certain dialects have not been treated, although this may be an interesting topic for further research. The pronunciation dictionary is intended to reflect the Budapest dialect-standard typically referred to as Educated Colloquial Hungarian (ECH). The ECH dialect-standard stands in contrast with Standard Literary Hungarian (SLH), which represents a rarer, idealized form of the language. Regional dialects also exist. I chose to describe the ECH standard not only due to its popularity, but also because the majority of current phonology literature focuses on this dialect-standard. For treatments of divergent phonological processes in the minority dialects of Hungarian, I can refer the reader to the general overviews provided by Rot (1994) and Kiss (2001).

3.2.4 Phonemic versus phonetic approaches

Pronunciation dictionaries vary as to whether they employ phonemic transcriptions of speech segments, phonetic transcriptions, or some variant along this spectrum. There exist alternatives, of course, to this symbolic, segment-based approach to representing the acoustic stream: Cohen (1995), objecting that traditional transcription systems were developed for written language instead of spoken language, uses a hybrid symbolic-neural network approach to phonetic transcription that captures graded information for each phoneme based on transitions to adjacent phonemes. Using syllables for the basic, atomic unit of transcription would also be a possibility; similarly, the use of Wickelphones[20] and Wickelfeatures (Wickelgren, 1969) would be in the same spirit as using basic units larger than segments. Finally, another alternative would be to transcribe the location of the articulators throughout the duration of the word, along the lines of the Browman and Goldstein gestural model (Browman and Goldstein, 1989).

Narrow phonetic transcriptions risk not being generalizable beyond a single speaker or speech instance. Broad phonemic transcriptions, on the other hand, risk omitting phonetic detail and can occasionally be theory-dependent. I chose to create a phonemic lexicon, and in doing so I have omitted four allophones from the dictionary. I have not included the velar nasal; its distribution is restricted to appearing immediately before velar stops. Another allophone, the alveo-palatal fricative [ɕ], is only found as an allomorph of /j/ in word-final position in the second person singular indefinite imperative

---

[20] A Wickelphone is a trigram containing the phone itself and its predecessor and successor. Hence in Wickelphones the word *bat* consists of *#ba*, *bat*, and *at#*. Hence the size of a Wickelphone inventory of a language is rougly the cube of the phone inventory. This captures the idea that phones have different realizations depending on their immediate context.

when preceded by a voiceless stop. The remaining two allophones are variants of /h/; when appearing in intervocalic position /h/ yields [ɦ], while geminate /h/ is realized as [x:].

In the present work, I use the term *phone* as a theory-neutral alternative to *phoneme* in order to refer to a single speech sound or segment. The terms biphone and triphone indicate pairs or triples of phones, just as a bigrams and trigrams are pairs and triples of graphemes (or words). For recent summaries of dissenting views against the phoneme, see (Port, 2007a, Port, 2007b).

3.2.5 Character representations of Hungarian sounds

Due to the fact that the Hungarian alphabet makes use of letters that are not included in the basic ASCII[21] character standard, it has often proved difficult to send Hungarian computer files between different machines without experiencing problems of encoding – an individual or their software must know the encoding of a file in order to interpret it properly. For example, Western European languages are encoded in ISO-8859-1 (Latin 1), while Eastern European languages are typically encoded in ISO-8859-2 (Latin 2). Despite the fact that the development and adoption of the Unicode standard promises to eliminate these hassles in the future, character encodings remain an issue at the present. Hence the default symbol representation scheme for this pronunciation dictionary should be based in 7-bit ASCII characters to ensure cross-platform compatibility. The alphabet selected is based on Péter Szigetvári's OGOB7, or one-grapheme-one-byte. This symbol

[21] ASCII is the American Standard Code for Information Interchange adopted during the 1960s. Using sequences of 0s and 1s of length seven (i.e. 7-bit sequences), the character set encodes $2^7 = 128$ symbols. Extended ASCII is an 8-bit encoding and contains $2^8 = 256$ symbols.

representation scheme enforces a principle of one-to-one mapping of sound to symbol. I will refer to my modified version of OGOB7 simply as OGOB.[22] While the name of this system might not seem readily transparent, the principle it is based on is straightforward. Just as English orthography uses digraphs such as /sh/ or /ch/ to denote a single sound, Hungarian uses digraphs or trigraphs to indicate sounds for which there is no single letter available in the Roman alphabet. An encoding scheme using a one-grapheme-one-byte principle represents each sound with a single character.

There are other advantages of a one-grapheme-one-byte system. Suppose one wants to search for all instances of consonant clusters of length exactly equal to two. Such a search performed on a text containing digraphs would yield false positive results such as /sz/ (which represents the single sound [s]); similarly, such a search could fail to find valid cases such as /nsz/ [ns]. The table in (3.4) gives the digraphs and trigraphs of Hungarian with their single character encoding equivalent.

---

[22] Szigetvári's purposes are somewhat different from mine, as he seeks to be able to convert back and forth between standard orthography and OGOB7. As a result, he requires a bijective mapping between the two encodings. My mapping from orthography to phones is not one-to-one because I collapse multiple ways of spelling a single phone into a single symbol. For instance, /ly/ and /j/ are both represented using [j] while Szigetvári introduces the symbol [L] in order to be able to recover spellings using /ly/. Here it would obfuscate pronunciations to use multiple symbols for a single sound, and hence it is not possible to recover the original spelling of a word given the pronounced form.

(3.4) Digraphs and trigraphs in OGOB

Hungarian Orthography	cs	ch[23]	dz[24]	dzs	gy	ly	ny	sz	ty	zs
IPA	tʃ[25]	x	--	ʤ[25]	ɟ	j	ɲ	s	c	ʒ
OGOB	C	H	--	D	G	j[26]	N	S	T	Z

As OGOB is not widespread in its use, the pronunciation dictionary is also made available in SAMPA, a transcription standard in computational linguistics. Conversion between OGOB and SAMPA is trivial and reversible because it consists of simple character string replacements. In this dissertation, I alternate between orthographic and OGOB representations for data.

For doubled or long segments, I retain the convention of the Hungarian writing system: long (geminate) consonants are written as a series of two consonants, while long vowels are represented by the capital letter version of the short vowel. Note that this is by no means a trivial decision, as there is some discussion in the Hungarian phonology literature about whether geminate consonants in Hungarian are "true" geminates or whether they are simply doubled versions of the basic segment (cf. Vago, 1992, Szigetvári, 2001). I might also add that in the vowel system, there are very few phonological processes that show convincingly that long vowels are truly lengthened variants of their short "counterparts", and hence the decision to use distinct symbols is

---

[23] The digraph /ch/ is not typically considered a digraph of Hungarian because it appears in only a few loanwords such as /pech/ 'bad luck' and a handful of proper nouns.

[24] Some grammars consider the digraph /dz/ to be a single (affricate) sound, but there is reason to believe it should simply be treated as a sequence of sounds – see the discussion in Siptár, P. & Törkenczy, M. (2000). *The Phonology of Hungarian.* Oxford: Oxford University Press. Hence instances of dz are considered separate phonemes and not affricates.

[25] Although represented by a sequence in IPA, we consider these affricates to be phonemes.

[26] The digraph /ly/ is equivalent to the modern Hungarian character /j/ and hence it is not necessary to introduce a symbol distinct from the one used for j.

not unwarranted. The orthography and corresponding character encoding in OGOB is given for the vowels in (3.5).

(3.5) Encoding of vowels with diacritics in OGOB

Hungarian Orthography	á	é	í	ó	ö	ő	ú	ü	ű
IPA	a:	e:	i:	o:	ø	ø:	u:	y	y:
OGOB	A	E	I	O	w	W	U	y	Y

Above note that OGOB uses lower case letters for short vowels and upper case letters for long vowels.

The remaining characters used in OGOB and not appearing in (3.4) or (3.5) are identical to the graphemes used in the Hungarian orthography. These characters are *b*, *c*, *d*, *e*, *f*, *g*, *h*, *i*, *j*, *k*, *l*, *m*, *n*, *o*, *p*, *r*, *s*, *t*, *u*, *v*, and *z*.

For my purposes, the amended OGOB encoding alphabet is ideal. However, in order to make the pronunciation dictionary widely useful to others, there are three populations of users that must be considered. First, some Hungarian specialists may be more accustomed to the Prószéky encoding, a system developed for earlier computers in which vowel diacritics are replaced by a letter followed by a digit as follows: 1 represents an acute accent, 2 is used for umlaut, and 3 is for the doubled acute accent (i.e. a long vowel with umlaut). For example the word *őrültség* 'insanity' would be rendered as *o3ru2ltse1g* in Proszéky encoding. This was an early means of working around the ASCII encoding limitations.

Meanwhile, for computational linguists, there exist standard transcription systems, such as SAMPA, the Speech Assessment Methods Phonetic Alphabet. SAMPA was developed in the 1980s and only uses 7-bit ASCII characters. SAMPA must provide

coverage for larger and more general phone sets, and hence SAMPA differs from OGOB in that multiple symbols often correspond to a single phoneme; transcriptions utilize white space to delimit phones. Individual SAMPA encodings exist for 29 languages, including Hungarian, while X- SAMPA (extended SAMPA) unifies the individual SAMPA alphabets.

Finally, International Phonetic Alphabet (IPA) symbols are most useful to the third concerned group – linguists with little or no knowledge of Hungarian. A table enumerating all the phonemes of Hungarian in each of the various transcription systems, including cross-references to IPA and Hungarian orthography, is given in

Appendix A.

3.2.6 Encoding of long segments

As was shown in (3.5), all long vowels are encoded using a single symbol which is related to but not identical to the base vowel. Conversely, long consonants are not assigned a unique symbol, but instead encoded by a sequence of two identical symbols. This bipartite approach to length representation may seem disconcerting, but it is parallel to the approach used in the Hungarian orthographic system. There are also linguistic reasons for the dual approach – a sequence of identical short vowels does not coalesce to create a long vowel, but instead the sequence remains distinct and straddles a syllable boundary. On the other hand, a sequence of two short consonants coalesces to form a long consonant (even in the case of affricates). Hence a sequence of two repeated consonants is indistinguishable from a long consonant, while the distinction between a sequence of short vowels and a long vowel motivates creating entirely separate categories (and hence new symbols) for the long vowels. This point is elaborated a bit further in the following section.

3.2.6.1 Single versus double root node for consonants in Hungarian

The dominant view in the cross-linguistic literature is to treat geminates as having single root nodes[27] (Hock, 1986, McCarthy and Prince, 1986/1996, Hayes, 1989, Hyman, 1992); under this view the geminate acts as a heavy (long) variant of the corresponding singleton

[27] A root node is a term adapted from computer science and tree graphs. The root node is the top node of the inverted tree that dominates all other nodes (as opposed to leaf nodes at the bottom). In phonology the root node the node which dominates all other features in a hierarchy. Here single phonemes dominate one or more moras.

consonant and is associated with a mora in an underlying representation (in contrast to single consonants which have no underlying moraic association). Meanwhile Selkirk has been the principal advocate of geminates as occupying a double root node (Selkirk, 1990). To support Selkirk, Tranel (1991) cites languages where geminates pattern as light (short) in the coda – Selkup, Malayalam, and Tubatulabal. Ringen and Vago provide a discussion of the relevant issues and characterize the single/double root debate as involving weight versus length interpretations of geminates (Ringen and Vago, 2006), and Davis (to appear) provides an overview on quantity surveys the different proposals for theoretical treatments of quantity in the literature.

One reason to treat geminates as having a double root node representation would be if they behave similarly to consonant clusters, which naturally occupy two consonant "slots". In the pattern of vowel epenthesis for verbs in Hungarian shown below, the forms in (3.6a) do not take epenthesis while the forms in (3.6b) do. This provides evidence that Hungarian geminates should be treated as occupying a double root node.

Crucially, note that the word for 'fear' in (3.6a) with a long vowel plus single consonant does not take epenthesis. However, the word for 'pour' in (3.6b) with a short vowel plus consonant cluster does take epenthesis. This led Vago (1992) to conclude that the epenthesis process in not sensitive to syllable weight per se, but rather the process counts the number of consonants. Given that geminates here pattern with consonant clusters, this gives reason to believe geminates occupy two C-slots or a double root node.

(3.6) Vowel epenthesis (data from Vago, 1992)

	3S	2S	infinitive	gloss
a.	kap	kapsz	kapni	receive
	Nő	nősz	nőni	grown
	Fél	félsz	félni	fear

b.	áld	áldasz	áldani	bless
	önt	öntesz	önteni	pour
	hall	hallasz	hallani	hear
	függ	függesz	függeni	depend

3.3 Converting orthography to pronunciation

Spelling conventions in the orthography of a language can be characterized as attempting to adhere to two competing standards. To language learners, a *pronunciation spelling* (or phonetic spelling) might be considered ideal, as the spelling of a given word can be directly deduced from its pronunciation. Hungarian orthography, however, at times conforms to what could be called the *etymological principle* (Keresztes, 1992:31, Vago, 1992). Here individual morphemes have a unified spelling across words, and morphophonological rules altering segments at morpheme boundaries are not reflected in the spelling. While etymological spelling may reflect the underlying morphological input, it does so at the expense of actual pronunciation. In practice, Hungarian orthography is not based wholly on pronunciation or etymology but is rather a combination of both. It is this tension that must be resolved in the creation of a pronunciation dictionary.

In this research, several sources were used to determine and verify standards for pronunciation in Educated Colloquial Hungarian (Deme, 1950, Kassai, 1989, Nádasdy, 1989b, Nádasdy, 1989a, Kontra, 1995, Pintzuk et al., 1995, Kenesei et al., 1998, Nádasdy and Siptár, 1998). Pronunciation and orthographical mismatches can be broadly grouped into one of three categories: (i) words which retain historical spellings, especially prevalent in place names and person names, (ii) orthographic issues relating to the alphabet, i.e. digraphs and trigraphs, and (iii) discrepancies resulting from the application

of phonological processes not reflected in orthography. The last category is by far the most extensive, and hence phonology is treated separately in Section 3.4, while the remaining issues are discussed in this section.

3.3.1 Many-to-one letter-to-sound relationships

Certain sounds and sound combinations in Hungarian have two possible spelling variants. The table in (3.7) shows that the OGOB encoding system used here consistently picks the more common orthographical convention. While most letters used in Hungarian are similar to the IPA or in this case OGOB symbol used in transcription, only a small number of sounds can be spelled in multiple ways:

(3.7) Multiple spelling strategies for consonants or clusters in Hungarian

Rare Hungarian Orthography	Standard Hungarian Orthography	OGOB
ly	j	j
q	kv	kv
w	v	v
x	ksz	kS

Order is important in mapping orthography to pronunciation. Because the OGOB encoding system uses /w/ and /y/ to stand in place of umlauted vowels, it is necessary to ensure that the replacements suggested by the table in (3.5) to eliminate the /w/ and /y/ graphemes with the replacements suggested by OGOB take place *before* introducing the characters for the vowels. Just as in any rule-ordered phonological grammar, the order of implementation of replacement rules in this project is also important.[28] The order of

[28] The relevant relationship between the two rules discussed here is one of counter-feeding.

presentation of phenomena in this report mirrors the actual order of implementation of the rules.

3.3.2 Divergent spelling conventions in the development of an orthography

The possibility of different or multiple spellings of a word represents an expansion of the issue described in the table in (3.7). It is necessary to consider several phonological, morphological, and historical factors. Attempts to standardize Hungarian spelling were not entirely successful until the late 19th and early 20th centuries (Benkő and Imre, 1972). As a result of the relatively recent standardization, Hungarian spelling accurately reflects modern pronunciation as the language has not had the chance to evolve and diverge greatly from its writing system over this relatively short period of time.

While writing standards had been proposed much earlier, the first time the Hungarian Academy of Sciences became involved in standardizing orthography and rules for writing Hungarian was in 1832. Essentially, the Hungarian writing system grew out of two traditions – the Catholic and Protestant writing systems (Benkő and Imre, 1972:565). The table below shortly summarizes a few of the crucial differences in the two writing systems and shows that present-day Hungarian orthography evolved, in part, from two separate traditions.

(3.8) Modern orthography as combination of two traditions

Modern orthography	IPA	Catholic tradition	Protestant tradition
cs	tʃ	cs	ts
c	ts	cz	tz
tj	cç	ty	tj

Most archaic spellings that survive today are typically found in place names and family names. Indeed, some names can have even more than two spellings, as in the variants of a particular family name: *Takács*, *Takáts*, and *Takách* 'Weaver'. Several additional letters are used in proper names or words of foreign origin (Keresztes, 1992: 30). These graphemes include ä, ae, c, ch, ie, oe, ph, q, sch, w, x, y. In general, names of foreign origin that were written in another script are transliterated. Foreign words written in the Roman script, however, generally retain their original spelling.

The issue of pronunciation of proper names is not inconsequential. In the AP Newswire corpus, Liberman and Church report that 21% of all tokens consist of proper names (Liberman and Church, 1992). The pronunciation of names follows simple phonetic rules, with the main exception being historic noble families such as *Dessewffy* or *Batthyány*. Németh et al. (2003) attempted to automatically guess the pronunciation of Hungarian names in the Hungarian phonebook. Their procedure involved creating 18 category labels such as "contains French word", "contains foreign word", "contains given name", "contains a single letter component", and so forth. A manually-labeled list of 300,000 proper names was used to label an entire database of 2,944,000 names. Using these very specific heuristics, the authors report 99% accuracy (precision). A summary of non-standard spellings is given in the table in (3.9).

(3.9) Non-standard spellings retained in family names

Historical Spelling	Modern Equivalent	Example
aa, áa	á	Gaal, Gaál
cz	c	Rákóczi
ee	é	Veer
eö, eő	ö, ő	Eötvös, Beőthy
ew	ő	Thewrewk [török]
oo	ó	Soos
uu	ú	Kuun
y	i	Ady
ch	cs	Madách
sz	s	Atzél
tz	c	Wessenlényi
th	t	Toth, Batthyány

For the present work, the Hungarian wordlist being used (Kornai, 1986) does not contain proper names. Most names such as those above contain the same orthographic-pronunciation discrepancies as the language as a whole. Hence there is no reason not to include rules to rewrite the foreign spellings first into modern orthography and then continue converting this representation to a phonetic transcription along with the rest of the wordlist.[29] As is likely the case for many languages, proper names in Hungarian display the greatest degree of divergence between pronounced and written form, owing to the influence of cross-cultural contact and population migrations. Fortunately, this topic occupies a minor sphere in the pronunciation dictionary – a more detailed treatment would only be required in the event the wordlist is expanded to include more proper nouns.

3.3.3 Digraphs and Trigraphs

As stated earlier, the Hungarian alphabet uses eight digraphs and one trigraph to represent single phones, and because of the expressed goal to have a one-to-one principle of sound-to-symbol correspondence for the pronunciation dictionary, I have elected to replace all

---

[29] In a few cases, it is not possible to apply general phonological rules developed here to the pronunciation of proper names; the patterns exhibited by the names conflict with more general spelling conventions. In the names Kossuth and Kiss, [ss] is pronounced [s]. However, it is not in general true that all cases of [ss] are reduced to [s]. For similar reasons, I can only count as exceptional such names as /papp/ [pap], /imreh/ [imre], and /cházár/ [CASAr].

digraphs with a single character. An initial step in the preprocessing of the dictionary is to convert any occurrence of an uppercase letter to lowercase. This ensures that the uppercase symbols used to represent long vowels and palatal consonants (digraphs) shown in the tables in (3.4) and (3.5) are unique; it also prevents having duplicate entries for a single word differing only in the capitalization of a single letter.

An interesting difficulty that arises from the particular set of digraphs and trigraphs in Hungarian involves the difficulty in disambiguating digraphs from consonant sequences. In computational linguistics this is related to the phoneme chunking problem. The following examples in (3.10) are from Péter Szigetvári, and they illustrate what might be called near-minimal pairs. Some caution is warranted, however, as some native speakers might find that the use of infrequent words in the following examples to be somewhat contrived.

(3.10) Examples of possible grapheme ambiguities involving clusters containing digraphs

[zs]		
Digraph:	*rézsűn* 'on the slope'	(*rézsű* 'slope', *-n* 'LOC')
Consonant cluster:	*rézsün* 'copper hedgehog'	(*réz* 'copper', *sün* 'hedgehog')
[szs]		
Monograph-digraph:	*sertészsír* 'pork grease'	(*sertés* 'pig', *zsír* 'grease')
Digraph-monograph:	*kertészsír* 'gardener's grave'	(*kertész* 'gardener', *sír* 'grave')
[cs]		
Digraph :	*lécsín* 'liquid beauty'	(*lé* 'liquid', *csín* 'beauty')
Consonant cluster:	*lécsín* 'slat track'	(*léc* 'slat', *sín* 'track')
[tty]		
Monograph-digraph:	*hattyúk* 'six hens'	(*hat* 'six', *tyúk* 'hens')
Long digraph:	*hattyúk* 'swans'	(*hattyú* 'swan', *-k* 'PL')[30]

One strategy that was not employed in the present work would be to use probabilistic heuristics or statistical machine learning to determine whether a potential digraph is a true digraph or simply a segment sequence. For example, consider the case of [j], spelled as both /j/ and /ly/. For historical reasons the digraph /ly/ (pronounced [j]) is more likely to occur at the end of polysyllabic words than word internally (Szemere, 1987). As a result, the word *muszáj* 'must' is incorrectly spelled *muszály* approximately ten percent of the time.[31] An alternate approach is to look up each component of the compound or derived form as a free-standing word in the dictionary. This approach takes care of all instances of grapheme ambiguity because the examples noted in (3.10) only involve compounds or derived words. (In the case of derived words only the stem can be looked

---

[30] In the final example, it must be noted that 'six hens' would typically be written *hat tyúk*, not as a single word.

[31] This data is based on a Google search of Hungarian web pages in late 2006 that found *műszály* occurring 118,000 times compared to the standard *műszáj* appearing 1,090,000 times. A Yahoo search returned similar results, as did a search in the Hungarian National Corpus. While the frequency data returned by search engines is only an approximation, in this case we can be reasonably certain that *muszáj* is mispelled as *muszály* with some regularity.

up in the dictionary.) In all other "simpler" cases of grapheme ambiguity, such as the /sz/ sequence being mistaken for independent /s/ and /z/ graphs, a principle of greedy grapheme chunking is used in which the potential grapheme sequence is always maximized.

3.4 Phonology and morphophonology unmarked in the orthography

In order to survey phonological processes of Hungarian, I consulted a variety of Hungarian grammars, dictionaries, and papers (e.g. Papp, 1969, Vago, 1980, Keresztes, 1992, van den Bosch and Daelemans, 1993, Törkenczy, 1994, Kenesei et al., 1998, Siptár and Törkenczy, 2000), as well as a number of stylistic guides to proper Hungarian writing and spelling conventions. Some phonological processes are already reflected in Hungarian orthography. For example, assimilation involving [v] is generally marked. However, voicing assimilation, palatalization, and affrication constitute a large number of the phonological processes that are not marked. In this section, each process is discussed in detail.

3.4.1 Assimilation of nasals to place of articulation

The nasal consonant [n] must agree with the specified value of the place of articulation feature of a following obstruent consonant. It actually may be the case that the following consonant need not be an obstruent (cf. e.g. Siptár and Törkenczy, 2000); however, the only sonorant which would provide crucial evidence is /j/ because all other non-obstruent consonants are alveolars, which already agree in place of articulation with [n]. In the case of /j/, there is palatal assimilation (see Section 3.4.3). While a backed variant of the

nasal appears before velars and palatals, all the examples in (3.11) involve fronting before a bilabial or dental segment. The velar nasal in Hungarian has dubious phonemic status because its appearance is always conditioned by a following velar consonant; because it is in complementary distribution with the alveolar nasal it can be considered an allophone of [n]. Hence at this time it is not used in the pronunciation dictionary.

(3.11) Examples of nasal place assimilation

Written Form	Pronounced Form	Gloss
szénpor	szémpor	'coal dust'
különben	külömben	'otherwise'
szenved	szemved	'suffer'

Let us now consider strategies used to create the dictionary. A linguist may seek to identify the most general statement of a rule in order to capture the generalization concerning nasal assimilation in (3.11). For example, a nasal must agree in place of articulation of a following consonant. This general rule is given in (3.12a), while a concrete instantiation concerning segments in Hungarian appears in (3.12b).

(3.12) a. $N \rightarrow [\alpha \text{ place}] / \_\_ C_{[\alpha \text{ place}]}$
b. $n \rightarrow m / \_\_$ (Ackerman, 1992 v)

The formulation in (3.12b) is more specific but formally equivalent, and I used something analogous to the latter in creating the dictionary. This is not a theoretical decision, but a practical one – implementing the rule in (3.12a) requires detailed feature data for each phoneme, while using the option in (3.12b) is less cumbersome and has the advantage of being very specific. Further implementation details are discussed in Section 3.5.

