On the creation of a pronunciation dictionary for Hungarian

Stephen M. Grimes
stgrimes@indiana.edu
August 2007

Abstract

This report describes the process of creating a pronunciation dictionary and phonological lexicon for Hungarian for the purpose of aiding in linguistic research on Hungarian phonology and phonotactics. The pronunciation dictionary was created by transforming orthographic forms to pronunciation representations by taking advantage of systematic deviations between Hungarian orthography and pronunciation. It is argued that the “automated” creation of such a dictionary is reasonably expected to be accurate due to the relative similarity of Hungarian orthography to actual pronunciation. This document includes discussion of goals and standards for creating a Hungarian pronunciation dictionary, and each phonological change creating a mismatch between orthography and pronunciation is highlighted. Future developments and additions to the current dictionary are also suggested as well as strategies for evaluating the quality of the dictionary. Finally, potential applications to linguistic research are discussed.

1 Introduction

While students of the English language quickly learn that English spelling is by no means 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/ [sabatʃ:a:g] ‘freedom, liberty’[1], in which no fewer than four orthographic-pronunciation discrepancies can be identified with the written form of this word:

(1) a. The sequence /sz/ is a digraph corresponding to the sound [s] while /s/ corresponds to [ʃ].

b. A general process of voicing assimilation applying between two consonants requires the [dʃ] to be pronounced [tʃ] at some intermediate level of representation.

c. The [tʃ] consonant cluster further undergoes affrication/coalescence and is pronounced [č:]. (The use *č* here is symbol manipulation to indicate the affricate is treated as a single sound instead of as a sound sequence.)

d. The acute accent on the vowel /á/ indicates vowel length – compare /a/ [ɔ] and /á/ [a:]. The issue is 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 systematic, as are the majority of the sound-symbol discrepancies in Hungarian. Hence one is able to develop a system of replacement rules which rewrite the grapheme strings into a phonemic transcription that is unambiguous with respect to pronunciation. Any exceptional word – one in which the deviation between orthography and pronunciation is not systematic – cannot be handled by a rewrite rule and instead will be listed as an