3.4.2 Voicing assimilation

In general, Hungarian obstruent consonant clusters must agree in voicing, and the assimilation process is anticipatory (also termed regressive assimilation). In instances of triconsonantal clusters across morpheme boundaries, this rule must apply iteratively. The segments /h/, /j/, /m/, /n/, /ny/, /l/ and /r/ do not undergo assimilation; note that the segments which do not undergo assimilation also do not have a voiced or voiceless counterpart. Furthermore, it is noted that [v] does not seem to trigger assimilation. Consonant clusters appearing in native stems already agree in their voicing features, and hence the interesting cases to look at involve processes of word creation. Examples of such ill-formed clusters with respect to voicing resolved through morphophonology either involve loanwords (3.13a), affixed forms (3.13b), or compound words (3.13c).

(3.13) Consonant assimilates to the voicing of a following consonant

Written Form	Pronounced Form	Gloss
a. abszolút	[ɔpsolu:t]	'absolute'
joghurt	[jokhurt]	'yogurt'
b. olvasd el	[olvɔʒdɛl]	'read it'
kútban	[ku:dbɔn]	'in the well'
c. népdal	[ne:bdɔl]	'folksong'
húsdarálo	[hu:ʒdɔra:lo]	'meat grinder'
kerékgyártó	[kɛreɟ:a:rto:]	'wheel maker'

A phonological rule requiring voicing agreement in consonants is given in (3.14). As stated above, the rule in (3.14) is understood not to apply in instances where the consonant does not have a counterpart of the appropriate voicing specification or if the second consonant is [v].

(3.14) $C \rightarrow C_{[\alpha \text{ voice}]} \ / \ \_\_\_ \ C_{[\alpha \text{ voice}]}$

3.4.3 Coronal palatalization

Morphology and phonology again interact in Hungarian in the case of coronal palatalization. A coronal stop is palatalized before the imperative morpheme or third person singular verbal suffix /j/. The result is coalescence of the two segments, but the moraic timing of the component segments is preserved. In other words, the resulting palatal is a long consonant.

(3.15) Palatalization of coronal stops involving [j] imperative morpheme

Written Form	Pronounced Form	Gloss
lát-ja	[la:c:ɔ]	see-3S.DEF 'he sees it'
ad-juk	[ɔɟ:uk]	give-1P.DEF 'we give it'
men-jen	[mɛɲ:ɛn]	go-IMP.S 'let him go'

3.4.4 Alveolar plosive affrication

When Hungarian morphology creates a sequence of an alveolar plosive and a following sibilant, these two segments coalesce into an affricate. The place of articulation of the resulting affricate is identical to the place of articulation of the sibilant. The new affricate is a long consonant unless reduced in length due to being adjacent to another consonant (a consonant cluster reduction rule is discussed in Section 3.4.8). A rule giving the relevant segments involved is given in (3.16).

(3.16) $t, t^y \rightarrow tʃ: \ / \ \_\_ \ ʃ$
$t, t^y \rightarrow ts: \ / \ \_\_ \ s$

The output of the process is a geminate affricate. The phonetic realization of gemination of affricates is lengthening of the stop closure of the affricate (i.e. duration expansion applies primarly to the first half of the affricate, not uniformly). These results were confirmed in a recent study of Hungarian affricates (Pycha, 2007). Examples of such affrication appear in (3.17).

(3.17) Examples of alveolar plosive affrication

Written Form	Pronounced Form	Gloss
váltson	[va:ltʃon]	'it should change'
szabadság	[szabatʃ:ág]	'freedom'
egyszer	[ɛts:ɛr]	'once'
maradsz	[mɔrɔts:]	'stay.2S'

With respect to rule ordering, it is crucial that this affrication take place after the voicing assimilation described in Section 3.4.2, as the voicing assimilation rule feeds affrication. For example, in the word *szabadság*, the devoicing of /d/ to /t/ is a necessary first step so that the word may meet the necessary input requirements to the affrication rule.

3.4.5 Hiatus resolution

A glide consonant [j] is inserted to interrupt so-called hiatus between a sequence of two vowels whenever one of the vowels is [i] or [i:]. The pattern is less clear for vowel sequences involving [e:], but in most cases it is optional (see Siptár and Törkenczy, 2000:282-284 for more details and examples). The process also acts across words in normal, fluid speech.

(3.18) Examples of hiatus resolution / glide insertion

Written Form	Pronounced Form	Gloss
tea	teja	‘tea’
szia	szija	‘hello’
hiába	hijába	‘in vain’
nénié	nénijé	‘the aunt’s’
dió	dijó	‘walnut’
kiöl	kijöl	‘extinguish’

Due to the optional nature of this rule and disagreements in native speaker judgments, I chose to only implement it for the clear cases of the high vowels [i] and [i:]. The rules in (3.19) state that an intervocalic empty string is rewritten as [j] in the environment preceding or following a short or long /i/.

(3.19) ø → j / V __ {i, i:}
ø → j / {i, i:}__ V

3.4.6 Phonotactics and syllable structure constraints

There is a phonotactic constraint stating that round, mid vowels are long in word-final position. This means that words end in [ó] and [ő] but do not end in [o] and [ö]. Hence there are many words such as *fogó* ‘pliers’ or *nő* ‘woman’, but I am aware of only are two exceptional function words: *no* ‘well [interjection]’and *ö* ‘ahh’. This relationship is formalized in (3.20).

(3.20) $$V \begin{bmatrix} - & lo \\ - & hi \\ + & rnd \end{bmatrix} \rightarrow [+long] \ / \ \_\_\_ \#$$

These long final vowels are almost always marked in the orthography. Many foreign borrowings, such as *unió* ‘(European) union’, indicate the proper vowel length; other

words such as *euro* or *Brno* are still pronounced with a long final vowel despite being written as short. Hence for the few foreign loan words in which vowel length is not indicated, a rule was included to ensure that the final mid vowel is round in these words. Even though the rule only applies to a handful of words in the present dictionary, it is more likely to be of use in more diverse corpora containing loanwords and foreign words.

3.4.7 High vowel lengthening in the primary syllable

High vowels may exhibit variable length in certain syllable positions, and this is likely to be related to the relatively low functional load of high vowels. High vowels are less frequent than other vowels, and the short-long vowel length distinction in high vowels is not used to make any meaningful contrasts. In the initial syllable, high vowels are invariantly long in open syllables. This phonotactic constraint is typically reflected in the orthography but is included here in (3.21) to apply to foreign borrowings such as *unió* [u:nijo:].

(3.21) V[+high] → [+long] / #C0__]σ

To determine syllabification for the purposes of this rule, in the case of a single, intervocalic consonant, the consonant is a member of the onset of the following syllable (V.CV). As noted before, an exception is in some compound words, where lexical similarity overrides the syllabification preference. In the case of intervocalic consonant clusters, VC.CV is the standard, preferred syllabification. Siptár and Törkenczy (2000) claim that Hungarian syllables do not allow onset clusters, although this is rather an

artifact of their analysis. Kenesei et al. (1998:413-5) report that V.CCV is allowed if the CC is rising in sonority.

In order to insert boundaries between syllables in the dictionary, a syllable boundary is inferred after the V if a word boundary or CV sequence follows. The syllable can also be open if a following sequence of CCV occurs in which the CC forms a possible onset according to a lookup table; the set of allowable rising sonority onsets was based on (Kornai, 1990).

The presumed sonority hierarchy in Hungarian is relatively similar to that proposed cross-linguistically:

(3.22) stops, affricates < fricatives < nasals < liquids < glides < vowels
(Siptár and Törkenczy, 2000: 10)

3.4.8 Consonant Shortening

Underlying geminate consonants are always shortened when appearing as part of a consonant cluster. The spelling of stems typically reflects this constraint. Again, processes of word formation are what give rise to geminate-singleton clusters in which the orthography reflects derivation, not pronunciation. Shortening can be found in three distinct situations: compound words such as in (3.23a), derived stems (3.23b), and loan words (3.23c); in each case the underlying geminate is realized as short.

(3.23)	a. orrhang [orhang]	'nasal'	(orr 'nose', hang 'sound')
	b. keddre [kedre]	'by Tuesday'	(kedd 'Tuesday' -re 'LOC')
	c. aggregátum [agregátum]	'aggregate'	

The rule in (3.24) formalizes the generalization stated above.

(3.24) C → [-long] / { __ C, C __ }

3.4.9 /l/-assimilation

The liquid /l/ assimilates to a following /r/ or /j/. The assimilation is typically only word-internal, but it can also occur across word boundaries in fluid speech (Kenesei et al., 1998:438). Representative examples appear in (3.25). Here hyphens represent morpheme boundaries but do not actually appear in written text.

(3.25)

Spelling	Assimilated Form	Gloss
tol-juk	[tojjuk]	'push-DEF.1PL'
gól-ja	[go:jja]	'goal-POSS.3SG'
bal-ra	[barra]	'to the left'
el-rejt	[errejt]	'conceal'

The rule is reflected in (3.26).

(3.26) l → r / __ r
l → j / __ j

3.4.10 Lexical pronunciation exceptions

In order to account for words with irregular pronunciations that cannot be described by the above phonological patterns, the pronunciation dictionary includes a list of hand-coded exceptions. One sub-pattern of exceptions is listed in (3.27) in which consonants are written short but pronounced long.

(3.27) Consonant length exceptions[32] (examples from Nádasdy, 1989a)

Written	Pronounced	Gloss
/egy/	[eggy]	'one'
/egyet/	[eggyet]	'one-ACC'
/lesz/	[lessz]	'will be'
/új/	[ujj]	'new'
/csat/	[csatt]	'battle'
/bridzs/	[briddzs]	'bridge'

Most examples of consonant length exceptions are monosyllabic, and this may be due to a minimal bimoraic constraint for Hungarian (cf. Grimes, 2007). Derived forms such as *egyet* may be based on analogy with the monosyllabic form.

In addition, all cases of the phoneme /dzs/ [ʤ] that occur in intervocalic position or word-final are long. As /dzs/ typically occurs word-initially, there are fewer than ten examples of intervocalic /dzs/ in the language.

Another sub-pattern of exceptions is given in (3.28) and involves back round vowels in loan words that are written short but pronounced long. The lengthening always occurs in an open syllable.

(3.28) Systematic vowel length exceptions in loanwords

Written Short	Pronounced Long	Gloss
kulturális	kultúrális	'cultural'
kulturált	kultúrált	'cultured'
ironikus	irónikus	'ironic'
melankolikus	melankólikus	'melankolikus'
kategorizál	kategórizál	'categorize'

As suggested by the order of presentation in this chapter, exceptional words are the final words in the dictionary to be assigned pronunciations. Hence the exception list overrides

[32] The pronunciations are transcribed in Hungarian orthography. The way to indicate length on consonants represented by digraphs or trigraphs is by doubling the first grapheme.

any rule outputs that may have previously applied to these exceptional words. This exception list functions analogously to a lexical entry overriding a grammatical pattern.

This brings to a close the discussion of the phonology of Hungarian in terms of the major processes not reflected in written Hungarian. I have omitted discussion of Hungarian's most widely known phonological process – vowel harmony – and any other process like vowel harmony that is always clearly marked in the orthography. This is also why the phonotactic constraints discussed in Chapter 2 do not coincide with the orthographic issues discussed in this chapter.

3.5 Implementation of the finite state pronunciation grammar

In this report I have been forced to be somewhat vague about some exact details involved in creating the pronunciation dictionary. The careful reader may also notice that in my description I have switched between orthographies and transcription systems to facilitate discussion of the phonological and orthographic issues. This section is intended to provide a few more concrete details concerning implementation.

3.5.1 Notes on rule ordering

To be clear about what is actually taking place, two sample derivations for *szabadság* 'freedom' and *egyszer* 'once' are given in (3.29). In each of the forms below, three rules apply in the following order: OGOB conversion, voicing assimilation, and finally affrication.

(3.29) Deriving the pronunciations of two words

a. szabadság	→ SabadsAg	→ SabatsAg	→ sabaCCAg
b. egyszer	→ eGSer	→ eTSer	→ eccer

From the output in OGOB encoding, it is possible to convert to and from all the other encodings listed in

Appendix A.

The phonological rules in the preceding section were implemented as string rewrite rules using regular expressions in Perl. An example of a Perl regular expression to handle nasal assimilation is given in (3.30). The use of regular expressions are nearly equivalent in terms of format and function to Chomsky and Halle-style context-sensitive rules (Chomsky and Halle, 1968). Johnson (1972) first observed that traditional phonological rewrite rules can be expressed as regular (finite-state) relations subject to the restriction that no rule may reapply directly to its own output. Hence the rule system implemented in this case is finite state; finite state transducers could equivalently be used to represent phonological rules, which would simplify the procedure of parsing the output of phonological rules in order to obtain the underlying forms.

(3.30) s/np/mp/g;

This command substitutes occurrences of the sequence "np" with "mp". The 'g' character instructs the regular expression interpreter to do this substitution globally – not just for the first occurrence of the pattern in a word.

The pronunciations assigned here for Hungarian are rule-based and do not incorporate statistical heuristics. Church (1986) expresses doubt about the viability of a strictly rule-based approach. According to Church, formulating letter-to-sound rules takes "a few years of intense effort by a highly skilled expert. The end result is often very difficult to debug and to maintain." Church's comments may not apply to Hungarian, but for English, rule-based pronunciation guessers alone have proven somewhat inadequate, and statistical language modeling techniques have also be used in

addition. Fisher (1999) used a combination of statistical and rule-based functions to produce a 94.5% accuracy when compared against the baseline transcriptions in PRONLEX. Neural networks have also been used to guess pronunciations by training on entries found in a commercial dictionary (Sejnowski and Rosenberg, 1987). It is not inconceivable that statistical techniques would be useful for letter to sound rules in Hungarian in the case of proper names, a realm in which Hungarian orthography is "deeper" or more opaque than in general.

3.6 Future developments to the dictionary

Certain phenomena were necessarily overlooked in the creation of the pronunciation dictionary. For each case in this section, a description of both the phenomenon itself and why it was not possible or desirable to implement it is given. Reasons for failing to implement a particular rule range from it representing the wrong dialect or register to it not being able to implement the desired rule due to computational restrictions. In particular, I have not implemented any rule referencing morpheme boundaries. In order to do so requires implementing a morphological parser. Open source morphological parsers for Hungarian are now apparently available (Trón et al., 2005, Trón et al., 2006), but I have not yet integrated these resources into the dictionary creation process.

3.6.1 Long vowel reduction before consonant clusters

Extra heavy syllables (i.e. greater than two moras) are not well-tolerated in Hungarian except across morpheme boundaries. A long vowel in an extra heavy syllable will shorten in certain instances, and this constraint is often abbreviated *VVCC, "prohibit long vowel-long consonant sequences". If the CC consonant sequence has falling sonority, the consonants straddle the syllable boundary. Conversely, if a CC sequence has rising sonority, then the consonants are grouped together into the onset and there is no vowel shortening. It is reported that this shortened vowel is not necessarily always short, but it is certainly shorter than a long vowel would be in a similar environment. I assume that this shortened vowel is a true short vowel, and I do not make allowances for a third gradation in vowel length. Complicating the matter is the fact that judgments about vowel reduction, in my experience and findings, may vary from person to person.

The examples in (3.31a) give forms where vowel reduction of VVCC must take place. Meanwhile, shortening is not necessary in the cases in (3.31b) due to the syllable not being extra heavy; this is presumably because syllabification of the consonant cluster into the following syllable takes place.[33] The divergent behavior of vowels exhibited in (3.31) may result from the contextual treatment consonants as moraic under a phenomenon known as weight-by-position by position (Rosenthall and Van der Hulst, 1999); this is also seen in Levantine Arabic where a coda consonant is moraic following a short vowel and non-moraic following a long vowel.

---

[33] The examples in (31a) and (31b) also differ as to the number of syllables, but I do not believe this fact directly bears on the problem.

(3.31) Vowel reduction according to sonority

Written Form	Pronounced Form	Gloss
a. őrs	örs	'patrol'
gyűjt	gyüjt	'collect'
b. ródli	ródli	'sled'
csúzli	csúzli	'slingshot'

3.6.2 Rapid speech processes

There is an optional process of consonant deletion in triconsonantal clusters (Dressler and Siptár, 1989, Siptár, 1989). As this process is optional, it is not implemented for the dictionary at this time as it is considered only a function of rapid speech. The process causes elision of the middle consonant of a tri-consonantal sequence, and it seems to act most frequently on coronals, as in *mindnyájan* 'all of them' being pronounced [minnyájan] and *kezdhetjük* 'let's get started' as [keszhettyük]. The elision is likely related to constraints on maximal syllable size.

Another rapid speech process involves deletion of a sonorant before a stop consonant with compensatory lengthening on the vowel. Hence there is an optional pronunciation of *zöld* 'green' as [ződ]. This pronunciation is more common in non-standard dialects. I mention the rapid speech processes here only to note that I am aware that they exist but decided to omit them because they are not observed under careful pronunciation conditions.

3.6.3 Non-standard spelling conventions

In corpora containing informal writing styles, such as web-based corpora, some regular, non-standard spellings are found. The non-standard spellings reflect casual pronunciation. However, I opted not to include such informal pronunciations in the

dictionary. Such spellings do not appear in the more formal genres of the corpora I am working with but rather on web pages and in emails. Keeping track of all variant pronunciations would be too difficult and is beyond the scope of this work.

(3.32) Non-standard spellings reflecting phonetic reductions of unstressed syllables

Standard	Non-standard	Gloss
azt hiszem	asszem	'I think (that)'
nem tudom	nemtom	'I dunno'
valószínűleg	valszeg	'probably'
tetszik	teccik	'I like'[34]

3.6.4 Additional future developments

Possible future developments include continuing to compile lexical exceptions to the grapheme-phoneme correspondences I have noted in Section 3.4.10. Interesting data to integrate in the future would be adding typical age of acquisition information for each word or familiarity ratings based on a psycholinguistic questionnaire. Data from confusion matrices indicating the likelihood of the word being mistaken for another lexical item in the language would also be useful; a similarly helpful addition would be to note the number of phonetic "neighbors" a given word has in order to indicate its density in the lexicon in terms of string edit distance.

In the future adding support for divergent dialects would also be very interesting. Unfortunately, this would require a more detailed understanding of the dialect variation than I currently possess. Ultimately such fine phonetic detail might be more appropriately handled by lexicographers.

[34] This example is only a non-standard spelling – not phonetic reduction.

Another more likely enhancement would be encoding suprasegmental information such as secondary stress placement or syllable weight to allow for further exploration of these patterns in Hungarian. For syllable weight, I am curious whether heavy syllables tend to occur in sequences or whether there is a preference for a light-heavy syllable alternation. None of these additions described in this section are planned at this time; rather a need for this additional data for use in a future research project would drive development.

The most likely future development would be to develop a web-based search interface for the dictionary. While I have experimented with distributing the pronunciation dictionary publicly via my university website, a web-based interface would allow more people to take advantage of this work than are currently able. Doing so would also promote comparative work.

3.7 Assessment of pronunciation correctness in the dictionary

In order for the pronunciation dictionary to be a valuable resource to other researchers, it is important to be able to assess the accuracy of the pronunciations in order to ensure quality control. I presented a Hungarian speaker with a list of 150 words, representing approximately 0.5% of total word count of the pronunciation dictionary. The words were chosen by beginning with a random number seed between 1 and 33,660 (the size of the lexicon). The random seed provided an index into the alphabetized lexicon, and later words were chosen at a fixed distance of 224 words from the initial entry (33600/150 = 224) so as to be spread out throughout the lexicon. The list of words appears in Appendix C. The informant was given the OGOB transcription alongside the

orthographical representation; the SAMPA transcription only appears for reference. By sampling the error rate determined by the informant using this list, a confidence estimate could be inferred for the precision of the dictionary as a whole.

Of the 150 words listed in Appendix C, there were three words found containing an error, and each of those contained exactly one error. Two of the three errors involved the morpheme *-ság*/*-ség* '-ness' in the words *igazság* 'truth' and *igazságtalanság* 'injustice'. The *zs* letter combination was incorrectly treated as a digraph indicating the single sound [Z] instead of the correct analysis in which the two sounds straddle the morpheme boundary between *igaz* 'true' and *ság* '-ness'. This error was in theory known to exist because the creation of the pronunciation dictionary did not involve parsing morpheme boundaries. It was unknown, however, how frequently the error might occur. For reference, the morpheme *-ság*/*-ség* '-ness' appeared six times in the 150 word test list – *igazság* 'truth', *igazságtalanság* 'injustice', *sajtószabadság* 'freedom of the press', *boldogságos* 'blessed', *örökség* 'heritage', and *közvetlenség* 'directness'. Only the first two contain an ambiguous digraph sequence, and these two cases are both actually based on the same stem. In response to this error, the pronunciation dictionary was updated to correctly handle all errors related to *-ság/-ség* in the dictionary.

The third error found could be considered an incorrect interpretation of the assimilation rule. While alveolar nasals assimilate to the place of articulation of a following palatal consonant, it may not be the case that a palatal nasal assimilates to the place of articulation of a following alveolar consonant, as in the case of the disputed word néhányszor [nEhAnSor] vs. [nEhANSor] 'a few times'.

After corrections were made to improve the systematic errors found on words in the test word list, a second test list of 150 words was selected at random from the improved dictionary. The second test list appears in Appendix D. The same informant checked the second data set and found no errors. Overall 300 words were presented to the informant, which constitutes roughly 1% of the lexicon we were using.

Unfortunately, while presentation of the informant with a sample list for error checking is likely the best way to get an unbiased assessment of errors, it does have drawbacks. The procedure requires training the informant about the unique transcription symbols used to represent sounds; in a way, this training destroys the unbiased nature of the informant. After training, the informant begins to make the same assumptions the linguist has made, but the informant does not yet possess enough linguistic training to adequately analyze potential problems in their training.

Another issue is that orthography can influence pronunciation correctness judgments. Recall that many Hungarians believe the Hungarian writing system is already phonetic. Nevertheless, using an informant did prove useful (as I am not a native Hungarian speaker) in improving the dictionary.

As a remedy to the deficiencies in the second evaluation method, a final possible evaluation technique suggested to me would involve presenting an informant with computer-synthesized speech based on the transcription appear in the pronunciation dictionary. An open source, free speech synthesizer called Festival contains a Hungarian voice that uses Hungarian biphones as part of Mbrola (Dutoit et al., 1996). I investigated this as a possible solution – to play the informant synthesized speech so that he could assess the accuracy of pronunciation. However, the speech synthesizer solution is

ineffective at present for Hungarian as the imbedded biphones in the program are actually bigrams – Mbrola simply attempts to process written texts as opposed to phonetically transcribed text. It currently lacks an orthographic-to-phonetic preprocessor – specifically the type void that this pronunciation dictionary project seeks to fill. However, even if the technology was working without a hitch, this approach does not seem to be without its own drawbacks. A good deal of synthesized speech sounds unnatural, and informants might cue their judgments to the unnaturalness of the synthesis rather than any problem with a particular pronunciation. I am not convinced such an evaluation method is feasible at this time. Indeed, the overall accuracy of the pronunciation dictionary is sufficient in order to conduct phonotactic research of the nature intended in this dissertation.

3.8 Potential applications of a pronunciation dictionary

This section briefly surveys potential applications of a pronunciation dictionary to phonological research. To be sure, an exhaustive listing of all potential applications is beyond the scope of this chapter, and no one application is addressed in great detail. In this section I only survey applications which are *not* treated later in the dissertation.

3.8.1 Studies in computational phonology

The pronunciation dictionary can be useful in the random, semi-automated selection of lexical materials for a variety of tasks, such as:

(3.33) Uses of the dictionary for lexical tasks

- word recognition or association experiments
- second language instruction
- a lexicon for a text-to-speech system (e.g. Gulikers and Willemse, 1992)
- study of the mental lexicon through the analysis of the distribution of wordlists using several deviation and uniqueness measures
- generation of frequency-based lists of words, graphemes, phonemes or syllables

Many experiments require careful selection of stimulus words in order to have a balanced distribution of words according to length or frequency. In the web-based version of the CELEX database, a tool is available to create lexicons of neighbors, calculate uniqueness points within words, or group words into various word cohorts. Phonological lexicons may serve as the basis for developing rule-based and stochastic grammatical taggers and parsers, spell checkers, or as a training tool for speech recognition and speech synthesis applications.

3.8.2 Phonological neighborhoods and structure of the mental lexicon

Linguists and psychologists have been interested in identifying what constitutes a phonological neighborhood and how a phonological neighborhood is influenced by word frequency (cf. Luce, 1986, Frauenfelder et al., 1993, Metsala, 1997, Luce and Pisoni, 1998, Barlow, 2000, Gruenenfelder and Pisoni, 2006). String edit distance is typically used as a measure of phonological similarity, but new measurements are being proposed in order to compensate for the observation that longer words inherently have fewer neighbors (cf. Kapatsinski, 2006). To this point, research attempting to connect properties of the phonological lexicon to data from language acquisition, speech errors, and word similarity judgments has not adequately addressed how results may diverge in unrelated languages; this is because English has little morphology, few morphemes per

word, and a relatively short average word length. It is not always clear whether and how conclusions based upon English can be generalized. A pronunciation dictionary for Hungarian would be used to compare an agglutinative language with several unique typological properties. Due to the high amount of inflectional and derivational morphology in Hungarian, I expect lexical neighbors to be more heavily influenced by morphological considerations in Hungarian than in English. Additionally, as Hungarian words are significantly longer than English words, the notion of a phonological neighborhood may also need refinement. Finally, a lexicon with frequency data may be used to study a variety of other psycholinguistic tasks (e.g. Jescheniak and Levelt, 1994).

3.8.3 Functional load of segments

Somewhat implicit in much phonological research is the view that all segments have equal standing as phones in the language. Instead, it is often the case that sounds occur at drastically different frequencies and in very distinct phonological contexts. Particular phonetic features may be more useful for contrastive purposes than others. For example, it may be the case that the voicing distinction in English is more important to phoneme recognition (or alternatively confusability) than place of articulation or manner. It would be noteworthy to see what patterns could be established for Hungarian for the sake of cross-linguistic comparison and for research on linguistic universals of articulation.

Elements of functional load were explored in the section in 2.5.1 on uniphone frequency. Another line of exploration would involve investigating the effect of morpheme frequency on the constitution of the lexicon. Specifically, frequent suffixes for Hungarian nouns are *-t* 'accusative' and *-k* 'plural'. In studying the distribution of [t]

and [k], one could expect that the Hungarian lexicon has evolved in order that nominative singular stems do not end in these sounds. This is to avoid confusion with plural or accusative endings; words previously ending in these sounds may have been subsequently reanalyzed. An example of this type of reanalysis or back formation occurred in Old English when *pis* 'pea' (plural *pisen* 'peas') was interpreted to be plural due the -s ending, giving rise to the modern *pea/peas* distinction. For Hungarian, this hypothesis could be tested by comparing the overall frequency of these sounds in all positions and coda positions to their observed frequency in word-final nominative stems.

3.8.4 Applications specific to theoretical phonology research in Hungarian

The pronunciation dictionary is not only useful for making cross-linguistic comparisons, but it is also useful in conducting Hungarian-specific research. A distributional, frequency-based method to determining the sonority hierarchy for Hungarian would be a useful line of investigation. A pronunciation dictionary could also inform the debate on the single or double root node representation of Hungarian geminates or be used to investigate the status of complex onsets in Hungarian (Törkenczy and Siptár, 1999). Concepts such as vowel length in present Hungarian (Nádasdy and Siptár, 1998) could also be investigated, but here a word of caution is necessary. As assumptions about vowel length and assimilation were programmed into the dictionary based on linguistic research, subsequent researchers must be careful to note such assumptions. As some regularities were enforced in the creation of the dictionary, some resulting patterns may be more regular than typically expected in natural language. However, given efforts to check the accuracy, I conclude such cases are quite rare.

This concludes the chapter on the development of the pronunciation dictionary. The focus of the dissertation now turns toward research employing the dictionary to solve specific linguistic challenges. Specifically, I seek insight into Hungarian phonotactics using segment frequency characteristics. In Chapter 4, I examine distribution of phonemes in the Hungarian within the framework of the syllable.

4 Syllable structure and phoneme distribution in Hungarian

This chapter aims to probe the sub-syllabic structure of Hungarian syllables and determine whether they can be described as possessing internal rhyme structure. English, for example, is often cited as a prototypical language with syllables having internal rhymes. Using the same methodologies (described herein) that have supported English rhyme structure, I will examine whether Hungarian internal syllable structure is comparable. I also compare the sub-syllabic structure of Hungarian to that of Korean, a language which has been shown to have syllable structure somewhat opposite to that of English.

In order to make this comparison, I have replicated, in part, studies by Kessler and Treiman (1997) and Lee and Goldrick (2008) that examined statistical properties of English and Korean monosyllabic CVC words. Prior to these studies, earlier research and anecdotal evidence had suggested that for adjacent phones in English syllables, stronger co-occurrence restrictions are found between vowel-consonant sequences than between consonant-vowel sequences (cf. Greenberg, 1950, Fudge, 1969, Selkirk, 1982a, Fudge, 1987, Blevins, 1995). Kessler and Treiman examined 2,001 English CVC words and looked at the frequency of occurrence of the CV and VC subsequences in those words. They determined that there is a significant connection (meaning either an attracting or repelling relationship) between the vowel and a following consonant. Specifically, the frequency of some VC biphone sequences was higher than expected given the base frequencies of those segments, while other VC sequences appeared less frequently than expected. Initial CV sequences, on the other hand, tended to be more equiprobable – the probability of a CV sequence tended to be closer to the product of the

individual C and V probabilities. (This is the general outline of the narrative as has been presented; however, it shall be shown that the argument is more nuanced when the many details are examined.)

The rhyme (alternatively "rime") debate has a long history of back-and-forth dialogue in the literature; the research of Kessler and Treiman sought to put to rest earlier disagreements of Clements and Keyser:

(4.1) There have occasionally been claims to the effect that syllable structure conditions never involve distributional constraints holding between the nucleus and preceding elements, while, on the other hand, they frequently are found to express co-occurrence restrictions between the nucleus and following elements….. Co-occurrence restrictions holding between the nucleus and the preceding elements of the syllable appear to be just as common as co-occurrence restrictions holding between the nucleus and following elements.
(Clements and Keyser, 1983: 19-20)

The debate originates with the comments of Fudge (1969: 272-273), who asserted that for English there exist no constraints between onset and nucleus and that no specific CV sequence should occur with greater frequency than another CV sequence for specifically "linguistic" reasons. Meanwhile, the VC rhyme, on the other hand, is proposed as a subdomain in which repelling or attracting constraints between adjacent phones may apply. This claim extends to languages beyond English – Harris (1983: 16-18) details a number of rhyme-internal co-occurrence restrictions for Spanish, but he found no corresponding restrictions between onset and nucleus.

In summary, a skewed frequency distribution is often used to provide justification for the existence of a rhyme or other sub-syllabic structures. The reasoning is, simply put, that syllable-final restrictions and gaps in segment combinations are evidence of

unexpected distributions of consonants in the syllable. In order to explain this asymmetry, internal structures are posited (cf. Goldsmith, 1990:123-127).

4.1 Brief typology of syllable structure

It is worthwhile to first review the historical development of the various proposals of syllable-internal structure. Attempts to model the structure of a syllable as a tree graph with the syllable node as the root of the tree may go back to attempts to create an isomorphism between phonology and syntax and between syllable and sentence structure (Kurylowicz, 1949). This subsection briefly reviews a typology of proposed syllable structures in order to provide historical context.

4.1.1 The no-structure hypothesis

The most basic and default proposal for the internal structure of a syllable is that it has no internal structure. Due to lack of internal hierarchy, this basic syllable grouping has also been termed *linear*. Here all segments attach directly to the syllable node with no intervening nodes. The no-structure hypothesis, along with most theories of syllable structure, does, however, make typical assumptions concerning association of segments to the syllable node (after Kahn, 1980):

(4.2) a. Each [+syllabic] segment (vowel) is associated with exactly one syllable.
b. Each [-syllabic] segment (consonant) is associated with at least one syllable.
c. Lines associating segments and syllables may not cross. (Segments are assigned to syllables in a linear fashion.)

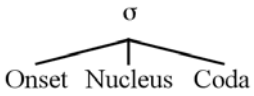
In cases in which a [-syllabic] segment is associated with more than one syllable, it is said to be ambisyllabic. For example, Kahn notes that for the word *atlas* a syllable break

between the two consonants [at.las] is uncontroversial, meanwhile in the case of the word *hammer*, choosing a syllable boundary to occur before or after the [m] is essentially an arbitrary choice. Kahn proposed this autosegmental view of syllable structure in order to solve the ambisyllabic problem. Specifically, he viewed the [m] in words like *hammer* as ambisyllabic, linked to both syllables (ending the first one and starting the second).

4.1.2 Level syllable structure

One additional layer of structure in the syllable collects consonants and vowels into onset, peak (nucleus), and coda categories, terms which were recognized and popularized by Hockett. The onset comprises the syllable-initial consonant or consonants; the nucleus comprises the vowel or peak of the syllable; and the coda comprises the syllable-final consonant or consonants. I will refer to this as "level" or "flat" syllable structure. The onset, nucleus, and coda are not organized hierarchically with respect to one another, but rather all are sisters that share a common syllable node, as depicted in (4.3).

(4.3)



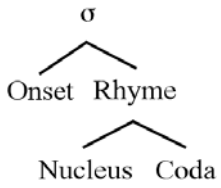
Davis (1988) examined Australian languages to argue for level syllable structure on the basis of onset-sensitive stress assignment. This is in contrast to a more general pattern of stress assignment that is sensitive to rhyme structure and in particular the tendency of heavy syllables to attract stress. I essentially treat level structure (as opposed to flat structure) as the null hypothesis for syllable structure in this dissertation.

4.1.3 Branching syllable structure

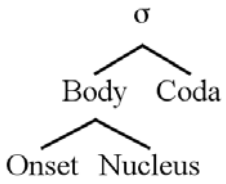
Many linguists assume that syllables with internally branching structure possess rhymes as depicted in (4.4a). Pike and Pike (1947) first suggested the possibility of nucleus and coda forming a constituent. Later, Selkirk (1978) and Halle and Vergnaud (1980) proposed the rhyme as a linguistic universal after observing that phonotactic constraints hold between nucleus and coda (obviating the fact that constraints in general hold between any two given segments).

The other logical possibility of a branching, body-coda structure in (4.4b) has also been suggested for certain languages (cf. e.g. McCarthy, 1979:455, Iverson and Wheeler, 1989). The main concern in (4.4) is whether the syllable peak is grouped with preceding or following consonants.

(4.4a)



(4.4b)



Non-linear syllable structures have also been proposed. For example, grouping non-adjacent segments using distinct vowel and consonant tiers would emphasize cross-consonant phonotactics or relationships between vowels such as vowel harmony. All proposed structures reflect some linguistic observation that a dependency exists between adjacent and/or non-adjacent segments.

4.1.3.1 The onset-rhyme structure

A large body of evidence and literature supports the rhyme hypothesis, although not all of the evidence is conclusive. Davis (1982) classifies types of evidence for rhymes into four distinct categories:

(4.5) Evidence for rhyme structure in syllables (Davis, 1982)
a. The existence of phonotactic constraints between nucleus and coda
b. Reference to the rhyme in stress assignment
c. Reference to the rhyme in other language specific rules
d. The existence of a durational relationship between peak and coda

In this chapter, the arguments for rhyme structure generally revolve around those of type (4.5a). However, in the ensuing few paragraphs I present alternative evidence to remind the reader about reasons rhyme structure is well-motivated.

It has been noted that speech errors often treat rhymes as units (MacKay, 1972, MacKay, 1973, Stemberger, 1983). The most widespread evidence of speech errors justifying rhyme structure is derived from speech spoonerisms, examples of which appear in (4.6). Speech errors are believed to reveal evidence of linguistic structure and of speech planning and organization. In the popular spoonerism examples below, the rhymes of the target words remain intact while the onsets are exchanged:

(4.6)

Utterance	Intended target
We'll have the hags flung out	(We'll have the flags hung out.)
Go and shake a tower.	(Go and take a shower.)
Fighting a liar.	(Lighting a fire)
Our queer old dean	(Our dear old queen)
The Lord is a shoving leopard.	(The Lord is a loving shepherd.)

It is said that, all things being equal, speech errors tend to result in existing words. Unfortunately, the data from speech errors does not *always* support rhyme unity. The following slips of the tongue demonstrate the cohesiveness of the syllable body:

(4.7)

Utterance	Intended target	(from Fromkin, 1971)
cassy put	(pussy cat)	
faust and lawned	(lost and found)	
piss and stretch	(stress and pitch)	

Language games are also often used to support rhyme structure. Examples from Cockney Rhyming Slang and Pig Latin imply major divisions between onset and rhyme. Furthermore, Burmese Disguised Speech (Haas, 1969) and a language game in Bengkulu (Burling, 1970: 136-137) also possess alternations indicating rhyme structure. However, at least one language game exists in which the onset and vowel are treated as a unit – there is a Finnish language game in which the first consonant and vowel of each succeeding pair of words are interchanged (Campbell, 1980). Hence language games can both support and deny the existence of the rhyme. See Davis (1994) for an overview of language games as they relate to arguments for syllable structure.

For English, most recent research supporting rhyme structure is derived from distributional data or speakers' explicit or implicit responses to well-formedness tasks. Treiman et al. (2000) found that English speakers' judgments of the well-formedness of nonce CVC words are affected more by the well-formedness of the vowel-coda sequences than by constraints on onset-vowel sequences. Treiman and colleagues also found that, in a word blending task where English speakers are asked to use the beginning of one monosyllabic word and the end of the second monosyllabic word to form a new word, speakers use the onset of the first word and the rhyme of the second and not the

body of the first and the coda of the second. The tendency is somewhat stronger for high-frequency rhymes than with low-frequency rhymes.

In summary, although there is much evidence for the rhyme in many individual languages, the proposal that rhyme structure is a linguistic universal may not have sufficient evidence. Furthermore, in an individual language the rhyme-type patterning is not always absolute. I next briefly examine evidence of languages with body-coda structure.

4.1.3.2 The body-coda structure

The study of markedness in syllable typology finds that CV syllables are common to nearly every language, and hence it is reasonable to assume some relationship could exist between a C and following V. While the rhyme hypothesis is more prevalent for English, support for body structure is found by examining other languages. For Hebrew, Share and Blum (2005) found that the body (CV) biphone unit is psycholinguistically more accessible than a rhyme (VC) biphone unit. They asked 6- to 8-year-old Hebrew speakers to perform structured and unstructured tasks to test their hypothesis; the tasks included several variations on the theme of splitting a CVC word into two parts to find which biphone unit would be "stickier". The results generally showed the preference to maintain the integrity of the syllable body.

The language that has been the subject of most recent research on body structure has been Korean. Body structure in Korean was initially proposed on the basis of word games, speech and writing errors, and orthography. Later, several psycholinguistic studies provided additional evidence. Yi (1999) conducted a phoneme exchange task for

Korean, structured after a similar task by Fowler (1987). Subjects were presented with pairs of written words, and in some cases asked to exchange the onsets of the two words as accurately as possible; in other cases they were asked to exchange codas. A second variant of the phoneme exchange task used oral word presentation. In both cases, the subjects exchanged codas more quickly and accurately than onsets. This result was interpreted as being consistent with the body-coda syllable organization.

Lee (2006) also conducted a study of Korean to test the body-coda syllable hypothesis. Following the experimental design of Treiman and Danis (1988), twenty-four native Korean speakers were familiarized with lists of six CVC nonwords by having the CVC syllables read to them and asking them to repeat the CVC syllable; mispronunciations were corrected. The participants then listened again to the six nonwords in a different random order. They were asked to orally recall the six syllables and judged on their performance. Lee found that nonwords with high probability CV sequences were recalled at a significantly higher rate (67.13%) than words with high probability VC sequences (43.21%). Again, the salience of the CV sequence was interpreted as an indication of body sub-syllabic structure.

It should be noted that the writing system of Korean is unlike the Roman alphabet system in that Korean allows letters to be arranged left-to-right, or top-to-bottom, depending on the vowel, in a square structure. In this organization, the onset and vowel appear together on top with the coda consonant below (or to the left, depending on the vowel), creating a visually salient association between the two body constituents. It is difficult to completely disassociate the effect the writing system has on speaker's performance. However, the authors of both of the Korean studies cited above made a

point to argue that the effects they found cannot be solely due to the fact that CV sequences are grouped together in Korean orthography. They reasoned that Korean speakers' error patterns did not significantly differ as a function of how the syllables were spelled. To be safe, however, they suggest repeating the tests on preliterate subjects.

4.1.4 Non-hierarchical syllable structure and emergent structure

While the syllable is often represented as a tree graph, this analogy and its associated terminology limits the range of relationships between syllable constituents. One alternate approach is to disregard proposed structures and treat the statistical patterns of co-occurrence as primary – no additional hierarchy is abstracted from these patterns. Under this view, all results derived from psycholinguistics which indicate syllable structure, including speech errors, are modeled using probability density functions or similar techniques.

A hybrid point of view adopted by Lee and Goldrick (2008) is one in which statistical correlations between segments serve as a basis for a type of emergent sub-syllabic structure. This is in line with a trend across linguistics which seeks to attribute classical linguistic structure as emerging from statistical properties of the language (Bybee, 1995, Gupta and Dell, 1999, Seidenberg and Gonnerman, 2000). As applied to syllable structure, categories such as onset, rhyme, or even syllable are not treated as linguistic primitives, but instead are assumed to be created by the language learner as generalizations in response to observations of the distribution of consonants and vowels. This hybrid approach allows for cross-linguistic generalization without appealing to universal grammar. It is to be viewed as a data-oriented, bottom-up creation of structure

based on distributional facts (as opposed to a top-down organization that imposes a relationship between syllable constituents). Under such a view, neither rhyme structure nor body structure is primitive or universal. In any language, either rhyme or body structure is in principle possible.

Finally, some linguists have allowed that rhymes and bodies are not mutually exclusive structures (e.g. Vennemann, 1988). For example, McCarthy (1976) proposed that Estonian syllables have rhyme structure if the coda consonant is an obstruent but have body structure in the case that the coda consonant is a sonorant. Hence, for McCarthy, a single language may have both types of structure. Meanwhile, Donegan and Stampe (1978) were early advocates of allowing both rhyme and body structure for the same word, although they use different terminology than is used here:

> Let us regard the syllables as having two 'slopes', one (the 'rise') including everything up through the syllabic, and the other (the 'fall') including the syllabic and everything which follows it. For example, the rise and fall of [klaonz] are [kla] and [aonz], respectively. The reason for including the syllabic in both slopes is simply that the principles governing both slopes include the syllabic.
>
> (Donegan and Stampe, 1978: 30)

Similarly, Fujimura defined the syllable as consisting of two demisyllables – an initial demisyllable consisting of initial consonants and the vowel and a final demisyllable consisting of the vowel and final consonants (Fujimura, 1976). In fact, Fujimura considered demisyllables and not phonemes to be the atomic units of the syllable. I return to the question of whether rhymes and bodies can co-exist at the end of this chapter after investigating the statistics of segment sequences in Hungarian.

4.2 Hungarian syllable structure

I now address Hungarian to determine the nature of its sub-syllabic structure. A number of research papers address Hungarian syllable structure, but I have found little research contrasting rhyme structure versus body structure; the following is a short survey of Hungarian research addressing syllable structure.

A series of papers have dealt with the topic of whether the onset can be branching in Hungarian (Törkenczy, 1989, Szigetvári, 1999, Szigetvári, 2001). Branching here means having two or more segments, and the term "complex" could also be used. Historically, Hungarian did not have word-initial consonant clusters, but acquired them under language contact and borrowing of loanwords. As noted previously, Siptár and Törkenczy (2000) do not consider the existence of words beginning in consonant clusters as evidence of complex onsets in Hungarian because these clusters are not found word internally. One proposal is to treat these consonants at word margins as syllable adjuncts, which allows one to preserve the uniformity of a simplex onset across all syllables. There are 50 possible CC complex onsets in Hungarian. Most of the CC onsets have s or š as their initial segment; common CC onsets not involving s or š and having more than 30 unique occurrences in words in the pronunciation dictionary are pl, pr, tr, kl, kr, bl, br, dr, gr, fl, and fr.[35] Consonants which are permitted to appear as part of a CCC cluster are str, skl, špr, štr, and škr (cf. Siptár and Törkenczy, 2000: 98-99).

As for the rhyme issue, Siptár and Törkenczy (2000: 9) assume that Hungarian sub-syllabic constituents are the onset, nucleus, rhyme, and coda. The rhyme is taken to

---

[35] I consider onsets occurring in fewer than 30 words to be marginal. The number 30 is somewhat arbitrary, but using it separates the presumably well-formed onsets from those involved in a few scientific terms and foreign borrowings such as [pt], [ng], [zr], [ft], or [gv].

be branching, as are the nucleus and coda nodes. While Siptár and Törkenczy describe properties of the rhyme, it is difficult to ascertain concretely why this structure is adopted. They do describe (2000: 104) one phonotactic constraint particular to the rhyme which has been mentioned earlier in this dissertation in Chapter 2:

(4.8) Vowels preceding the nasal + stop clusters /mp, mb/ must be rounded if the vowel and the entire consonant cluster are within the same rhyme.

There are not many examples of such words, but the list includes *különbség* 'difference', *gömb* 'sphere', *tömb* 'block', *gomb* 'button', *comb* 'thigh', *domb* 'hill', *lump* 'carouser' and *krumpli* 'potato'.[36] Again, note that if the consonant cluster straddles a syllable boundary, then the vowel does not need to be round. For example, *ember* 'human being', *csempész* 'smuggler', and *templom* 'church' are syllabified such that the entire consonant cluster is not contained in the same rhyme (*em.ber*, *csem.pész*, and *temp.lom*) and hence do not meet the structural prerequisites for the constraint to apply. I do not view the existence of a constraint such as (4.8) as particularly strong evidence for a rhyme structure – while the segments under consideration consist of a nucleus + coda sequence (thus rhyme), this could rather be seen as a fact about syllable boundaries in Hungarian phonology rather than a phenomenon particular to the rhyme.

Some Hungarian linguists have followed a research program attempting to reduce all syllable structure in Hungarian to a strict CV skeleton within the framework of strict CV phonology (Lowenstamm, 1996), a descendant of the theory of Government Phonology in which all words are generated from one or more CV units. In this

[36] Many of these examples are loanwords, and the size of this category indicates the relative unimportance of this particular constraint.

framework, a consonant cluster arises from a $C_1V_1C_2V_2$ sequence in which V_1 is not licensed (yielding $C_1C_2V_2$), and similarly a long vowel arises from a $C_1V_1C_2V_2$ in which C_2 is not licensed ($C_1V_1V_2$). In Government Phonology, the coda constituent does not exist. As there is no clean solution of how to express sub-syllabic structures such as the rhyme in this framework, it is not surprising that the rhyme question is not prominent, and nothing similar to a statistical study like Kessler and Treiman's has been conducted. As an aside, it is interesting to observe how the theoretical framework, here Strict CV and Government Phonology, can influence the nature of the research questions.

Finally, it is typical to find statements in grammars of Hungarian to the effect that segments freely co-occur in the language and a wide range of syllable types are possible. Indeed, Siptár and Törkenczy remark that within the rhyme there is "no restriction on nuclei in branching or non-branching rhymes in Hungarian: any vowel can occur in a closed or an open syllable" (2000: 104). I wish to examine whether this generalization is actually true or whether a more nuanced stance must be adopted.

4.2.1 Distribution of voiced consonants in the Hungarian syllable

In examining the distribution of phonemes in Hungarian syllables, I will initially restrict my study to monosyllabic CVC words. The choice to use CVC words for this study was primarily for ease of comparison to previous results. There are a number of reasons earlier researchers chose to only examine CVC monosyllables. First, stress assignment is consistent in single syllable words, and so confounding issues of stress placement can be safely disregarded. A problem posed by multisyllabic words is that segments can be ambisyllabic or have ambiguous syllable constituency. Additionally, some have

suggested that some or all word final consonants should not be considered part of a syllable (e.g. Kenstowicz, 1994: 260-261). Finally, studying single syllables allows for a more straightforward statistical analysis.

In Hungarian there are not any absolute restrictions on the nature of the initial or final consonant. For example, many languages have devoiced final consonants, and such a restriction should be considered a word-level constraint and not a syllable-level constraint. Here I briefly investigate this issue with regard to voicing as a prerequisite to emulating the Kessler and Treiman (1997) study.

In response to the concern of word-level restrictions being impossible to disentangle from syllable-level restrictions, below in (4.9) I examine the distribution of voiceless consonants in Hungarian to ensure that they are permitted in all positions. The reason for doing so is that cross-linguistically voiced segments tend to appear more frequently in intervocalic position while voiceless segments appear at word margins. The following data are derived from the pronunciation dictionary (Chapter 3), which includes corpus frequency data for each lexical item.

(4.9) Frequency of voiceless consonants at word edges in Hungarian

	Type Frequency	Token frequency
Word-initial position (only CVC words)	50.4%	45.9%
Word-initial position (all words)	53.5%	51.2%
Word-final position (only CVC words)	34.4%	17.8%
Word-final position (all words)	55.1%	42.7%
All positions throughout word	42.9%	39.7%

The data in (4.9) should generally dispel the notion that there are any categorical restrictions on the distribution of voiceless consonants – voiced and voiceless consonants are permitted in all positions. Nonetheless, there are a few observations to be made.

First, words containing voiceless consonants tend to be less frequent – the token frequency of such words is always lower than the corresponding type frequency, indicating that the word type is underrepresented in corpora. Second, a voiceless consonant is more likely to appear on the periphery of the word (in initial or final position) than word-internally. I make this claim because frequencies of word-initial and word-final voiceless consonants (type or token) are higher than the frequencies of voiceless consonants in all positions. This is not surprising – consonants in environments such as intervocalic position are under pressure to be voiced.

A last observation based on (4.9) is that a word-final consonant in CVC words is voiced about twice as often as it is voiceless. The effect of word frequency in this case is also remarkable – CVC words with voiceless consonants are not frequent. The only explanation I can offer for this fact is that voiced coda consonants may be more likely to be moraic (i.e. contribute to syllable weight) than voiceless consonants because of their increased duration. In the CVC context, a word without a moraic final consonant may not meet the requirements of the Minimal Word Condition – see Grimes (2007) for more on this issue.[37]

I now turn to replicating the study on the distribution of consonants for Hungarian as previously done by Kessler and Treiman (1997) for English.

---

[37] In Grimes (2007) it was proposed that all content words in Hungarian must have at least two moras. It was also observed that word-final consonants do not appear to contribute to syllable weight, unlike word-internal consonants at the end of a syllable. Hence it was proposed that CVC words do not meet the minimum length condition to be valid content words of the language. While many CVC content words do exist, including examples ending in voiced consonants, the paper illustrates that they are not as common as statistically expected.

4.3 Methodology

In selecting a word list to study, I chose to examine the 678 CVC words found in the Hungarian National Corpus (HNC) in order to maximize the number of words under examination. There were 556 CVC words in the Hungarian pronunciation dictionary (see Section 3.2.1 for relative sizes of the corpora). The comparable studies for English contained many more CVC words – Kessler and Treiman's study was based on 2,001 CVC words, while Lee and Goldrick's study was based on 2,521 English CVC words in the CELEX database. The number of CVC words in Lee and Goldrick's study of Korean is closer to Hungarian – 940 in total.

According to Lee and Goldrick, the asymmetry between English CV and VC sequences still holds when considering only a reduced set of the 940 most frequent CVC words in English. The number of English CVC words cited in other studies – 2,521 and 2,001 – is somewhat misleading. In my examination of the English CELEX2 database, I found 2,430 monomorphemic CVC words (out of 2,613 CVC words total). However, after eliminating duplicate homophones from this list such as *bat* 'flying mammal', *bat* 'wooden club' and *bat (an eye)* 'to blink', only 674 unique monomorphemic word forms remain. By contrast, the Hungarian list has no homophones. Allowing homophones in the list artificially increases the word count without creating unique phone sequences, and it is akin to allowing a backdoor way for token frequencies to be included in the type frequency count. When excluding homophones from English CVC monosyllable word counts, Hungarian and English actually contain comparable numbers of CVC monosyllables in their lexicons.

Kessler and Treiman attempted to cull any monosyllabic word that was polymorphemic; hence *this* and *that* were omitted on the basis that "th" could be a demonstrative morpheme. They also omitted words with foreign phonemes or accented letters. I made no such attempt to restrict the Hungarian word set for this study.

4.4 Results on investigations of Hungarian sub-syllabic structure

4.4.1 Experimental frequency results

The frequencies of the consonants and vowels for CVC words are given in (4.10). Only word types were considered, unweighted by their frequency. The vowel frequency column sums to 678, while the consonant column sums to 2 * 678 = 1356, as each word contains exactly two instances of a consonant.

A comparison of the uniphone frequencies of CVC words to the uniphone frequencies of the entire language at large (data which appeared earlier in Section 2.5.1) shows overall patterns of uniphone frequency are consistent between monosyllables and the lexicon at large. For example, the five most frequent consonants in an unrestricted full corpus are /l t r k ʃ/; when examining only CVC syllables the most frequent consonants are /r l k t s/. This provides some encouragement that examining CVC segment frequencies could potentially be representative of patterns of the broader lexicon.

(4.10) Frequencies of consonants and vowels in 689 CVC words

Consonants	Frequency
r	119
l	96
k	89
t	86
s	81
ʃ	80
m	73
b	70
n	66
j	61
g	60
p	57
h	56
d	57
v	52
f	47
z	44
tʃ	41
d^y	40
n^y	33
c	24
x	10
ʒ	9
t^y	3
dʒ	2

Vowels	Frequency
á	93
é	90
a	81
e	81
o	63
ú	44
i	38
í	37
ó	37
ö	33
u	28
ő	24
ű	21
ü	8

4.4.2 Preferences for onset and coda distribution

I now examine whether certain segments tend to appear in a particular syllable position. The table in (4.11) shows the frequency of consonants in the onset and coda position. The table is arranged according to the strength of association – that is, consonants which have distributions skewed towards primarily initial or final position are listed first, while consonants with balanced distributions appear at the bottom of each list. In order for

comparison to English, Kessler and Treiman's consonantal distribution data is included in Appendix E.

(4.11) Distribution of consonants in descending order of strength of association

Onset or no preference		
Phone	Onset	Coda
h	56	0
ʤ	2	0
f	41	6
b	53	17
v	37	15
ʧ	26	15
s	49	32
n^y	19	14
m	41	32
ʒ	5	4
c	12	12
ʃ	40	40
t	43	43

Coda preference		
Phone	Onset	Coda
x	0	10
z	10	34
n	18	48
r	34	85
g	18	42
t^y	1	2
j	21	40
l	37	59
d^y	18	22
p	26	31
d	26	30
k	44	45

From (4.11) it is clear that there is an association between consonant type and syllable position. At the outset, it was already known that /h/ and /ʤ/ can only appear in onsets and /x/ (an allophone of /h/) in the coda – this is common knowledge. However, the skew present for other consonants has not been generally recognized – for example, /z/, /n/, /r/, and /g/ show strong preference for appearing in the coda, while /f/, /b/, and /v/ have strong preference for onset position. In fact, excepting /p/, all labials show a preference for onset position. This is a peculiar fact that I have not seen noted by others previously.

Treiman and Kessler found that in English, coronals show a strong preference for coda position. They claim that "when languages restrict codas or word endings to consonants of a particular place of articulation, anterior coronals are the least likely to be excluded." Spanish is further cited as having a core vocabulary consisting mostly of words ending in coronals despite a wide range of other consonants being frequent at the beginning of words. Hungarian has similar tendencies – 54.6% of coda consonants are coronals compared with only 38.7% of onsets.

Allow me to drill down further into the question of coronal consonant distribution. Berg (1994) conducted a statistical study of VC sequences in British English. He found that long vowels tend to precede short consonants. In this context, short consonants do not mean non-geminate singletons, but rather coronal consonants. It turns out that for Hungarian, coronals are somewhat more likely to appear after a long vowel than after a short vowel in CVC words, but there is not much effect in the majority of other conditions. The table in (4.12) summarizes their distribution.

(4.12) Likelihood of coronal consonants after a short (V) or a long (VV) vowel

	V	VV	Total sample size
Final coronal in monosyllable	51.5%	58.9%	687
Final coronal across all words	70.7%	71.6%	22698
Coronal appearing as first C in word-internal CC cluster	69.4%	69.4%	25735

Many disyllabic nouns in Hungarian end in -t, as this suffix is often used to derive nouns from verbs. However, because the table in (4.12) shows that a final coronal is almost equally likely after a long or short vowel, we surmise that the distribution of final coronal does not seem to be conditioned by the length of the preceding vowel. Hence, unlike for British English where Berg reported that long vowels tend to precede short consonants, it

appears that Hungarian does not use place of articulation as a cue to consonant length (again, here meaning coronality); Hungarian already possesses a more robust short-long consonant distinction, namely the singleton-geminate distinction.

4.4.3 Strength of association of vowel to preceding and following segments

The previous section provided results concerning the distributions of consonants in the syllable, but it only indirectly addressed the question of whether Hungarian has rhyme or body syllable structure. This section is designed to determine whether the vowel and the coda or the vowel and the onset are more closely associated. Before I do that, however, I must introduce the statistical metrics used.

4.4.3.1 Statistical measures of association

The strength of the association between adjacent phones is assessed here using correlation coefficients. I follow the strategy used by Lee and Goldrick and others in analyzing dichotomic data by using the normative measure of contingency, r_φ ("r phi"). It is a correlation statistic comparable to Pearson's *r* and provides a correlation value between -1 and 1, with 1 being perfect correlation, 0 being no correlation, and -1 being inverse correlation. Perruchet and Peereman (2004) surveyed a number of statistical measures, including simple co-occurrence frequency, forward transitional probability P(C/V), and backward transitional probability P(V/C); they showed that the contingency between Vs and Cs in French rhymes is best assessed using r_φ. They determined that r_φ correctly predicted judgments of word-likeness by children and adults as a function of the frequencies of rhyme segments, and hence it is useful to adopt it here as well. As I do not

have judgments of word-likeness by children and adults for Hungarian, I assume that whatever property of French that allows for r_φ to be a good measure of contingency also holds for Hungarian. Note that a more thorough evaluation to determine the best measure of contingency for Hungarian would require word-likeness judgments from Hungarians; such judgments would be useful to obtain in the future but are presently unavailable.

The statistic is defined as follows for CV sequences but is similar for VC sequences. I treat a CV biphone sequence as a sequence of two events, a consonant event and a vowel event – let us refer to these as C_i and V_j. From this I create a 2 x 2 contingency matrix, depicted in (4.13), where *a* stands for the number of C_iV_j occurrences, *b* for the number of occurrences of C_i followed by a vowel different from V_j, *c* for the number of occurrences of V_j preceded by a consonant different from C_i, and *d* for the number of onset-nucleus events comprising neither C_i nor V_j.

(4.13) A contingency matrix for a CV biphone event

		V_j	
		+	-
C_i	+	a	b
	-	c	d

Based on the contingency matrix above, the r_φ correlation coefficient is now defined according to the following formula:

$$(4.14)\quad r_\varphi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

This represents the two-way dependency between the C and a following V. Alternatively and equivalently, r_φ can be expressed as the geometric mean of the forward and backward

transitional probabilities. The correlation statistic for a VC biphone sequence is defined analogously.

4.4.3.2 Strength of association of biphone sequences in the Hungarian syllable

I can now use the r_φ correlation coefficient to assess whether there are more restrictions on CV or VC sequences in Hungarian words. In Appendix F, I have included the most frequent CV and VC sequences found in CVC words. For reference, I have also included as Appendix G the most frequent CV and VC sequences across all words.

The analyses contained herein use the absolute (non-negative) value of r_φ, as the strength of the association is paramount here and not whether the correlation is positive or negative. That is, in seeking evidence of either body or rhyme structure, attracting or repelling tendencies between adjacent phones are considered to be equally important, and I am interested primarily in the strength of associations and not the overall directionality. Later I will discuss the interpretation of taking the absolute value of r_φ and whether it is reasonable.

Following Lee and Goldrick, the absolute value of r_φ was computed for all 235 CV and 219 VC sequences attested in the Hungarian CVC wordlist. The results are presented below in (4.15), along with the comparable results from the Lee and Goldrick study for English and Korean for ease of comparison. I draw the reader's attention to the "Mean absolute r_φ" column. The higher value of mean absolute r_φ for Korean CVs indicates greater dependencies within CV sequences, and it is this value that Lee and Goldrick interpret as implying body-coda syllable structure. The situation is reversed for English – higher mean absolute r_φ values are found for VC sequences, which suggests

onset-rhyme syllable structure. For Hungarian, the mean absolute r_φ values are similar, suggesting for neither rhyme nor body dominance in Hungarian.

(4.15) Mean absolute value for r_φ for CV and VC sequences in Korean and English

		N	Mean absolute r_φ	Mean type frequency
Korean	CV	152	0.05	2.40
	VC	76	0.039	6.19
English	CV	280	0.022	7.02
	VC	222	0.034	7.08
Hungarian	CV	235	0.032	2.89
	VC	219	0.034	3.10

Lee and Goldrick measure the statistical significance of differences in mean absolute r_φ between CV and VC. The non-parametric Mann-Whitney U test was used because it does not require normality of the data, unlike the t-test. The test showed that the differences between strengths of constraints is statistically significant for Korean ($U=6723$; $p < .05$) and English ($U=21{,}582$; $p < .0001$). For the Hungarian CVC word forms, the mean absolute r_φ values of 0.032 and 0.034 are close to identical, and no statistical significance was achieved ($U=27{,}206$; $p = .27$). Also note that in general, the mean absolute r_φ values for Korean are higher than for Hungarian and English – this indicates that in Korean consonants tend to appear exclusively in either the onset or the coda. Meanwhile English has the most balanced consonant distribution of the three languages, based on overall lower r_φ values.

As I indicated earlier, Lee and Goldrick's mean absolute r_φ measure takes the average of the absolute values of r_φ scores. However, another approach is to keep

positive (attracting) and negative (repelling) values separate. For example, the mean absolute r_φ value of 0.032 for Hungarian CVs is a weighted average of the absolute values of 0.038 and -0.018. Similarly, the VC r_φ value of 0.034 is a weighted average of 0.042 and -0.020. In both cases when I am measuring the co-occurrence of phones, it is the attracting relationships (positive values) which serve to raise the weighted average and contribute more to the r_φ average. Later in the chapter I offer an explanation for this relationship.

4.5 CVC versus entire phonological words

I observed earlier that previous segment distribution and syllable structure studies have primarily examined CVC forms. While the simplifying assumption to restrict the study to monosyllabic words may make the study design simpler, it risks conflating syllable and word properties. Because syllable and word boundaries coincide for monosyllables, what seems to be a property of the syllable may actually be a property of the word; polysyllabic words must be examined to resolve this ambiguity (cf. Davis, 1989b). Despite these concerns, however, Lee and Goldrick point to a study of English (Thorn and Frankish, 2005) which demonstrated a significant correlation for biphone frequency between the CVC and full word lexicons. The Thorn and Frankish study suggests that English CVC forms may have a phone distribution representative of the entire lexicon.

However, concerns over differences between monosyllables and full words remain. Thorn and Frankish were not comparing rhyme versus body structure using r_φ, for example. I was concerned that monosyllabic words risk being not representative of the entire language, and hence I expanded my investigation to include all lexical items for

both Hungarian and English; both languages were included first to confirm that the Thorn and Frankish results generalize to the full English lexicon and then to test this pattern for the full Hungarian lexicon. Part of the issue with using polysyllabic words is that syllable boundaries are not always annotated correctly or predictable, and hence in my examination of longer words I give results for both full syllables and for simple CV and VC biphone sequences.

The results of the Thorn and Frankish study are supported by my investigations of all CVC syllables in the English CELEX2 database. I examined all CVC syllables in two environments: in monosyllabic words and in all words. After eliminating homophones, there were 674 CVC monosyllables and 4,672 CVC syllable tokens in all words.

The consonant strength of association results for CELEX2 are given in (4.16). Note the considerable range of variation in mean absolute r_φ across the different conditions; also note that $+r_\varphi$ and $-r_\varphi$ are given for reference.

In addition to CV and VC sequences in CVC syllables (4.16a,b), I also report data for all CV and VC biphone sequences regardless of syllable type in (4.16c). Not surprisingly, the "N" (number of each type) is dramatically higher in this condition. As there are 30 consonants and 24 vowels (when including diphthongs) in English CELEX2, in theory the upper bound on N is 30 x 24 = 720. However, if a given consonant, say, never appears in onset position, then certainly no CV sequence containing that consonant has a chance of appearing. Hence in parentheses in (4.16) I give a more realistic upper bound for N based on the product of the number of actually occurring segments.

(4.16) Mean absolute value for r_φ for CV and VC sequences in English

		N (possible)	Mean absolute r_φ	+rφ	-rφ	Mean type frequency
a. English Celex2	CV	117 (161)	0.041	0.048	-0.020	5.7
only monosyllable cvc	VC	93 (140)	0.056	0.069	-0.026	7.2
b. English Celex2	CV	139 (168)	0.035	0.044	-0.019	33.6
all cvc syllables	VC	108 (140)	0.055	0.068	-0.032	43.3
c. English Celex2	CV	505 (624)	0.019	0.027	-0.013	247.7
all biphones	VC	439 (624)	0.022	0.03	-0.014	281.5

In all cases there were statistically significant stronger associations between the vowel and a following consonant.[38] This is in keeping with previous findings for English. For the all biphones case (all words in the lexicon), there is a greater variety of CV (505) and VC (439) sequences. The greater sample size seems to have averaged away many of the differences in associations seen in the monosyllabic cases. Hence the findings are not necessarily parallel when attempting to scale up to the multi-syllabic word, and the fact that subsyllabic distributions vary depending on word length should be kept in mind whenever data is only reported for the monosyllabic lexicon.

One consequence of (4.16) and shown throughout the results section is the failure of statistical measures to strictly partition languages into solely rhyme or body-based syllable structures. While there are more restrictions for the syllable rhyme than syllable body in English, this is not as absolute a constraint as has been portrayed. In fact, it is not hard to see how the back and forth between Fudge and Clements and Keyser emerged.

---

[38] The results of statistical correlation tests are as follows:
(16a): U=9,194, p<.005 ; (16b): U=6,636 ; p<.01 ; (16c): U=199,624; p <.02

Clements and Keyser's claim (quoted earlier in this chapter in (4.1)) that there are as many restrictions between onset-vowel as vowel-coda could almost be considered qualitatively appropriate – it is certainly not the case that relationships only exist between rhyme segments, even though there are a higher proportion of such relationships in English.

The investigation of full-length English CVC syllables and biphones can serve as a basis for comparison to Hungarian. Biphones come in four broad types: CV, VC, CC, and VV. For Hungarian, the VV and CC sequences are first discarded (as I am interested in investigating onset-nucleus and nucleus-coda transitional probabilities only) and the remaining CV and VC sequences contribute to frequency counts. The biphone counts appearing in (4.17c) are not a direct measure of syllable structure (as some VC sequences, for example, may span a syllable boundary), but instead they are a measure of strength of association between vowels and consonants. In (4.17b) I extracted all Hungarian syllables from full-length words matching the type CVC, which were determined using the syllable parsing algorithm described in Chapter 3. The results are presented alongside the earlier data on Hungarian monosyllables from (4.15) – repeated below as (4.17a) for easier comparison. The mean absolute r_{φ} values are similar to one another in each condition, but show a considerable degree of variation across conditions:

(4.17) Mean absolute value for r_φ for CV and VC sequences in Hungarian

		N	Mean absolute r_φ	$+r_\varphi$	$-r_\varphi$	Mean type frequency
a. Hungarian	CV	235	0.032	0.038	-0.018	2.89
monosyllable cvc	VC	219	0.034	0.042	-0.02	3.1
b. Hungarian	CV	320	0.021	0.028	-0.014	143.2
all cvc	VC	292	0.028	0.039	-0.017	157
c. Hungarian	CV	334	0.019	0.025	-0.014	296.2
all biphones	VC	329	0.022	0.03	-0.015	297

For each of (4.17a,b,c), none of the differences in mean absolute r_φ between CV and VC achieve statistical significance using Mann-Whitney; nonetheless there is a general pattern of fewer restrictions across CV sequences compared to VC sequences that is consistent if not statistically significant. Note that both the mean absolute r_φ values when considering all biphones in (4.17c) are lower than when examining CVC syllables (4.17a,b) – this was also seen for biphone sequences in English. This indicates that when examining the lexicon as a whole without regard to syllable divisions, consonants demonstrate a more balanced distribution. Examining the larger lexicon has a smoothing effect on any idiosyncratic distribution demonstrated in the CVC monosyllables; the CVC syllable distribution, meanwhile, may exhibit greater variation and can be skewed by a small handful of word forms.

One other meaningful distinction emerges when comparing these data. If there are harder phonotactic constraints for CVC syllables than for biphone, this indicates the language may have a syllable domain and a constraint that applies in that domain. If the

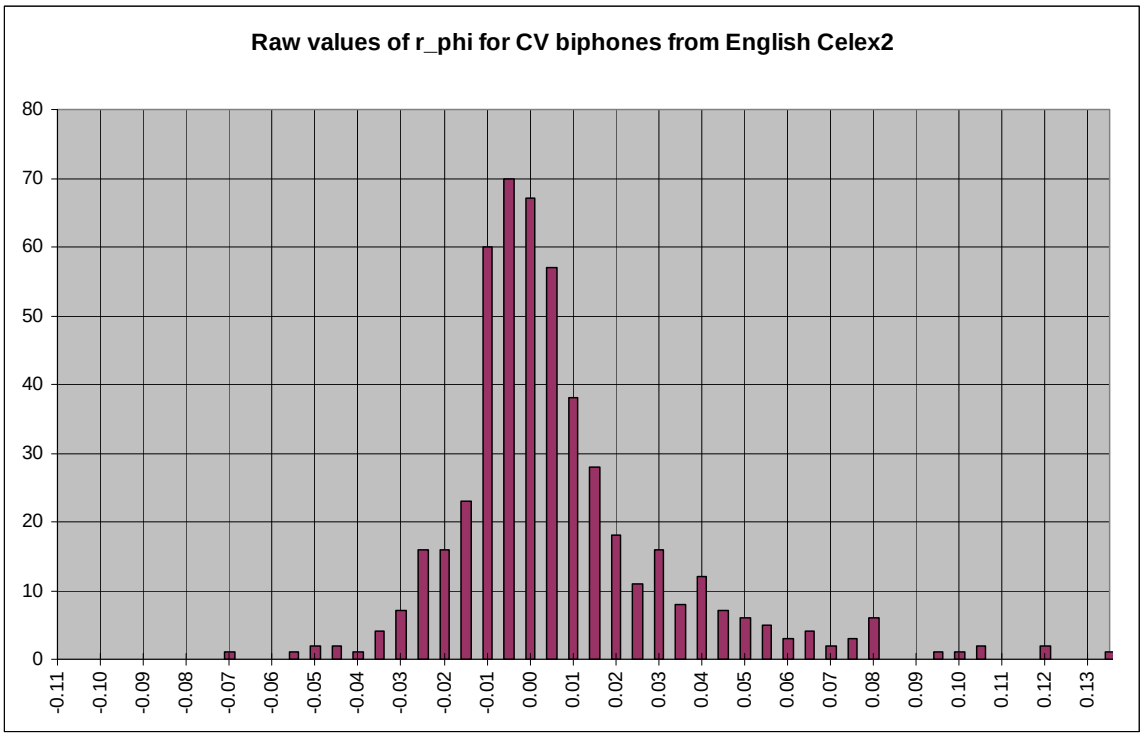
language shows no significant difference between syllable r_φ and that of all biphones, then the language does not have a syllable constraint and the relationships governing segment sequences hold regardless of the position of the segments with respect to syllable boundaries.

In summary, the patterns attested in CVC words are not always generalizable to full phonological words or at least are not as robust – this is another reason researchers seeking to find evidence of the rhyme have avoided longer words. Attempting to study all syllable types (not just CVC) and longer words can be confounding. Nevertheless, it is necessary to look beyond the syllable and consider the entire lexicon if the results are to be considered of broad importance and influential on syllable structure theories.

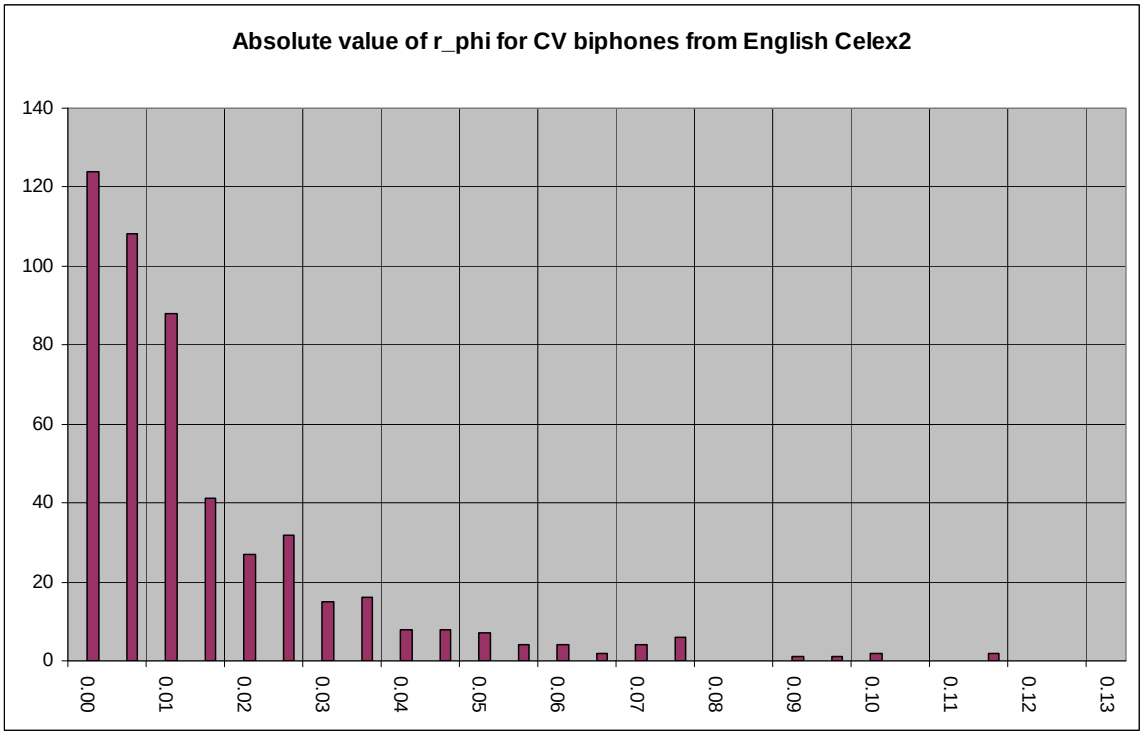
4.6 Distribution of r_φ values

The use of r_φ in order to calculate strength of association values is not one that is familiar to many phonologists, and it is unclear how Lee and Goldrick applied this measure and whether its use has been appropriate and straightforward. If the reader has been examining the positive and negative values of r_φ before the absolute value and average is taken, it can be observed that the mean of the positive r_φ values is usually greater (in absolute value) than the mean of the negative r_φ values. This can be seen by examining the graphs of the distribution of these r_φ for a few instances of data reported earlier in this chapter. Below, for CV biphones in English, I examine both actual r_φ values (which may be either positive and negative) and the absolute value of r_φ values (that is, their positive magnitude).

(4.18) Raw values of r_φ for CV biphones from English Celex2



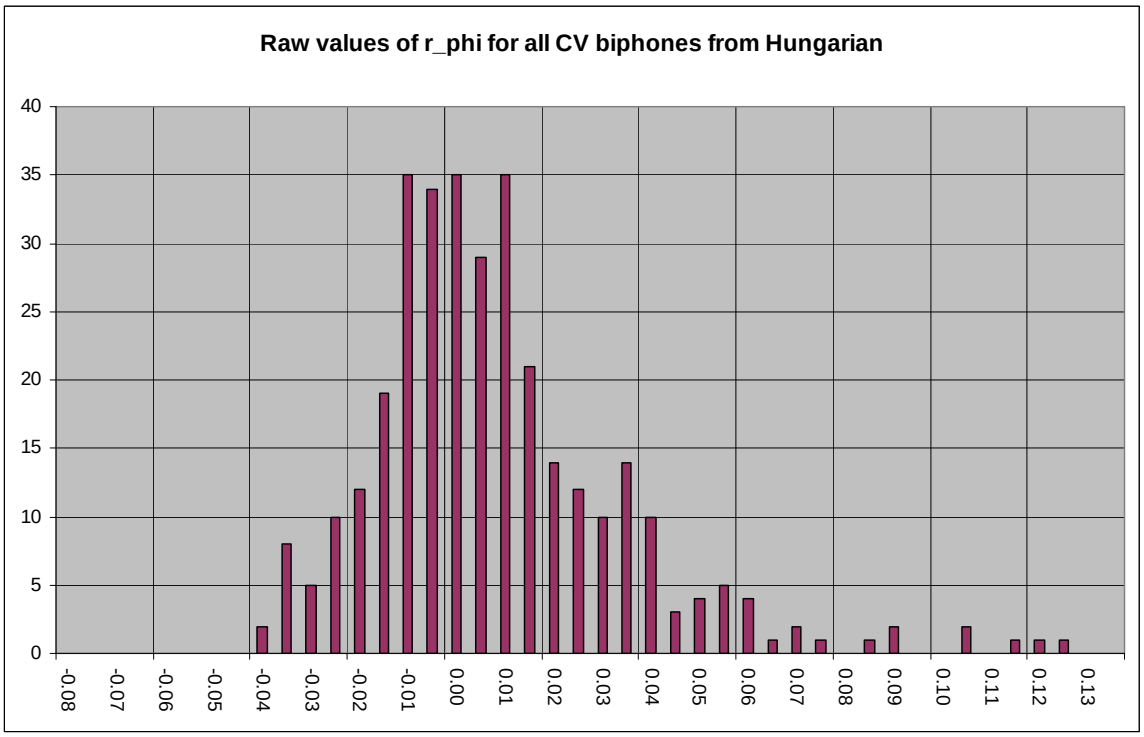
(4.19) Absolute value of r_φ for CV biphones from English Celex2



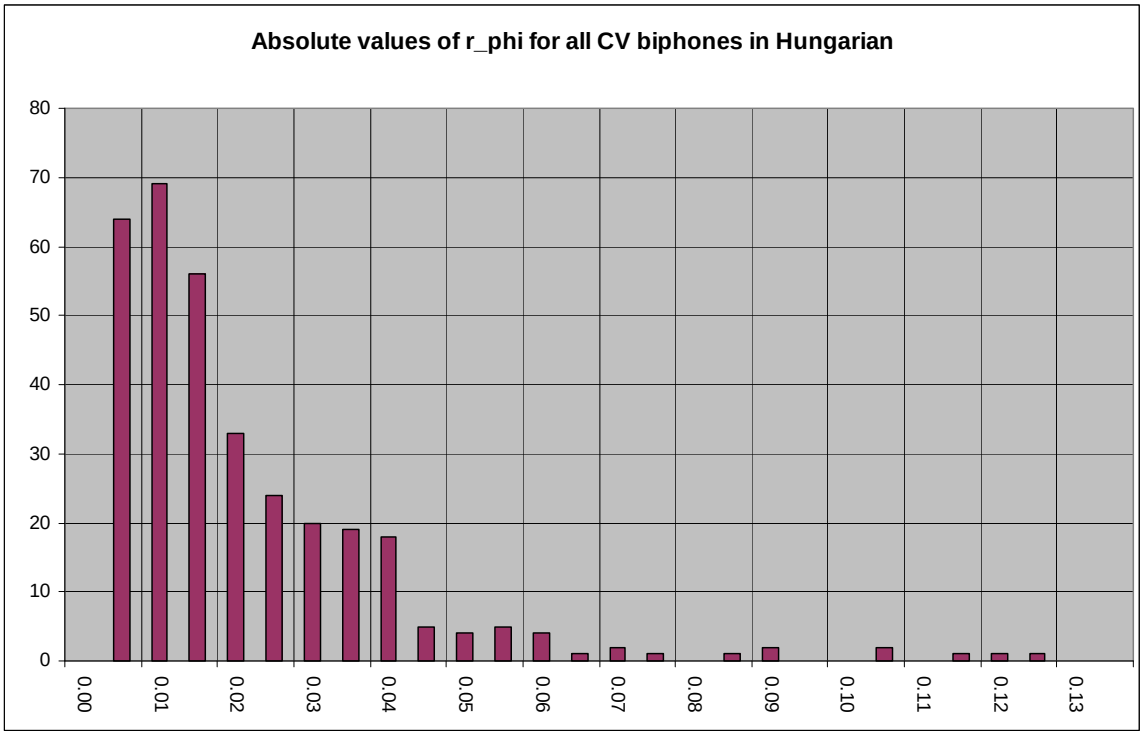
There is some skew to the distribution that is obscured by taking the absolute value. This same pattern is repeated in other data sets. The bar charts here present histograms for CV r_φ values; for comparison I have included VC bar charts at the end of the dissertation in Appendix H.

To demonstrate the nature of the r_φ distributions for another langguage, I also present graphs of the r_φ distribution for all CV biphones in Hungarian.

(4.20) Raw values of r_φ for all CV biphones from Hungarian



(4.21) Absolute values of r_φ for all CV biphones in Hungarian



The distribution of r_φ is not quite centered on the origin. That is, the mean of the raw r_φ values can fall anywhere between 0.007 and 0.040 across the various data sets I examined. From this, I surmise that attracting relationships between segments are more prevalent than repelling relationships in English and Hungarian; there do not appear to be a balanced number of each type. This indicates a predetermined relationship between the median and mode in my dataset – most biphones contain two phones whose appearance together is positively correlated, while only few biphones contain phones which are negatively correlated.

4.7 Categorical versus gradient biphone constraints

This study has not attempted to distinguish between the linguistic nature of the gaps and patterns in phone distribution throughout the syllable. There are essentially two types of gaps. A *lexical* gap is accidental and not prohibited by any principle or linguistic constraint; these accidental gaps can also be called "false zeroes". Meanwhile a *grammatical* gap is systematic and is related to phonotactic restrictions. (Many times it is tempting to assign a grammatical interpretation to otherwise lexical gaps.) In addition, some sequences may be considered ungrammatical despite actually occurring in a few instances; these "false positives" have linguistic or grammatical reasons to not exist, but the restrictions are waived for certain loanwords or another restricted subset of the lexicon. Further work using frequency of suffixes (morphotactics) could be instrumental in determining the nature and causes of lexical gaps. For example, as stated earlier, my hypothesis is that the ubiquity of -t and -k suffixes in Hungarian as accusative and plural markers create a dispreference for nouns to end in these segments.

I made the distinction concerning segment gaps in the previous paragraph because the present study (and previous ones after which it was modeled) does not take into account non-occurring biphone sequences. That is, the mean absolute r_φ measure cited above is the average r_φ *only over occurring biphones*; the statistic ignores biphones with zero frequency. Hence an absolute constraint against co-occurrence – the strongest form of the otherwise gradient constraints considered here – appears to be set aside. Non-occurring sequences do not contribute to the measure of strength of association or repulsion between adjacent segments.

This absence would seem a serious drawback that applies to the present and previous statistical studies of syllable structure. For example, as Hungarian has 14 vowels and 24 consonants, there are 336 possible CV and VC sequences. As reported in (4.15), there are 235 CV and 219 VC sequences which were actually attested, which leaves around one-third of all possible sequences as non-occurring lexical gaps. This is a meaningful quanity, and the calculation of r_φ needs to account for the lexical gaps. This is done in the next section.

4.7.1 Accounting for non-occurring biphones

The question arises as to why there exist more attracting relationships than repelling ones between phonemes. One reason may be because systematic gaps are not included in the distribution of r_φ values. To find whether this caused skew in the r_φ distribution, I decided to recalculate the r_φ for CVC monosyllables in Hungarian, this time including biphone pairs with zero frequency. The table below in (4.22) shows results of the calculations. It can be seen that there is only a small effect of including zero-probability biphones.

(4.22) r_φ in Hungarian CVC monosyllables with and without zero probability biphones

	mean absolute r_φ	positive component of mean absolute r_φ	negative component of mean absolute r_φ	mean absolute r_φ with zeros added	positive component of mean absolute r_φ with zeros added	negative component of mean absolute r_φ with zeros added
CV	0.032	0.038	-0.018	0.030	0.038	-0.023
VC	0.034	0.042	-0.020	0.030	0.041	-0.022

All of the newly added, zero-probability biphone sequences had negative r_φ values. This indicates that the added biphones had negative correlations with one another, which is another way of stating that they were unlikely to appear together. In general, low-frequency biphones would tend to have negative r_φ values while high-frequency biphones have positive r_φ values; however, this is not a monotonic relationship because other factors, such as the underlying uniphone probabilities, will affect the r_φ calculation also.

Nevertheless, the relationship observed previously seemed to hold – the mean of the negative component is generally smaller than the mean of the positive component (in absolute value). It was necessary to retry the experiment by adding zero-probability biphones into the r_φ distributions in order to better understand how these scores are calculated and to ensure that the original deisgn was not flawed. However, it did not appear that including or omitting the non-occurring phoneme sequences has a significant impact on the relative values of mean absolute r_φ.

4.8 The unique syllable structures of Hungarian, Korean, and English

Finally, I return to the main finding disclosed in this chapter, which is the result that Hungarian biphones do not show strong evidence suggestive of either body or rhyme

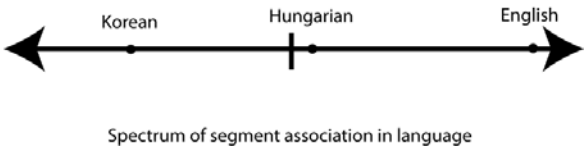
syllable structure. This is not to say that there are no internal associations in the Hungarian syllable – that is clearly not the case, as the biphones listed in Appendix F and Appendix G are examples of biphones with strong attracting associations; strong repelling relationships also exist between certain phone pairs. The finding is simply that these associations are not particularly biased towards appearing in either the syllable body or rhyme.

Lee and Goldrick's (2008) account of syllable structure is that it is emergent based on the statistical distribution of segments in the word. In the case of Hungarian, I conclude that this language presents a case in which no hierarchical structure is emergent – Hungarian syllable structure is best described as flat with no internal branching, if one were forced to pick one of the structures described in Section 4.1 of this chapter.

In a typology of language syllable structures, Hungarian could be described as falling in the middle of a body-rhyme continuum, depicted below in (4.23). At the center of this continuum one finds languages with no particular tendencies to be categorized as either predominantly body or rhyme; to the left, predominantly body-coda languages; to the right, predominantly onset-rhyme languages.

The data points along this continuum can be found by taking the logarithm of the ratio of mean absolute r_{φ} values of the rhyme to the body. This calculation is fairly straightforward, but the values used to do this calculation are presented in (4.24) for clarity.

(4.23) A body-rhyme typology of the sub-syllabic structure of languages



The data points plotted in (4.23) are the ln(r/b) scores from the last column in the table in (4.24). I created these measures to give a visual representation to the intuition that many phonologists have as to the spectrum of possible sub-syllabic structures that may exist across languages. In (4.24), the values for *r* (rhyme) and *b* (body) are taken directly from the r_{φ} scores listed in the table in (4.15) for Korean, Hungarian, and English, and then the r/b and ln(r/b) values are derived from the r_{φ} scores.

The logarithm function is used because it maps numbers between 0 and 1 onto the set of negative numbers while mapping numbers greater than 1 onto the set of positive real numbers. Hence predominantly body-oriented languages have negative ln(r/b) values, while predominantly rhyme-oriented languages obtain positive ln(r/b) scores.[39] Again, the ln(r/b) scores in the final column in (4.24) were plotted to obtain the continuum in (4.23).

[39] For strict CV languages or languages with no codas, the value of b is equal to 0 and hence the ln(r/b) score is undefined in such cases. In such cases the rhyme node would be essentially meaningless. Changing the measure to ln (b/r) here does nothing to alleviate the issue of division by zero, as the real problem is that strict CV languages have no meaningful rhyme (or at least VC relationships within the rhyme as assumed in this chapter).

(4.24) Data used to create the body-rhyme continuum

Language	Body – b (r_φ for CV)	Rhyme – r (r_φ for VC)	r/b	ln(r/b)
Korean	0.050	0.039	0.78	-0.248
Hungarian	0.032	0.034	1.06	0.061
English	0.022	0.034	1.55	0.435

4.8.1 Refining the body-rhyme continuum

There is still an element lacking in the descriptive approach proposed above to categorize languages as either body or rhyme languages. Interesting qualitative data or linguistic generalizations could fail to be recognized. For instance, by averaging across all syllables of given type in Hungarian, I may have failed to notice significant subpatterns of association that tend strongly towards either body or rhyme structure and depend on particular vowels or consonants. This issue is representative of a more general problem with averaging data and why it is often necessary to give more descriptive statistics than an average – such as median, quartiles, or standard deviation – in order to provide a more detailed illustration of data distribution.

One type of linguistic generalization I may have failed to notice is similar to a situation for English concerning lax vowels in CVC syllables. In a CVC syllable where the vowel is lax, there is reason to believe that the syllable structure is strongly onset-rhyme (/C/ /V C/) instead of body-coda (/C V/ /C/) in English (cf. e.g. Kapatsinski, 2007). While English is in general onset-rhyme, in this subset of cases the relationship becomes absolute, and this could be called a *strongly* onset-rhyme sublexicon. Syllables containing lax vowels in English must be followed by a consonant but are optionally preceded by a consonant. This evidence provides a natural reason to support the rhyme structure for this subset of the English lexicon. While Hungarian does not have a tense-

lax distinction in its vowel system, there is a short-long vowel length distinction for all vowels. Generalizations across other natural classes are also possible.

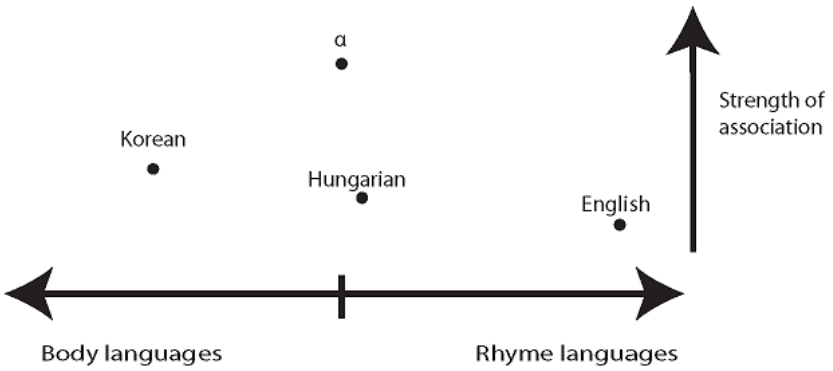
For the reasons stated above, I propose augmenting the body-rhyme continuum in (4.23) to add a second descriptive dimension. Let us denote as *S* the *mean strength of association* across CV and VC subsequences. This measure *S* is designed to indicate the tendency of a language to have strong associative relationships between adjacent phonemes, and it is defined as follows:

(4.25) $S = (r + b) / 2$

Without a measure such as *S*, one could mistakenly assume that Hungarian, for example, has no associations between its phones. To the contrary, it has many prominent associations, but these associations do not appear exclusively in either the rhyme or the coda.

I use *S* to create a vertical dimension for the typology of sub-syllabic structure, as shown in (4.26). Strength of association values for the three languages under question are $S_K = 0.045$, $S_E = 0.028$, $S_H = 0.033$. Languages with larger *S* values appear higher in the illustration below.

(4.26)



It is not possible to have $S < 0$ because it is the average of two correlation scores that I took to be positive. A score of $S = 0$ is possible in theory, but in this case the language would be plotted at the origin above – it is not possible to favor body or rhyme structure without having non-zero association strength scores.

I devised the mean association strength indicator in order to describe cases such as the hypothetical language α, also plotted in (4.26). Language α is a language which shows no overall evidence of a preference for body or rhyme structure. However, subsets of this language show strong tendencies for both rhyme- and body-like patterns. Language α – along with other languages with high mean association strength measures – is a language in which further investigation of the nature of the linguistic associations would likely yield rich qualitative linguistic information. Put another way, while language α would not be typically defined as a body or rhyme language, it is not the case that the phonemes have random distributions within the syllable.

Ultimately, speakers may possess sensitivity to individual CV or VC statistics of co-occurrence that are more useful than the generalization of whether the language is broadly classified as a rhyme or body language. Indeed, this is borne out in the data of

Lee (2006), who asked speakers to memorize nonce or atypical syllables and then later recall them. Focusing on errors in recall, when asked to repeat from memory a CVC sequence in which the nucleus co-occurs more often with the onset, English-speaking study participants tended to produce body-coda recombinations of previously presented bodies and codas – the opposite of the general pattern expected for English participants in which body-coda sequences are more easily recalled. By the same token, if Korean speakers are presented with syllables with frequent nucleus-coda Korean sequences, speakers tended to produce onset-rhyme recombinations, despite Korean being a body-coda language. This shows that the strength of association between individual phones wins out over the broad tendencies of association for the language. A language can have rhyme-like patterns and still be overall a body-structure language (and the opposite is also true).

In summary, I have shown that Hungarian sub-syllabic structure cannot be characterized as either English-like or Korean-like. This result may mean that one may make predictions that Hungarian speakers will perform differently from English and Korean speakers on a variety of psycholinguistic tasks.

4.9 Directions going forward

A question that remains a subject for future research is to determine how Hungarian speakers perform in psycholinguistic tasks requested of them along the lines of previous research (such as the repetition, word-blending, or word-breaking tasks described earlier in Section 4.1.3.2). For example, when asked to break a CVC sequence into two parts, would a Hungarian informant not prefer creating syllable bodies or rhymes, or would the

informant instead prefer to keep high probability biphones together as a unit? As Hungarian does not appear to be strongly rhyme- or body-based, my hypothesis is that speakers would rely on biphone probabilities when performing this task.

What is apparent to me is that trying to distinguish between body structure and rhyme structure for Hungarian may not be such an insightful endeavor. Writing about the structure of syllables in all languages, Vennemann arrived at a similar conclusion regarding a similar debate on internal syllable structure twenty years ago:

> Syllables [...] actually have all the sorts of structure that have been proposed; more precisely, that they can assume any one of those structures depending on the syllable-related phenomenon under study. Regularities of accent, rime, and meter are typically sensitive only to that part of the syllable which consists of the peak and what follows; they only look at the 'rhyme projection' of a text.
>
> I think it would be a methodological error to insist that, even despite all the conflicting evidence cited, the syllable must have [...] one or the other of the structures discussed. In linguistics, this error has caused a lot of unfruitful discussion. Perhaps the best known is the controversy over whether affricates are mono- or diphonemic. The answer is they may be either or both, depending on the regularity under study. (Vennemann, 1988: 269-70)

The "false dichotomy" of body versus rhyme structure distracts our attention from other issues. Being forced to adopt one structure at the expense of another may obscure other interesting patterns of co-occurrence. Indeed, in order to retain all insightful notions of sub-syllabic structure – which are at times contradictory – a model needs to be found which would group constituents based on relevant phonological properties. This seems especially apparent when investigating Hungarian, which does not seem to strongly prefer one syllable hierarchy over another.

5 Concluding remarks

This dissertation advocated a frequency-based approach to studying the phonotactics of language. The body of work may serve not only as a resource for those interested in new Hungarian insights, but it also can be a basis for similar explorations of lexical and phonotactic explorations of other languages. In this final chapter, I wish to review some of the more significant findings of the dissertation. I also discuss research topics that remain unaddressed and proposals for extending the questions raised by the present work.

5.1 Results of the dissertation

One concrete output of the current research is the Hungarian pronunciation dictionary. After making presentations about the pronunciation dictionary at linguistic conferences and distributing it online through a website, other researchers were motivated to download it for their comparative research. Later in this chapter, examples of how to further enhance the dictionary are provided.

The other major output of the dissertation was the work on syllable structure presented in Chapter 4. I showed that Hungarian can neither be classified as a strong rhyme language (such as English) nor as a strong syllable body language (such as Korean). This finding could be interpreted as problematic for theorists who propose one type of universal sub-syllabic structure for all languages. At the same time, such a finding may spur other investigations of sub-syllabic structure in other languages to confirm that other such rhyme-neutral languages exist.

5.2 Future directions

5.2.1 Phonological lexicon

In Section 3.6, I proposed a number of possible improvements to the Hungarian pronunciation dictionary. The most pressing and substantial improvement, in my view, is to add morphological analyses for each word. This requires a morphological parser to identify stems and suffixes for each word, and the resulting analyzed word will contain morpheme boundary information. Fortunately a resource already exists that could be deployed: morphdb.hu (Trón et al., 2006) is a Hungarian lexical database and morphological grammar. Morphdb.hu itself was created by merging three lexical databases, and it is capable of handling inflective and productive morphological derivation. Knowledge of morpheme boundaries would allow the implementation of specific improvements for pronunciation rules that only apply to stems such as, for example, the constraint on allowable syllable weight (*VVCC). As noted in Chapter 2, *VVCC does not apply to derived (suffixed) forms. As morphdb.hu is also a lexical resource, new entries for the pronunciation dictionary can be harvested from this list, assigned pronunciations, and added to the dictionary to enhance its scope.

Another method for checking the correctness of pronunciations has also now presented itself. An online database of speech duration for Hungarian words (Olaszy, 2003, Olaszy and Kálmán, 2005) gives pronunciations for over one million words.[40] One could compare pronunciations for words in my pronunciation dictionary against this resource, and the cases in which the outputs do not agree could be flagged for closer examination by hand.

---

[40] Available at http://fonetika.nytud.hu/hitint.

In Section 3.4.10, I provided a partial list of words with exceptional pronunciations. The online speech duration database from Olaszy and colleagues correctly handles such exceptional words as *lesz* 'will be', *nagy* 'large', and *egyet* 'large.ACC', but it appears to have incorrect pronunciations listed for words such as *csat* 'battle', *új* 'new', *kulturális* 'cultural', *csehek* 'Czech.PL', and others. By referring to this additional pronunciation resource for Hungarian, one can automatically identify discrepancies and flag potential errors. These potential errors could be later adjudicated by hand.

Aside from making the pronunciation dictionary available on my personal website, there are possibilities for wider distribution. For example, the natural language toolkit (Loper and Bird, 2002, Bird, Klein and Loper, 2009) is an educational resource to teach natural language processing. NLTK includes widely available versions of corpora such as the Brown corpus; distribution of a phonological lexicon with accompanying exercises would establish and promote this type of research in classrooms. Creating a web-based interface to search the dictionary would also be useful for less technically-savvy users.

Finally, there are many more possibilities for extending the pronunciation dictionary. One is to encode the data according to format requirements of other popular digital lexicons such as CELEX2. In this way, people familiar with the UNIX tools required for searching CELEX databases could naturally transition from only studying English, German, and Dutch to also studying Hungarian, thereby expanding their research focus to historically unrelated languages.

5.2.2 Syllable structure of whole words

There are further directions to pursue with respect to researching sub-syllabic structure. Chapter 4 proposed a typology of sub-syllabic structure in which each language falls along a body-rhyme continuum. In order to refine the proposed categorization and better understand its implications, plotting additional language data points is necessary. This is especially pertinent for languages believed to have rhyme- or body-based syllable organization based on independent evidence (such as evidence from phonological processes, insights from slips of the tongue, or patterning evidenced by language games).

Chapter 4 included a discussion on the pitfalls encountered when trying to scale up from the monosyllabic word to full-length, polysyllabic representations. To account for differences in resulting r_{φ} measures, my conclusion was that monosyllabic lexicons may possess more extreme phonotactics as compared to the language at large. Aside from only studying bigram connectivity of segments, examining the full lexicon crucially requires having correct syllable parses. It would be useful to compare separately the phonotactics of initial, medial, and final syllables, as their characteristics are likely not to be uniform. For certain languages, stress also influences syllabification, and further investigation of the relationship between stress and phonotactics could prove useful in this framework.

Appendix A. Transcription systems and symbol equivalents

ORTHOGRAPHY	IPA	OGOB	SAMPA	PRÓSZÉKY
a	ɑ	a	O	a
á	aː	A	a:	a1
b	b	b	b	b
c	ts	c	ts	c
cs	ʧ	C	tS	cs
d	d	d	d	d
dzs	ʤ	D	dZ	dzs
e	ɛ	e	E	E
é	eː	E	e:	e1
f	f	f	f	f
g	g	g	g	g
gy	ɟ	G	d'	gy
h	h	h	h	h
i, y	i	i	i	i
í	iː	I	i:	i1
j, ly	j	j	j	j
k	k	k	k	k
l	l	l	l	l
m	m	m	m	m
n	n	n	n	ny
ny	ɲ	N	J	ny
o	o	o	o	o
ó	oː	O	o:	o1
ö	ø	w	2	o2
ő	øː	W	2:	o3
p	p	p	p	p
r	r	r	r	r
s	ʃ	s	S	s
sz	s	S	s	sz
t	t	t	t	ty
ty	c	T	t'	ty
u	u	u	u	u

ú	uː	U	u:	u1
ü	y	y	y	u2
ű	yː	Y	y:	u3
v, w	v	v	v	v
z	z	z	z	zs
zs	ʒ	Z	Z	zs

Appendix B. A screenshot of the first thirty entries in the pronunciation dictionary

ORTH	PRON	SAMPA	FREQ	VOWELS	CONS	CV_STRUCT	POS	#SYLL
abba	abba	O b: O	28965	aa	bb	VCCV	Pre	2
ablak	ablak	O b l O k	18756	aa	blk	VCCVC	N	2
ablakos	ablakos	O b l O k o S	607	aao	blks	VCCVCVC	A	3
ablaktörlő	ablaktwrlW	O b l O k t 2 r l 2:	589	aawW	blktrl	VCCVCCVCCV	N	4
abnormális	abnormAlis	O b n o r m a: l i S	797	aoAi	bnrmls	VCCVCCVCVC	A	4
abortusz	abortuS	O b o r t u s	3024	aou	brtS	VCVCCVC	N	3
abroncs	abronC	O b r o n tS	816	ao	brnC	VCCVCC	N	2
abszolutizmus	apSolutizmus	O p s o l u t i z m u S	897	aouiu	pSltzms	VCCVCVCVCCVC	N	5
abszolút	apSolUt	O p s o l u: t	25184	aoU	pSlt	VCCVCVC	N	3
abszolúte	apSolUte	O p s o l u: t E	829	aoUe	pSlt	VCCVCVCV	Adv	4
absztrakció	abzdrakcijO	O b z d r O k ts i j o:	778	aaiO	bzdrkcj	VCCCCVCCVCV	N	4
absztrakt	abzdrakt	O b z d r O k t	3264	aa	bzdrkt	VCCCCVCC	A	2
abszurd	apSurd	O p s u r d	3967	au	pSrd	VCCVCC	A	2
acél	acEl	O ts e: l	5100	aE	cl	VCVC	N	2
acéllemez	acEllemez	O ts e: l: E m E z	614	aEee	cllmz	VCVCCVCVC	N	4
ad	ad	O d	121452	a	d	VC	V	1
adag	adag	O d O g	6838	aa	dg	VCVC	N	2
adakozó	adakozO	O d O k o z o:	605	aaoO	dkz	VCVCVCV	MIF	4
adalék	adalEk	O d O l e: k	2030	aaE	dlk	VCVCVC	N	3
adandó	adandO	O d O n d o:	2372	aaO	dnd	VCVCCV	A	3
adat	adat	O d O t	24903	aa	dt	VCVC	V	2
adatfeldolgozó	adatfeldolgozO	O d O t f E l d o l g o z o:	940	aaeooO	dtfldlgz	VCVCCVCCVCCVCV	MIF	6
adatgyûjtés	adadGYjtEs	O d O d d' y: j t e: S	2745	aaYE	ddGjts	VCVCCVCCVC	N	4
adathalmaz	adathalmaz	O d O t h O l m O z	471	aaaa	dthlmz	VCVCCVCCVC	N	4
adatik	adatik	O d O t i k	2286	aai	dtk	VCVCVC	V	3
adatszolgáltatás	adaccolgAltatAs	O d O ts: o l g a: l t O t a: S	4006	aaoAaA	dcclgltts	VCVCCVCCVCCVCVC	N	6
adattár	adattAr	O d O t: a: r	1152	aaA	dttr	VCVCCVC	N	3
addigi	addigi	O d: i g i	6748	aii	ddg	VCCVCV	A	3
adekvát	adekvAt	O d E k v a: t	1985	aeA	dkvt	VCVCCVC	A	3
adjunktus	aGGunktus	O d': u n k t u S	1445	auu	GGnkts	VCCVCCCVC	N	3
adminisztratív	adminizdratIv	O d m i n i z d r O t i: v	9552	aiiaI	dmnzdrtv	VCCVCVCCCVCVC	A	5

Appendix C. Initial pronunciation dictionary error-checking list

Orthography	OGOB transcription	SAMPA transcription
néhányszor	nEhAnSor	n e: h a: n s o r
hasonlatos	hasonlatos	h O S o n l O t o S
kedd	kedd	k E d:
negyed	neGed	n E d' E d
örökre	wrwgre	2 r 2 g r E
részesít	rESesIt	r e: s E S i: t
arisztokratikus	ariStogratikus	O r i s t o g r O t i k u S
hangzás	hangzAs	h O n g z a: S
állatkert	Allatkert	a: l: O t k E r t
virul	vIrul	v i: r u l
igazán	IgazAn	i: g O z a: n
délután	dElutAn	d e: l u t a: n
bukik	bUkik	b u: k i k
kompetencia	kompetencija	k o m p E t E n ts i j O
kapu	kapu	k O p u
félig	fElig	f e: l i g
futó	fUtO	f u: t o:
rettenetes	rettenetes	r E t: E n E t E S
paraszti	paraSti	p O r O s t i
emelet	emelet	E m E l E t
lovaglás	lovaglAs	l o v O g l a: S
marxizmus	markSizmus	m O r k s i z m u S
igazságtalanság	IgaZAktalansAg	i: g O Z a: k t O l O n S a: g
téved	tEved	t e: v E d
utólagos	UtOlagos	u: t o: l O g o S
fejedelmi	fejedelmi	f E j E d E l m i
jog	jog	j o g
kapkodás	kapkodAs	k O p k o d a: S
személyszállító	SemEjSAllItO	s E m e: j s a: l: i: t o:
indulási	indulAsi	i n d u l a: S i
étkező	EtkezW	e: t k E z 2:
hangfal	hankfal	h O n k f O l
szomorú	SomorU	s o m o r u:
leír	lejIr	l E j i: r
tápláló	tAblAlO	t a: b l a: l o:
jogosítvány	jogosItvAN	j o g o S i: t v a: J
gyomor	Gomor	d' o m o r
magántőke	magAntWke	m O g a: n t 2: k E
presztízs	breStIZ	b r E s t i: Z
foglalkoztatott	foglalkoStatott	f o g l O l k o s t O t o t:
áramvonalas	Aramvonalas	a: r O m v o n O l O S
gyűjtő	GYjtW	d' y: j t 2:
mélyhűtött	mEjhYtwtt	m e: j h y: t 2 t:
szérum	SErum	s e: r u m
hogy	hoG	h o d'

lojalitás	lojalitAs	l o j O l i t a: S
menetjegy	meneTTeG	m E n E t': E d'
recepció	recepcijO	r E ts E p ts i j o:
csokor	Cokor	tS o k o r
alperes	alperes	O l p E r E S
sivár	sIvAr	S i: v a: r
sajtószabadság	sajtOSabaCCAg	S O j t o: s O b O tS: a: g
teknős	tegnWs	t E g n 2: S
ritmusú	ridmusU	r i d m u S u:
értelmes	Ertelmes	e: r t E l m E S
boldogságos	boldoksAgos	b o l d o k S a: g o S
megdöbbentő	megdwbbentW	m E g d 2 b: E n t 2:
függ	fygg	f y g:
link	link	l i n k
mostan	mostan	m o S t O n
démon	dEmon	d e: m o n
mogorva	mogorva	m o g o r v O
sodort	sodort	S o d o r t
kimondhatatlan	kImonthatadlan	k i: m o n t h O t O d l O n
őrzés	WrzEs	2: r z e: S
szójáték	SOjAtEk	s o: j a: t e: k
irt	irt	i r t
kanyar	kaNar	k O J O r
bagázs	bagAZ	b O g a: Z
milliomos	millijomos	m i l: i j o m o S
termelés	termelEs	t E r m E l e: S
felkészítés	felkESItEs	f E l k e: s i: t e: S
szelíd	SelId	s E l i: d
örömest	wrwmest	2 r 2 m E S t
írástudó	IrAstudO	i: r a: S t u d o:
remete	remete	r E m E t E
egyházszakadás	eThASSakadAs	E t' h a: s: O k O d a: S
bűzlik	bYzlik	b y: z l i k
metaforikus	metaforikus	m E t O f o r i k u S
alapít	alapIt	O l O p i: t
örökség	wrwksEg	2 r 2 k S e: g
akkorára	akkorAra	O k: o r a: r O
hátrafelé	hAdrafelE	h a: d r O f E l e:
egyszer	eccer	E ts: E r
óvatos	Ovatos	o: v O t o S
törvénycikk	twrvEncikk	t 2 r v e: n ts i k:
klubtag	gluptag	g l u p t O g
szivacs	SIvaC	s i: v O tS
tiltakozó	tiltakozO	t i l t O k o z o:
kormányfő	kormANfW	k o r m a: J f 2:
agilis	agilis	O g i l i S
visszamenőleg	viSSamenWleg	v i s: O m E n 2: l E g
hírlap	hIrlap	h i: r l O p

folklór	folglOr	f o l g l o: r
hu	hu	h u
irányítás	IrANItAs	i: r a: J i: t a: S
főként	fWkEnt	f 2: k e: n t
konföderáció	komfwderAcijO	k o m f 2 d E r a: ts i j o:
tök	twk	t 2 k
fürdik	fyrdik	f y r d i k
igazság	IgaZAg	i: g O Z a: g
billentyűzet	billeNTYzet	b i l: E J t' y: z E t
írás	IrAs	i: r a: S
kaszt	kaSt	k O s t
szerencse	SerenCe	s E r E n tS E
esedékes	esedEkes	E S E d e: k E S
tüzes	tYzes	t y: z E S
megépül	megEpyl	m E g e: p y l
csillagászat	CillagASat	tS i l: O g a: s O t
króm	grOm	g r o: m
átkerül	Atkeryl	a: t k E r y l
tűrhetetlen	tYrhetedlen	t y: r h E t E d l E n
fehér	fehEr	f E h e: r
apai	apaji	O p O j i
elvi	elvi	E l v i
sarki	sarki	S O r k i
ovális	ovAlis	o v a: l i S
vakáció	vakAcijO	v O k a: ts i j o:
meghatározatlan	mekhatArozadlan	m E k h O t a: r o z O d l O n
képviselet	kEpviselet	k e: p v i S E l E t
hegyoldal	heGoldal	h E d' o l d O l
fül	fyl	f y l
ered	ered	E r E d
mérő	mErW	m e: r 2:
játszi	jAcci	j a: ts: i
rohamos	rohamos	r o h O m o S
káposzta	kApoSta	k a: p o s t O
kiemelés	kIjemelEs	k i: j E m E l e: S
felkérés	felkErEs	f E l k e: r e: S
foglalt	foglalt	f o g l O l t
kutatómunka	kUtatOmunka	k u: t O t o: m u n k O
szerves	Serves	s E r v E S
formai	formaji	f o r m O j i
leküzdhetetlen	lekySthetedlen	l E k y s t h E t E d l E n
kényelmetlen	kENelmedlen	k e: J E l m E d l E n
tökös	twkws	t 2 k 2 S
igyekvő	IGekvW	i: d' E k v 2:
súgó	sUgO	S u: g o:
lemerül	lemeryl	l E m E r y l
mérhetetlen	mErhetedlen	m e: r h E t E d l E n
természetfeletti	termESetfeletti	t E r m e: s E t f E l E t: i

közélet	kwzElet	k 2 z e: l E t
fellegi	fellegi	f E l: E g i
kifejlődött	kIfejlWdwtt	k i: f E j l 2: d 2 t:
közvetlenség	kwzvedlensEg	k 2 z v E d l E n S e: g
munkaadói	munkaadOji	m u n k O O d o: j i
csereszerződés	CereSerzWdEs	tS E r E s E r z 2: d e: S
tárgyalási	tArGalAsi	t a: r d' O l a: S i
gyomorszáj	GomorSAj	d' o m o r s a: j
lefordítható	lefordIthatO	l E f o r d i: t h O t o:

Appendix D. Followup pronunciation dictionary error-checking list

Orthography	OGOB transcription	SAMPA transcription
naivitás	najivitAs	n O j i v i t a: S
szürkeség	SyrkesEg	s y r k E S e: g
középkor	kwzEpkor	k 2 z e: p k o r
vámterület	vAmterylet	v a: m t E r y l E t
dotáció	dotAcijO	d o t a: ts i j o:
átrendezés	AtrendezEs	a: t r E n d E z e: S
lecsatol	leCatol	l E tS O t o l
irodaépület	IrodaEpylet	i: r o d O e: p y l E t
legyőzött	leGWzwtt	l E d' 2: z 2 t:
gyékény	GEkEN	d' e: k e: J
kertajtó	kertajtO	k E r t O j t o:
kitermelés	kItermelEs	k i: t E r m E l e: S
átszivárog	AccivArog	a: ts: i v a: r o g
sarkú	sarkU	S O r k u:
útelágazás	UtelAgazAs	u: t E l a: g O z a: S
vásárló	vAsArlO	v a: S a: r l o:
kapkodás	kapkodAs	k O p k o d a: S
kipécéz	kIpEcEz	k i: p e: ts e: z
étkező	EtkezW	e: t k E z 2:
közjogi	kwzjogi	k 2 z j o g i
alliteráció	alliterAcijO	O l: i t E r a: ts i j o:
csallóközi	CallOkwzi	tS O l: o: k 2 z i
vaskó	vaskO	v O S k o:
lekicsinyel	lekiCiNel	l E k i tS i J E l
peron	peron	p E r o n
kókuszdió	kOkuzdijO	k o: k u z d i j o:
körbenéz	kwrbenEz	k 2 r b E n e: z
delfin	delfin	d E l f i n
baracklekvár	baracklekvAr	b O r O ts k l E k v a: r
népszerű	nEpSerY	n e: p s E r y:
alibi	alibi	O l i b i
visszapillant	viSSapillant	v i s: O p i l: O n t
íves	Ives	i: v E S
csodál	CodAl	tS o d a: l
anyanyelvi	aNaNelvi	O J O J E l v i
kapitalista	kapitalista	k O p i t O l i S t O
helység	hejsEg	h E j S e: g
nívótlan	nIvOtlan	n i: v o: t l O n
évad	Evad	e: v O d
bozót	bozOt	b o z o: t
tisztán	tiStAn	t i s t a: n
feltétlen	feltEtlen	f E l t e: t l E n
cserkész	CerkES	tS E r k e: s
vágóállat	vAgOAllat	v a: g o: a: l: O t
szétver	SEtver	s e: t v E r

lopó	lopO	l o p o:
fejvadász	fejvadAS	f E j v O d a: s
alakos	alakos	O l O k o S
fecseg	feCeg	f E tS E g
kikapcsol	kIkapCol	k i: k O p tS o l
homoszexuális	homoSekSuAlis	h o m o s E k s u a: l i S
sírdogál	sIrdogAl	S i: r d o g a: l
szállásol	SAllAsol	s a: l: a: S o l
cigányzenekar	cIgANzenekar	ts i: g a: J z E n E k O r
magyarság	maGarsAg	m O d' O r S a: g
megfélemlít	mekfElemlIt	m E k f e: l E m l i: t
jövőre	jwvWre	j 2 v 2: r E
szállásadó	SAllAsadO	s a: l: a: S O d o:
leleményesség	lelemENessEg	l E l E m e: J E S: e: g
jelképi	jelkEpi	j E l k e: p i
zöldség	zwlCEg	z 2 l tS e: g
vagyontárgy	vaGontArG	v O d' o n t a: r d'
szétkerget	SEtkerget	s e: t k E r g E t
sirály	sIrAj	S i: r a: j
csúcspont	CUCpont	tS u: tS p o n t
direkt	dIrekt	d i: r E k t
proklamáció	proklamAcijO	p r o k l O m a: ts i j o:
világhírű	vIlAkhIrY	v i: l a: k h i: r y:
élettelen	Elettelen	e: l E t: E l E n
elcsúszás	elCUSAs	E l tS u: s a: S
filmipar	filmipar	f i l m i p O r
beleönt	belewnt	b E l E 2 n t
kulcs	kulC	k u l tS
mindenható	mindenhatO	m i n d E n h O t o:
kesztyűtartó	keSTYtartO	k E s t' y: t O r t o:
zenit	zenit	z E n i t
hangzik	hangzik	h O n g z i k
fonal	fonal	f o n O l
fonnyadt	foNNatt	f o J: O t:
sérelmez	sErelmez	S e: r E l m E z
említés	emlItEs	E m l i: t e: S
társasági	tArsasAgi	t a: r S O S a: g i
észlelő	ESlelW	e: s l E l 2:
csucsor	CUCor	tS u: tS o r
főnyeremény	fWNeremEN	f 2: J E r E m e: J
álmélkodik	AlmElkodik	a: l m e: l k o d i k
sokszorosít	sokSorosIt	S o k s o r o S i: t
kanyargó	kaNargO	k O J O r g o:
küzdelem	kyzdelem	k y z d E l E m
továbbindul	tovAbbindul	t o v a: b: i n d u l
üldözési	yldwzEsi	y l d 2 z e: S i
akasztófa	akaStOfa	O k O s t o: f O
vérszerződés	vErSerzWdEs	v e: r s E r z 2: d e: S

fegyverviselés	feGverviselEs	f E d' v E r v i S E l e: S
legyilkol	leGilkol	l E d' i l k o l
az	az	O z
jobbágyfelszabadítás	jobbATfelSabadItAs	j o b: a: t' f E l s O b O d i: t a: S
motyogás	moTogAs	m o t' o g a: S
védegylet	vEdeGlet	v e: d E d' l E t
platió	platO	p l O t o:
fényképész	fENkEpES	f e: J k e: p e: s
mozdít	mozdIt	m o z d i: t
konyhaajtó	koNhaajtO	k o J h O O j t o:
fiúcska	fIjUCka	f i: j u: tS k O
rendszertelen	rencertelen	r E n ts E r t E l E n
adóbehajtás	adObehajtAs	O d o: b E h O j t a: S
féreg	fEreg	f e: r E g
lehúz	lehUz	l E h u: z
furikázik	fUrikAzik	f u: r i k a: z i k
lám	lAm	l a: m
egészségügyi	egEssEgyGi	E g e: S: e: g y d' i
tuskó	tuskO	t u S k o:
pinceajtó	pinceajtO	p i n ts E O j t o:
elpirul	elpirul	E l p i r u l
párosodik	pArosodik	p a: r o S o d i k
lecsendesít	leCendesIt	l E tS E n d E S i: t
bányamérnök	bANamErnwk	b a: J O m e: r n 2 k
apróz	aprOz	O p r o: z
párkány	pArkAN	p a: r k a: J
elhárít	elhArIt	E l h a: r i: t
cirógatás	cIrOgatAs	ts i: r o: g O t a: S
hattyúdal	haTTUdal	h O t': u: d O l
trágyadomb	trAGadomb	t r a: d' O d o m b
hangoztatás	hangoStatAs	h O n g o s t O t a: S
visszanyer	viSSaNer	v i s: O J E r
meglóg	meglOg	m E g l o: g
beleesik	beleesik	b E l E E S i k
pattogás	pattogAs	p O t: o g a: S
iszap	ISap	i: s O p
ugrat	ugrat	u g r O t
rugó	rUgO	r u: g o:
életfilozófia	EletfilozOfija	e: l E t f i l o z o: f i j O
vetkőztet	vetkWStet	v E t k 2: s t E t
padlós	padlOs	p O d l o: S
igazságügy	IgassAgyG	i: g O S: a: g y d'
elküld	elkyld	E l k y l d
felvonó	felvonO	f E l v o n o:
tűlevél	tYlevEl	t y: l E v e: l
amiképpen	amikEppen	O m i k e: p: E n
bog	bog	b o g
haszonélvező	haSonElvezW	h O s o n e: l v E z 2:

diákvezér	dIjAkvezEr	d i: j a: k v E z e: r
megfertőz	mekfertWz	m E k f E r t 2: z
kocsikísérő	koCikIsErW	k o tS i k i: S e: r 2:
kiút	kIjUt	k i: j u: t
határtalan	hatArtalan	h O t a: r t O l O n
precedens	precedens	p r E ts E d E n S
felgyújt	felGUjt	f E l d' u: j t
számtani	SAmtani	s a: m t O n i
nyomorú	NomorU	J o m o r u:

Appendix E. Distribution of consonants within onset and coda for English (data from Kessler and Treiman, 1997)

Phone	Onset	Coda	Chi^2	Theta
j	30	0	30	1
ʍ	24	0	24	1.009
w	82	0	82	1
ŋ	0	46	46	1
h	105	0	105	1
ð	1	14	11.27	0.867
ʒ	1	5	--	0.667
z	13	58	28.52	0.634
b	154	62	39.19	0.426
θ	17	39	8.64	0.393
n	99	207	38.12	0.353
ǰ	74	41	9.47	0.287
t	119	204	22.37	0.263
l	135	230	24.73	0.26
ʃ	65	44	4.05	0.193
f	92	68	3.6	0.15
r	163	124	5.3	1.36
g	88	67	2.85	0.135
k	142	182	4.94	0.123
v	45	54	0.82	0.091
p	128	112	1.07	0.067
d	126	142	0.96	0.06
m	116	127	0.5	0.045
s	126	116	0.41	0.041
č	56	59	0.08	0.026

Appendix F. Most frequent CV and VC sequences in CVC words

Transcribed in OGOB – capital letters indicate long vowels

Top 13 most frequently occurring CV sequences in CVC words

CV sequence	Type frequency	Token frequency
kE	11	428095
vE	9	152500
SA	8	150915
mE	8	423642
tA	7	30548
lA	7	193615
va	7	1814406
ha	7	104928
hA	7	89186
rE	7	183130
Se	7	66962
kw	7	94855
lE	7	86363
vA	7	98172

Top 12 most frequently occurring codas in CVC words

VC sequence	Type frequency	Token frequency
Ar	13	722413
Er	8	165512
aj	8	56033
El	8	222792
Aj	8	33929
el	8	199073
Az	7	82032
Ur	7	5019
Et	7	341908
ak	7	503202
Ep	7	185755
Eg	7	572851

Appendix G. Most frequent CV and VC sequences as strings in all words
Transcribed in OGOB – capital letters indicate long vowels

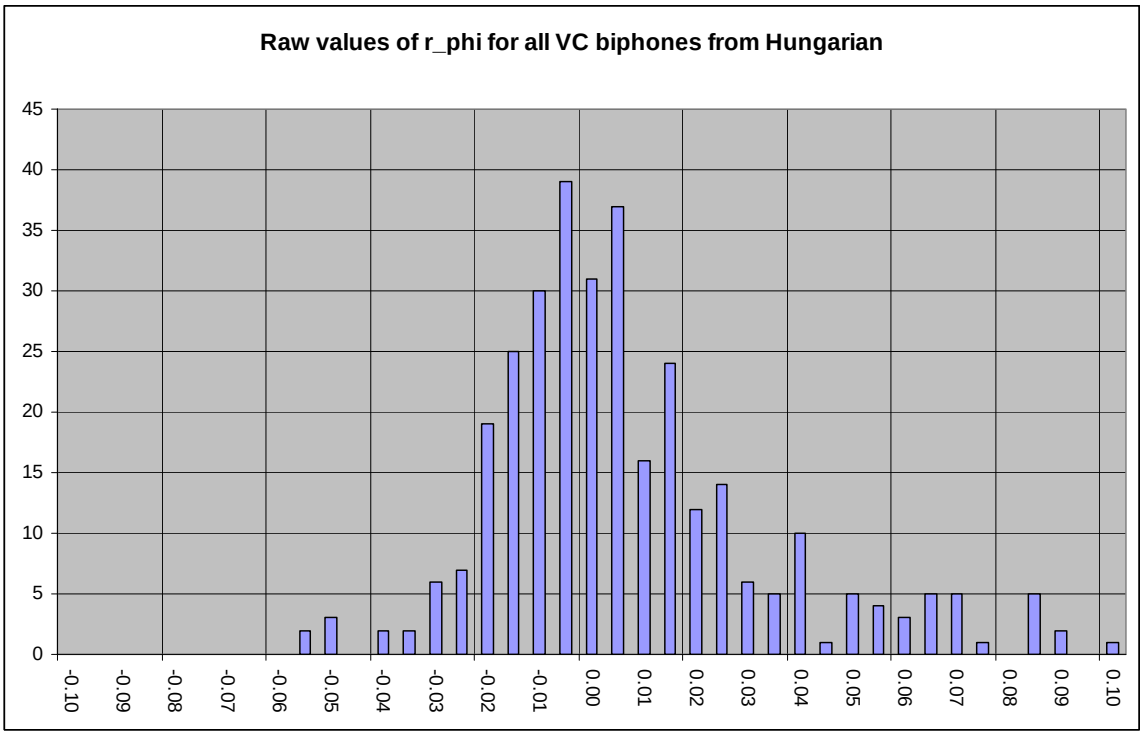
Top 15 most frequent CV sequences across all words

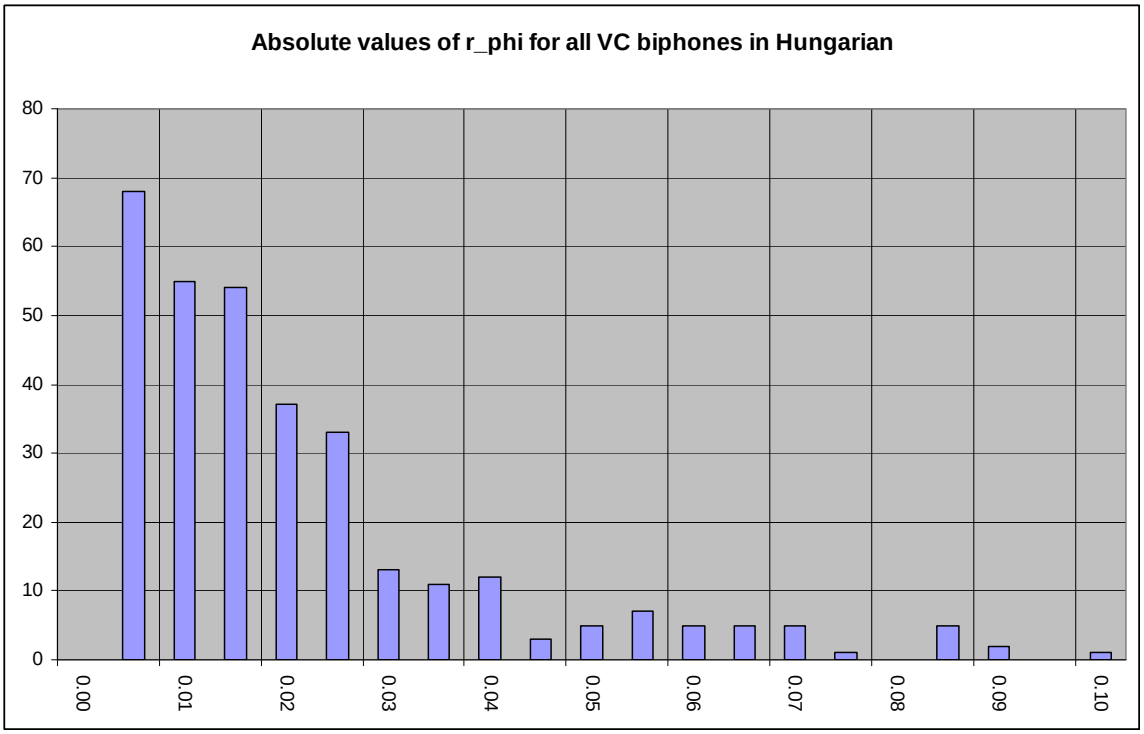
CV sequence	Type frequency	Token frequency
me	2604	5098900
ke	2210	4423779
fe	1819	2621885
be	1625	1773062
te	1556	2985189
ka	1555	1694416
ha	1449	2901338
ma	1399	3287880
ko	1248	1499730
Se	1205	3102570
va	1183	4073641
kI	1181	1271074
le	1176	2348505
ta	1120	2425833
kE	1108	2662121

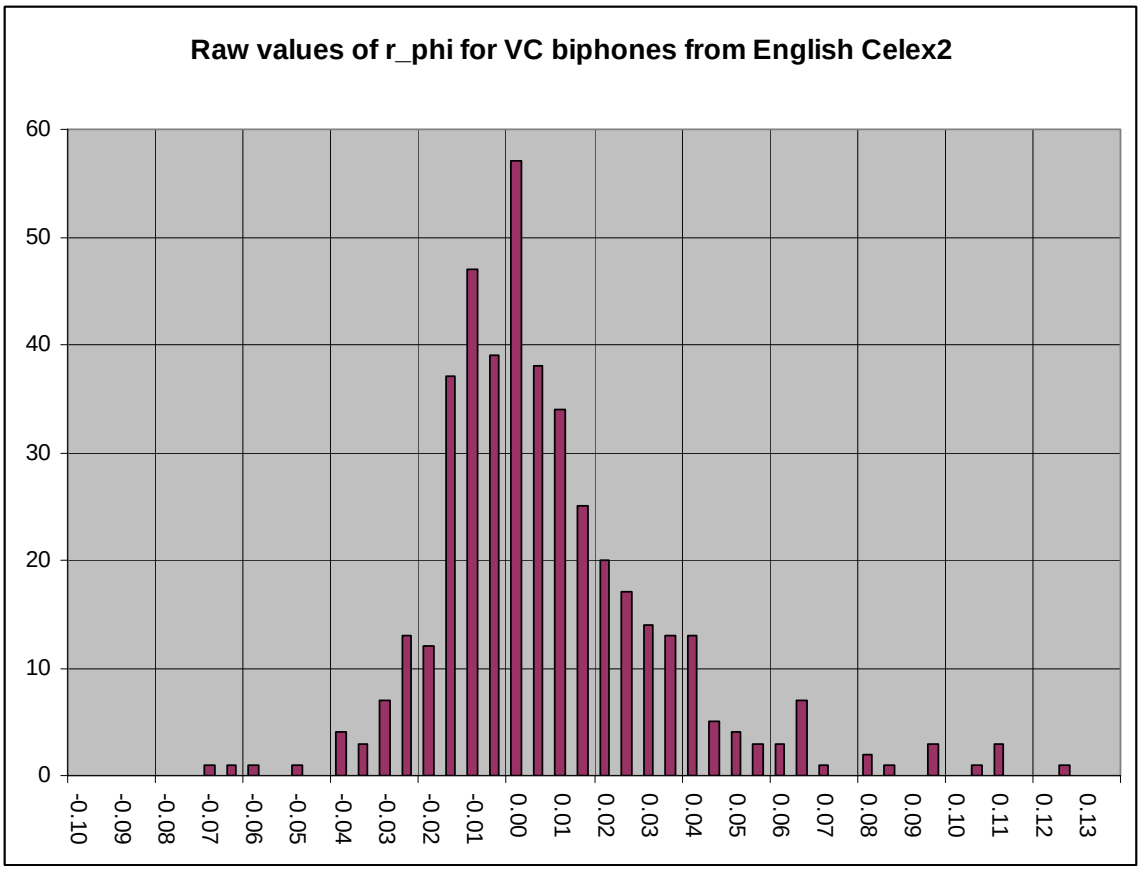
Top 15 most frequent VC sequences across all words

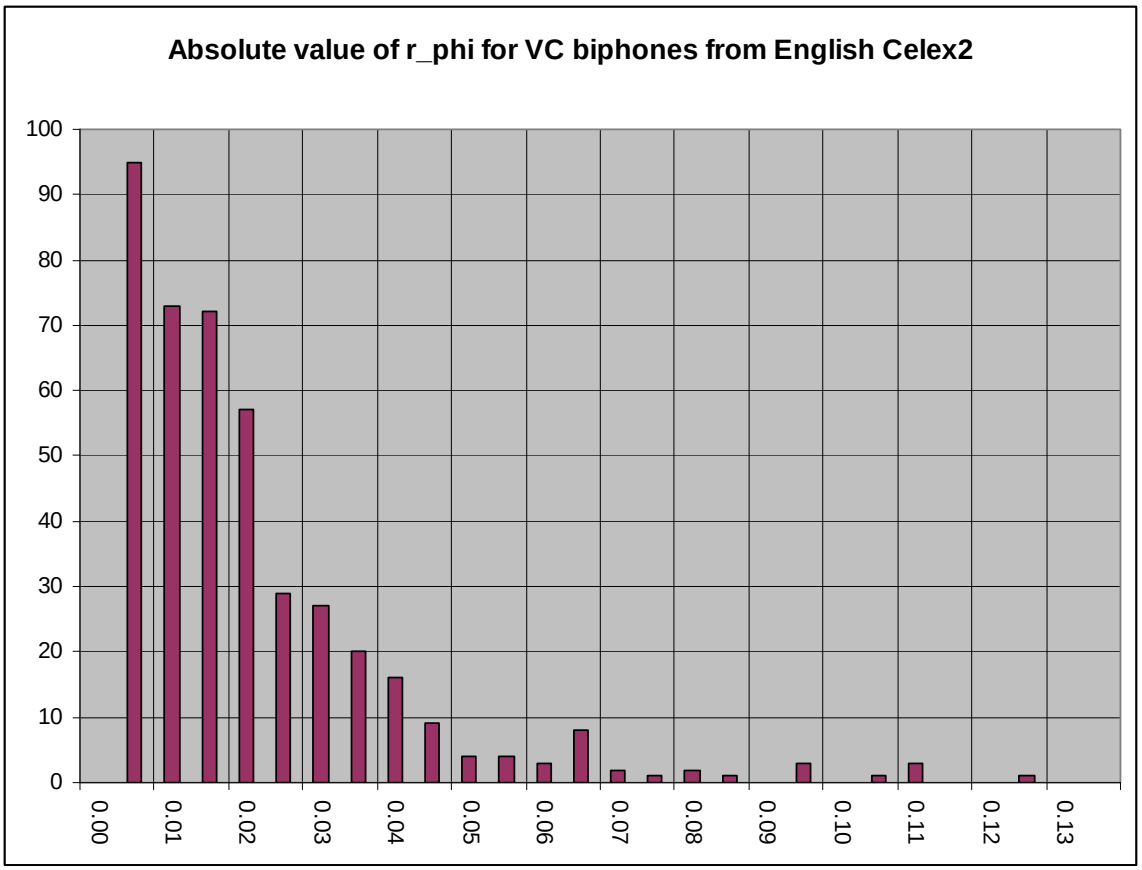
VC sequence	Type frequency	Token frequency
el	4169	8006446
er	2692	5199034
et	2499	4843211
As	2496	3313395
at	2380	4038511
al	2213	4271878
en	2211	4149651
or	1854	2724009
Al	1763	2904183
ik	1744	2505815
Ar	1712	3287903
ol	1596	2284911
Es	1555	5150065
ar	1521	1937264
an	1518	3710459

Appendix H. Bar charts of r_{φ} values of VC biphones from Hungarian and English. These are provided for comparison with CV bar charts on pages 130-132.






Raw values of r_phi for VC biphones from English Celex2
60
50
40
30
20
10
0
-0.10
-0.09
-0.08
-0.07
-0.06
-0.05
-0.04
-0.03
-0.02
-0.01
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.10
0.11
0.12
0.13


Absolute value of r_phi for VC biphones from English Celex2
100
90
80
70
60
50
40
30
20
10
0
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.10
0.11
0.12
0.13

References

Ackerman, F. (1992). On the Domain of Lexical Rules: Hungarian Causatives and Wordhood. In I. Kenesei & C. Pléh (eds.), *Approaches to Hungarian IV,* Szeged: JATE.

Albright, A. (2006). How many grammars am I holding up? Discovering phonological differences between word classes. *West Coast Conference on Formal Linguistics*.

Anderson, J. M. (1969). Syllabic or non-syllabic phonology. *Journal of Linguistics,* 5, 136-143.

Anttila, A. (2008). Gradient phonotactics and the complexity hypothesis. *Natural Language and Linguistic Theory,* 26 (4), 695-729.

Baayen, H., Piepenbrock, R. & Gulikers, L. (1996). The CELEX lexical database (release 2) [CD-ROM]. Linguistic Data Consortium, University of Pennsylvania.

Bailey, T. M. & Hahn, U. (2001). Determinants of wordlikeness: phonotactics or lexical neighborhoods? *Journal of Memory and Language,* 44 (4), 568-591.

Barlow, J. A. (2000). A preliminary typology of word-initial clusters with an explanation for asymmetries in acquisition. In R. Kirchner, J. Pater & W. Wikely (eds.), *Papers in Experimental and Theoretical Linguistics: Proceedings of the Workshop on the Lexicon in Phonetics and Phonology,* Edmonton: Department of Linguistics, University of Alberta.

Beesley, K. R. & Karttunen, L. (2000). Finite-state non-concatenative morphotactics. *Proceedings of the Fifth Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON-2000),* pp. 1-12.

Benkő, L. & Imre, S. (1972). *The Hungarian Language.* Budapest: Mouton, Akadémiai Kiadó.

Berg, T. (1994). The sensitivity of phonological rimes to phonetic length. *Arbeiten aus Anglistik und Amerikanistik,* 19, 63-81.

Bird, S., Klein, E. & Loper, E. (2009). *Natural Language Processing with Python.* Cambridge: O'Reilly Media.

Bisani, M. & Ney, H. (2002). Investigations on joint-multigram models for grapheme-to-phoneme conversion. *Proceedings of ICSLP-2002,* pp. 105-108.

Blevins, J. (1995). The syllable in phonological theory. In J. Goldsmith (ed.), *The Handbook of Phonological Theory,* Oxford: Blackwell.

Boersma, P. & Hayes, B. (2001). Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry,* 32, 45-86.

Booij, G. (2000). Morpheme structure constraints and the phonotactics of Dutch. In H. G. van der Hulst & N. A. Ritter (eds.), *The Syllable: Facts and Views,* pp. 69-92. Berlin: Mouton de Gruyter.

Booij, G. E. (1977). *Dutch Morphology: A Study of Word Formation in Generative Grammar.* Lisse: Peter de Ridder Press.

Browman, C. & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology,* 6, 201-51.

Burling, R. (1970). *Man's many voices.* New York: Holt, Rinehart & Winston.

Butskhrikidze, M. (2002). *The Consonant Phonotactics of Georgian.* Utrecht: LOT.

Bybee, J. L. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes,* 10, 425-455.

Bybee, J. L. (2001). *Phonology and language use.* Cambridge [England]: Cambridge University Press.

Campbell, L. (1980). The psychological and social reality of Finnish vowel harmony. In R. Vago (ed.), *Issues in vowel harmony,* pp. 245-270. Amsterdam: John Benjamins.

Carson-Berndsen, J., Kelly, R. & Neugebauer, M. (2004). Automatic acquisition of feature-based phonotactic resources. *Seventh Meeting of the ACL Special Interest Group on Computational Phonology*.

Cebrian, J. (2002). Phonetic similarity, syllabification and phonotactic constraints in the acquisition of a second language contrast. Doctor dissertation, University of Toronto.

Celex (1993). The CELEX lexical database. Centre for Lexical Information, Max Planck Institute for Psycholinguistics.

Chomsky, N. & Halle, M. (1968). *The sound pattern of English.* New York: Harper & Row.

Church, K. W. (1986). Stress assignment in letter to sound rules for speech synthesis. *ICASSP,* pp. 2423-2426.

Clements, G. N. & Keyser, S. J. (1983). *CV phonology: A generative theory of the syllable.* Cambridge: MIT Press.

CMU (1993). The Carnegie Mellon Pronouncing Dictionary v0.1. *Carnegie Mellon University*.

Coetzee, A. (2008). Grammaticality and ungrammaticality in phonology. *Language,* 84, 218-257.

Coetzee, A. & Pater, J. (2008). Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Natural Language and Linguistic Theory,* 26 (2), 289-337.

Cohen, A. (1995). Developing a non-symbolic phonetic notation for speech synthesis. *Computational Linguistics,* 21, 567-575.

Coleman, J. S. & Pierrehumbert, J. (1997). Stochastic phonological grammars and acceptability. *Meeting of the ACL Special Interest Group in Computational Phonology,* Somerset NJ: Association for Computational Linguistics.

Content, A., Mousty, P. & Radeau, M. (1990). BRULEX: Une base de données lexicales informatisée pour le Français écrit et parlé [A lexical computerized database for written and spoken French]. *L'Année Psychologique,* 90 (551-566).

Davis, S. (1982). Rhyme, or reason? A look at syllable-internal constituents. *The Annual Meeting of the Berkeley Linguistic Society,* pp. 525-532.

Davis, S. (1988). Syllable onsets as a factor in stress rules. *Phonology,* 5, 1-19.

Davis, S. (1989a). Cross-vowel phonotactic constraints. *Computational Linguistics,* 15, 109-111.

Davis, S. (1989b). On a non-argument for the rhyme. *Journal of Linguistics,* 25, 211-217.

Davis, S. (1994). Language games. In R. E. Asher & J. M. Y. Simpson (eds.), *The Encyclopedia of Language and Linguistics (1st ed.),* pp. 1980-1985. Oxford: Pergamon Press.

Davis, S. (to appear). Quantity. In J. A. Goldsmith, J. Riggle & A. Yu (eds.), *The Handbook of Phonological Theory (2nd ed.)*: Blackwell.

Deme, L. (1950). Kiejtésünk néhány kérdésről [A few questions on Hungarian pronunciation]. *Magyar Nyelv* 46.
Donegan, P. J. & Stampe, D. (1978). The syllable in phonological and prosodic structure. In A. Bell & J. B. Hooper (eds.), *Syllables and segments,* pp. 25-34. Amsterdam: North Holland.
Dressler, W. & Siptár, P. (1989). Towards a natural phonology of Hungarian. *Acta Linguistica Hungarica,* 39, 29-51.
Dutoit, T., Pagel, V., Pierret, F., Bataille, O. & van der Vrecken, O. (1996). The MBROLA project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. *Proceedings of ICSLP96,* pp. 1393-1396. Philadelphia.
Fekete, L. (1995). *Magyar Kiejtési Szótár [Hungarian pronunciation dictionary]*. Budapest: Gondolat.
Fisher, W. M. (1999). A statistical text-to-phone function using n-grams and rules. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing,* pp. 649-652.
Fowler, C. A. (1987). Consonant-vowel cohesiveness in speech production as revealed by initial and final consonant exchanges. *Speech Communication,* 6, 231-244.
Frauenfelder, U. H., Baayen, R. H., Hellwig, F. & Schreuder, R. (1993). Neighborhood density and frequency across languages and modalities. *Journal of Memory and Language,* 32, 781-804.
Frisch, S., Pierrehumbert, J. & Broe, M. B. (2004). Similarity avoidance and the OCP. *Natural Language and Linguistic Theory,* 22, 179-228.
Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Language,* 47, 27-52.
Fudge, E. C. (1969). Syllables. *Journal of Linguistics,* 5 (2), 53-86.
Fudge, E. C. (1987). Branching structure within the syllable. *Journal of Linguistics,* 23 (2), 359-77.
Fujimura, O. (1976). Syllables as concatenated demisyllables and affixes. *Journal of the Acoustical Society of America,* 59 (Suppl. 1), S55.
Gathercole, S. E., Frankish, C. R., Pickering, S. J. & Peaker, S. (1999). Phonotactic influences on short-term memory. *Journal of Experimental Psychology,* 25, 84-95.
Goldsmith, J. (1995). Phonological theory. In J. A. Goldsmith (ed.), *Handbook of Phonological Theory,* pp. 1-23. Cambridge: Blackwell.
Goldsmith, J. A. (1990). *Autosegmental and Metrical Phonology.* Oxford: Blackwell.
Goldsmith, J. A. & Riggle, J. (2007). Information theoretic approaches to phonological structure: the case of Finnish vowel harmony. University of Chicago.
Good, I. J. (1953). The population of frequencies of species and the estimation of population parameters. *Biometrika,* 40 (3-4), 237-264.
Greenberg, J. H. (1950). The patterning of root morphemes in Semitic. *Word,* 5, 162-181.
Greenberg, J. H. (1978). Some generalizations concerning initial and final consonant clusters. In J. H. Greenberg (ed.), *Universals of Human Language, vol. 2, Phonology,* pp. 243-280. Stanford: Stanford University Press.
Greenberg, J. H. & Jenkins, J. J. (1964). Studies in the psychological correlates of the sound system of American English. *Word,* 20 (2), 157-177.

Grimes, B. F. (1996). Ethnologue: Languages of the World [thirteenth edition]. Dallas: SIL.
Grimes, S. (2005). Moraic weight, extraprosodic word-final consonants, and morphophonological length alterations in Hungarian. *Presentation at the 7th International Conference on the Structure of Hungarian,* Veszprém, Hungary.
Grimes, S. (2007). Word final consonant extrametricality in Hungarian. Indiana University.
Gruenenfelder, T. & Pisoni, D. B. (2006). Modeling the mental lexicon as a complex system: some preliminary results using graph theoretic measures. *Speech Research Laboratory Progress Report*: Indiana University.
Gulikers, L. & Willemse, R. (1992). A lexicon for a text-to-speech system. *ICSLP,* Banff.
Gupta, P. & Dell, G. S. (1999). The emergence of language from serial order and procedural memory. In B. MacWhinney (ed.), *The emergence of language,* pp. 447-481. Hillsdale, NJ: Lawrence Earlbaum Associates.
Haas, M. R. (1969). Burmese disguised speech. In A. S. Dil (ed.), *Language, Culture and History: Essays by Mary S. Haas,* pp. 27-38. Stanford, CA: Stanford University Press.
Halácsy, P., Kornai, A., Németh, L., Rung, A., Szakadát, I. & Trón, V. (2004). Creating open language resources for Hungarian *4th International Conference on Language Resources and Evaluation (LREC2004)*.
Hall, T. A. (1999). Phonotactics and prosodic structure of German function words. In T. A. Hall & U. Kleinholz (eds.), *Studies on the Phonological Word,* Amsterdam/Philadelphia: John Benjamins.
Halle, M. (1962). Phonology in generative grammar. *Word,* 18, 54-72.
Halle, M. (1978). Knowledge unlearned and untaught: What speakers know about the sounds of their language. In M. Halle, J. Bresnan & G. A. Miller (eds.), *Linguistic Theory and Psychological Reality,* pp. 294-303. Cambridge, MA: MIT Press.
Halle, M. & Vergnaud, J. R. (1980). Three dimensional phonology. *Journal of Linguistic Research,* 1, 83-105.
Hammond, M. (2004). Gradience, phonotactics, and the lexicon in English phonology. *International Journal of English Studies,* 4 (2), 1-24.
Hansson, G. (2001). Theoretical and typological issues in consonant harmony. Ph.D. thesis: University of California, Berkeley.
Harris, J. (1994). *English Sound Structure.* Oxford: Blackwell.
Harris, J. W. (1983). *Syllable structure and stress in Spanish: a non-linear analysis.* Cambridge, Mass.: MIT Press.
Haugen, E. (1956). The syllable in linguistic description. In M. Halle, H. B. Lunt, H. McLean & C. H. van Schooneveld (eds.), *For Roman Jakobson, Essays on the Occasion of his Sixtieth Birthday,* The Hague: Mouton.
Hay, J., Pierrehumbert, J. B. & Beckman, M. E. (2003). Speech perception, well-formedness, and the statistics of the lexicon. In J. Local, R. Ogden & R. Temple (eds.), *Phonetic interpretation: papers in laboratory phonology VI,* pp. 58-74. Cambridge: Cambridge University Press.
Hayes, B. (1986). Inalterability in CV Phonology. *Language,* 62 (2), 321-351.

Hayes, B. (1989). Compensatory lengthening in moraic phonology. *Linguistic Inquiry,* 20, 253-306.
Hayes, B. & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry,* 39 (3), 349-440.
Heinz, J. (2006). Learning phonotactic grammars from surface forms. In D. Baumer, D. Montero & M. Scanlon (eds.), *Proceedings of the 25th West Coast Conference on Formal Linguistics,* pp. 186-194. Cascadilla Press.
Heinz, J. (2007). Inductive learning of phonotactic patterns. Ph.D. dissertation, UCLA.
Hill, A. A. (1958). *Introduction to Linguistic Structures.* New York: Harcourt, Brace, and Company.
Hock, H. (1986). Compensatory lengthening: in defense of the concept "mora". *Folia Linguistica,* 20, 431-460.
Hooper, J. B. (1972). The syllable in phonological theory. *Language,* 48, 525-540.
Hyman, L. (1985). *A Theory of Phonological Weight.* Dordrecht: Foris.
Hyman, L. (1992). Moraic mismatches in Bantu. *Phonology,* 9, 255-265.
Ito, J. (1989). A prosodic theory of epenthesis. *Natural Language and Linguistic Theory,* 7, 217-259.
Ito, J. & Mester, A. (1995). Japanese phonology. In J. Goldsmith (ed.), *The Handbook of Phonological Theory,* pp. 817-847. Oxford: Blackwell.
Iverson, G. K. & Wheeler, D. W. (1989). Phonological categories and constituents. In R. Corrigan, F. Eckman & M. Noonan (eds.), *Linguistic Categorization,* pp. 93-114. Amsterdam; Philadelphia: Benjamins.
Jescheniak, J. & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 20 (4), 824-843.
Johnson, C. D. (1972). *Formal Aspects of Phonological Description.* The Hague: Mouton.
Jusczyk, P. W., Luce, P. A. & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language,* 33, 630-645.
Kahn, D. (1980). *Syllable-based Generalizations in English Phonology.* New York: Garland.
Kapatsinski, V. M. (2006). Sound similarity relations in the mental lexicon: Modeling the lexicon as a complex network. *Speech Research Lab Progress Report,* pp. 133-52. Indiana University.
Kapatsinski, V. M. (2007). Implementing and testing theories of linguistic constituency I: English syllable structure. Research on Spoken Language Processing Progress Report No.28, Indiana University Speech Research Lab.
Kassai, I. (1989). On vowel length variability in Hungarian. In T. Szende (ed.), *Proceedings of Speech Research '89,* pp. 96-99. Budapest: Linguistics Institute.
Kawasaki-Fukumori, H. (1992). An acoustical basis for universal phonotactic constraints. *Language and Speech,* 35, 73-86.
Kaye, J., Lowenstamm, J. & Vergnaud, J. R. (1985). The internal structure of phonological elements: A theory of charm and government. *Phonology Yearbook,* 2, 303-326.
Kelly, M. H. (1991). Using sound to solve syntactic problems: the role of phonology in grammatical category assignments. *Psychological Review,* 99, 349-364.

Kenesei, I., Vago, R. M. & Fenyvesi, A. (1998). *Hungarian.* New York: Routledge.
Kenstowicz, M. (1994). *Phonology in Generative Grammar.* Cambridge, Mass.: Blackwell.
Keresztes, L. (1992). *A practical Hungarian grammar.* Debrecen: Debreceni Nyári Egyetem.
Kessler, B. & Treiman, R. (1997). Syllable structure and the distribution of phonemes in English syllables. *Journal of Memory and Language,* 37, 295-311.
Kiss, J. (2001). *Magyar dialektológia [Hungarian dialectology]*. Budapest: Osiris Kiadó.
Kisseberth, C. W. (1973). Is rule ordering necessary in phonology? In B. e. a. Kachru (ed.), *Papers in linguistics in honor of Henry and Renee Kahane,* Urbana: University of Illinois Press.
Kontra, M. (1995). On current research into spoken Hungarian. *International Journal of the Society of Language,* 111, 9-20.
Kornai, A. (1986). Szótári adatbázis az akadémiai nagyszámitógépen [A dictionary database of Hungarian]. *Working Papers,* pp. 65-79. Budapest: Hungarian Academy of Sciences Institute of Linguistics
Kornai, A. (1990). The sonority hierarchy in Hungarian. *Nyelvtudományi Közlemények,* 91, 139-146.
Kornai, A. (1991). Hungarian vowel harmony. In I. Kenesei (ed.), *Approaches to Hungarian III,* pp. 183-240. Szeged: JATE.
Kuĉera, H. & Francis, W. N. (1967). *Computational analysis of present-day American English.* Providence, R.I.: Brown University Press.
Kurylowicz, J. (1949). La notion de l'isomorphisme. *Travaux du Cercle Linguistique de Copenhague* 5,48-60.
LDC (1995). COMLEX English Pronouncing Lexicon version 0.2. Philadelphia: Linguistic Data Consortium.
Lee, Y. (2006). Sub-syllabic constituency in Korean and English. Doctoral dissertation, Northwestern University.
Lee, Y. & Goldrick, M. (2008). The emergence of sub-syllabic representations. *Journal of Memory and Language,* 59, 155-168.
Levelt, W. J. M. (1992). Accessing words in speech production: stages, processes and representations. *Cognition,* 42, 1-22.
Liberman, M. & Church, K. W. (1992). Text analysis and word pronunciation in text-to-speech synthesis. In S. Furui & M. M. Sondhi (eds.), *Advances in Speech Signal Processing,* pp. 791-832. New York: Marcel Dekker.
Loper, E. & Bird, S. (2002). NLTK: the natural language toolkit. *ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics,* pp. 62-69. Philadelphia: Association for Computational Linguistics.
Lowenstamm, J. (1996). CV as the only syllable type. In J. Durand & B. Laks (eds.), *Current Trends in Phonology: Models and Methods,* pp. 419-441. CNRS, ESRI, Paris X.
Luce, P. A. (1986). Neighborhoods of words in the mental lexicon *Research on Speech Perception,* Bloomington, IN: Speech Research Laboratory, Indiana University.
Luce, P. A. & Pisoni, D. B. (1998). Recognizing spoken words: the neighborhood activation model. *Ear and Hearing,* 19, 1-36.

Lutz, A. (1988). On the historical phonotactics of English. In D. Kastovsky & G. Bauer (eds.), *Luick Revisited,* pp. 221-239. Tübingen: Narr.
MacKay, D. G. (1972). The structure of words and syllables: Evidence from errors in speech. *Cognitive Psychology,* 3, 210-227.
MacKay, D. G. (1973). Spoonerisms: The structure of errors in the serial order of speech. In V. A. Fromkin (ed.), *Speech errors as linguistic evidence,* pp. 164-194. The Hague: Mouton.
Mattys, S. L. & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition,* 78, 91-121.
McCarthy, J. J. (1976). On hierarchical representation within syllables. Cambridge, Massachusetts: MIT. Unpublished manuscript.
McCarthy, J. J. (1979). On stress and syllabification. *Linguistic Inquiry,* 10, 443-466.
McCarthy, J. J. & Prince, A. (1986/1996). Prosodic Morphology 1986. New Brunswick, NJ: Rutgers University Center for Cognitive Science.
Menn, L. (2004). Saving the baby: making sure that old data survive new theories. In R. Kager, J. Pater & W. Zonneveld (eds.), *Constraints in Phonological Acquisition,* pp. 54-72. Cambridge University Press.
Metsala, J. L. (1997). An examination of word frequency and neighborhood density in the development of spoken word recognition. *Memory and Cognition,* 25, 47-56.
Nádasdy, Á. (1989a). Consonant length in recent borrowings into Hungarian. *Acta Linguistica Hungarica,* 39, 195-213.
Nádasdy, Á. (1989b). The exact domain of consonant degemination in Hungarian. In K. Bolla, M. Gósy, J. Herman, G. Olaszy & T. Szende (eds.), *Speech Research '89,* Budapest: MTA Nyelvtudományi Intézete.
Nádasdy, Á. & Siptár, P. (1998). Vowel length in present-day Hungarian. *The Even Yearbook,* 3, 149-172.
Neef, M., Neijt, A. & Sproat, R. W. (eds.) (2002). *The Relation of Writing to Spoken Language.* Tübingen: Niemeyer.
Németh, G., Zainkó, C., Kiss, G., Fék, M., Olaszy, G. & Gordos, G. (2003). Language processing for name and address reading in Hungarian. *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering,* pp. 238-243.
Nespor, M. & Vogel, I. (1986). *Prosodic phonology.* Dordrecht: Foris.
Nusbaum, H. C., Pisoni, D. B. & Davis, C. K. (1984). Sizing up the Hoosier Mental Lexicon: Measuring the familiarity of 20,000 words. *Research on Speech Perception Progress Report,* 10, 357-376.
Obendorfer, R. (1975). The ambiguous status of Hungarian long consonants. *Lingua,* 36 (4), 325-336.
Ohala, J. J. & Kawasaki-Fukumori, H. (1997). Alternatives to the sonority hierarchy for explaining the shape of morphemes. In S. Eliasson & E. Hakon Jahr (eds.), *Studies for Einar Haugen,* pp. 343-365. Berlin: Mouton de Gruyter.
Ohala, J. J. & Ohala, M. (1986). Testing hypotheses regarding the psychological reality of morpheme structure constraints. In J. J. Ohala & J. J. Jaeger (eds.), *Experimental Phonology,* pp. 239-252. San Diego Academic Press.

Olaszy, G. (2003). Magyar szóalakok hangidő-térképei [Duration charts of Hungarian word forms]. In M. Gósy (ed.), *Beszédkutatás 2003,* pp. 113-134. Budapest: MTA Nyelvtudományi Intézet.
Olaszy, G. & Kálmán, A. (2005). Adatbázisok és számítógépprogramok a magyar beszéd időszerkezeti vizsgálatához *Alkalmazott Nyelvtudomány,* V (1-2), 41-62.
Papp, F. (1969). *A Magyar Nyelv Szóvégmutató Szótára [Reverse-Alphabetized Dictionary of the Hungarian Language]*. Budapest: Akadémiai Kiadó.
Perruchet, P. & Peereman, R. (2004). The exploitation of distributional information in syllable processing. *Journal of Neurolinguistics,* 17, 97-119.
Pike, K. & Pike, E. (1947). Immediate constituents of Mazateco syllables. *International Journal of American Linguistics,* 13, 78-91.
Pintzuk, S., Kontra, M., Sándor, K. & Borbély, A. (1995). The effect of the typewriter on Hungarian reading style. *Working Papers in Hungarian Sociolinguistics*.
Polgárdi, K. (2005). Geminates and degemination in Hungarian. *International Conference on the Structure of Hungarian* Veszprém, Hungary.
Port, R. (2007a). How are words stored in memory? Beyond phones and phonemes. *New Ideas in Psychology,* 25, 143-170.
Port, R. F. (2007b). The graphical basis of phones and phonemes. In O.-S. Bohn & M. J. Munro (eds.), *Language Experience in Second Language Learning: In honor of James Emil Flege,* pp. 349-365. Amsterdam: John Benjamins.
Prince, A. & Tesar, B. (2004). Learning phonotactic distributions. In R. Kager, J. Pater & W. Zonneveld (eds.), *Constraints on Phonological Acquisition,* pp. 245-291. Cambridge University Press.
Pycha, A. (2007). Phonetic vs. phonological lengthening in affricates. *Proceedings of the 16th International Conference on the Phonetic Sciences*, 1757-1760.
Rebrus, P. & Trón, V. (2002). A fonotaktikai általánosításokról: Kísérlet a magyar mássalhangzó-kapcsolatok nem-reprezentációs újabb módszerei [On phonotactic generalizations: an attempt at a non-representational description of Hungarian consonant clusters]. In M. Maleczki (ed.), *A mai magyar nyelv leírásának újabb módszerei [New methods in present-day Hungarian language description],* pp. 17-63. Szeged: Szegedi Tudományegyetem.
Rebrus, P. & Trón, V. (2005). Re-presenting the past: Contrast and uniformity in Hungarian past tense suffixation. In I. Kenesei, C. Pinón & P. Siptár (eds.), *Approaches to Hungarian,* Budapest: Akadémiai Kiadó.
Ringen, C. O. & Vago, R. M. (2006). Geminates: Heavy or Long? : manuscript.
Roelofs, A. (1996). Serial order in planning the production of successive morphemes of a word. *Journal of Memory and Language,* 35 (854-876).
Rose, S. & Walker, R. (2004). A typology of consonant agreement as correspondence. *Language,* 80 (3), 475-531.
Rosenthall, S. & Van der Hulst, H. (1999). Weight-by-position by position. *Natural Language and Linguistic Theory,* 17 (3), 499-540.
Rot, S. (1994). *Hungarian. Its Origins and Originality.* Budapest: Korona Publishing House.
Saporta, S. (1963). Phoneme distribution and language universals. In J. H. Greenberg (ed.), *Universals of Language,* pp. 61-67. Cambridge, Massachusetts: MIT Press.

Seidenberg, M. S. & Gonnerman, L. M. (2000). Explaining derivational morphology as the convergence of codes. *Trends in Cognitive Science,* 4, 353-361.
Sejnowski, T. & Rosenberg, C. (1987). NETtalk: a parallel network that learns to pronounce English text. *Complex Systems,* 1, 145-168.
Selkirk, E. (1978). On prosodic structure and its relation to syntactic structure. *The Conference on Mental Representation of Phonology*.
Selkirk, E. (1982a). Prosodic domains in phonology: Sanskrit revisited. In M. Aronoff & M.-L. Kean (eds.), *Juncture (Studia Linguistica et Philologica 7),* Sarasota, CA: Anma Libri.
Selkirk, E. (1982b). The syllable. In H. G. van der Hulst & N. Smith (eds.), *The Structure of Phonological Representations, volume II,* Dordrecht: Foris.
Selkirk, E. (1984). *Phonology and Syntax: The Relationship between Sound and Structure.* Cambridge: MIT Press.
Selkirk, E. (1990). A two-root theory of length. *University of Massachusetts Occasional Papers,* (14).
Share, D. L. & Blum, P. (2005). Syllable splitting in literate and preliterate Hebrew speakers: Onsets and rimes or bodies and codas. *Journal of Experimental Child Psychology,* 92 (2), 182-202.
Siptár, P. (1989). On fast speech. *Acta Linguistica Hungarica,* 39, 215-224.
Siptár, P. (1994). The vowel inventory in Hungarian: its size and structure. *The even yearbook*, 175-184.
Siptár, P. & Törkenczy, M. (2000). *The Phonology of Hungarian.* Oxford: Oxford University Press.
Smith, J. (2001). Lexical category and phonological contrast. In R. Kirchner, J. Pater & W. Wikely (eds.), *Papers in experimental and theoretical linguistics 6: Workshop on the Lexicon in Phonetics and Phonology,* pp. 61-72. Edmonton: University of Alberta.
Smolensky, P. & Legendre, G. (2006). *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar.* Cambridge, MA: MIT Press.
Souter, C. (1993). Harmonising a lexical database with a corpus-based grammar. In C. Souter & E. Atwell (eds.), *Corpus-based Computational Linguistics,* pp. 181-193. Amsterdam, Atlanta: Rodopi.
Spencer, A. (1996). *Phonology: theory and description.* Oxford: Blackwell.
Sproat, R. W. (1992). Morphology and Computation. Cambridge, MA: MIT Press.
Sproat, R. W. (2000). *A Computational Theory of Writing Systems.* Cambridge: Cambridge University Press.
Stemberger, J. P. (1983). *Speech errors and theoretical phonology: a review.* Bloomington: Indiana University Linguistics Club.
Steriade, D. (1999). Alternatives to syllable-based accounts of consonantal phonotactics. In O. Fujimura, B. D. Joseph & B. Palek (eds.), *Proceedings of LP '98: Item Order in Language and Speech,* pp. 205-242. Karolinum Press: Prague.
Storkel, H. L. & Rogers, M. A. (2000). The effect of probabilistic phonotactics on lexical acquisition. *Clinical Linguistics and Phonetics,* 14 (6), 407-425.
Szemere, G. (1987). *Hogy is írjuk? [How should we write?].* Budapest: Gondolat.
Szende, T. (1994). Hungarian: Illustrations of the IPA. *International Journal of Phonetics,* 4, 91-94.

Szigetvári, P. (1999). VC Phonology: A theory of consonant lenition and phonotactics. Doctoral dissertation. Eötvös Loránd University/MTA, Budapest.
Szigetvári, P. (2001). Dismantling syllable structure. *Acta Linguistica Hungarica,* 48, 155-181.
Thorn, A. & Frankish, C. R. (2005). Long-term knowledge effects on serial recall of nonwords are not exclusively lexical. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 31, 729-735.
Törkenczy, M. (1989). Does the onset branch in Hungarian? *Acta Linguistica Hungarica,* 39, 272-292.
Törkenczy, M. (1994). A szótag [The syllable]. In F. Kiefer (ed.), *Strukturális magyar nyelvtan, 2. kötet: Fonológia [Structural Hungarian Grammar volume 2: Phonology],* pp. 272-392. Budapest: Akadémiai Kiadó.
Törkenczy, M. (2001). Phonotactic grammaticality and the lexicon. *Acta Linguistica Hungarica,* 48, 137-153.
Törkenczy, M. (2006). The phonotactics of Hungarian verbs. *The Even Yearbook,* 7.
Törkenczy, M. & Siptár, P. (1999). Hungarian syllable structure: Arguments for/against complex constituents. In H. van der Hulst & N. A. Ritter (eds.), *The Syllable: Views and Facts,* pp. 249-284. Berlin: Mouton de Gruyter.
Tótfalusi, I. (2006). *Kiejtési szótár: Idegen nevek, szavak helyes kiejtése [Pronunciation dictionary: The correct pronunciation of foreign names and words]*. Budapest: Tinta Kiadó.
Tranel, B. (1991). CVC light syllables, geminates, and moraic theory. *Phonology,* 8 (291-302).
Treiman, R. & Danis, C. (1988). Short-term memory errors for spoken syllables are affected by the linguistic structure of the syllables. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 14, 145-152.
Treiman, R., Kessler, B., Tincoff, R. & Bowman, M. (2000). English speakers' sensitivity to phonotactic patterns. In M. B. Broe & J. B. Pierrehumbert (eds.), *Papers in Laboratory Phonology V: Acquisition and the Lexicon,* pp. 269-282. Cambridge: Cambridge University Press.
Trnka, B. (1936). General laws of phonemic combinations. *Travaux du Cercle Linguistique de Prague,* 6.
Trón, V., Halácsy, P., Rebrus, P., Rung, A., Vajda, P. & Simon, E. (2006). Morphdb.hu: Hungarian lexical database and morphological grammar. *In proceedings of LREC 2006,* pp. 1670-1673.
Trón, V., Németh, L., Halácsy, P., Kornai, A., Gyepesi, G. & Varga, D. (2005). Hunmorph: open source word analysis. *ACL*.
Trón, V. & Rebrus, P. (2001). Morphophonology and the hierarchical lexicon. *Acta Linguistica Hungarica,* 48, 101-135.
Trubetzkoy, N. S. (1939). *Grundzüge der Phonologie. [Principles of phonology]*. (1st edition 1939, translated from German by C. A. M. Baltaxe). Berkeley: University of California Press, 1969.
Vago, R. M. (1980). *The Sound Pattern of Hungarian.* Washington, D.C.: Georgetown University Press.

Vago, R. M. (1992). The root analysis of geminates in the moraic phonology of Hungarian. In I. Kenesei & C. Pléh (eds.), *Approaches to Hungarian,* Szeged: JATE.

van den Bosch, A. & Canisius, S. (2006). Improved morpho-phonological sequence processing with constraint satisfaction inference. *Eighth Meeting of the ACL Special Interest Group on Computational Phonology at HLT-NAACL 2006,* New York City: Association for Computational Linguistics.

van den Bosch, A. & Daelemans, W. (1993). Data-oriented methods for grapheme-to-phoneme conversion. *Proceedings of the European Chapter of ACL,* pp. 45-53. Utrecht.

Váradi, T. (2002). The Hungarian National Corpus. *Proceedings of the Third International Conference on Language Resources and Evaluation,* pp. 385-389. Las Palmas, Spain.

Vennemann, T. (1978). Universal syllabic phonology. *Theoretical Linguistics,* 5, 175-215.

Vennemann, T. (1988). The rule dependence of syllable structure. In C. Duncan-Rose & T. Vennemann (eds.), *On language: Rhetorica, Phonologica, Syntactica: A Festschrift for Robert P. Stockwell from his friends and colleagues,* pp. 257-283. London: Routledge.

Vitevitch, M. S. & Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers,* 36 (3), 481-487.

Vitevitch, M. S., Luce, P. A., Charles-Luce, J. & Kemmerer, D. (1997). Phonotactics and syllable stress: implications for the processing of spoken nonsense words. *Language and Speech,* 40, 47-62.

Vogel, I. (1988). Prosodic constituents in Hungarian. In P. M. Bertinetto & M. Loporcaro (eds.), *Certamen Phonologicum. Papers from the 1987 Cotrona Phonology Meeting,* pp. 231-50. Torino, Italy: Rosenberg & Sellier.

Wickelgren, W. A. (1969). Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review,* 86 (44-60).

Yi, K. (1999). The internal structure of Korean Syllables. *Second International Conference on Cognitive Science and 16th Annual Meeting of the Japanese Cognitive Science Society,* Tokyo.

Zipf, G. K. (1935). *The psycho-biology of language: An introduction to dynamic philology.* Cambridge, M.A.: Houghton Milton.

VITA

Stephen Matthew Grimes attended North Ridgeville High School in North Ridgeville, Ohio. In 1995 he matriculated at Bucknell University in Lewisburg, PA. He spent the summers of 1997 and 1998 performing theoretical mathematics research at Lafayette College in Easton, PA and Rutgers University in New Brunswick, NJ. The spring of 1998 was spent studing at the Budapest Semesters in Mathematics in Budapest, Hungary. He received the degree of Bachelor of Science magna cum laude from Bucknell University in May, 1999 while minoring computer science. He was member of the Phi Beta Kappa honorary society and a Barry S. Goldwater Scholar.

In August 1999, Stephen entered the Graduate School at Indiana University in Bloomington as a Ph.D. student in mathematics (but later switched to linguistics). He was supported by the Government Assistance in Areas of National Need and the Fellowship for Language and Area Study programs. He taught as an Associate Instructor in the Mathematics, Linguistics, and Psychology departments and held positions at the Library Electronic Text and Resource Archive and the American Indian Studies Research Institute. As a graduate exchange fellow, he spent eight months studying at the University of Debrecen in Hungary. The Master of Arts degree in Mathematics was earned in December 2000 and the Master of Arts degree in Linguistics was conferred in May 2003.

Stephen worked briefly in 2001 at Lernout and Hauspie in Burlington, MA and Dragon Systems in Newton, MA as a language model engineer for speech recognition. Since late 2008 he has worked as a programmer analyst at the Linguistic Data Consortium at the University of Pennyslvania in Philadelphia.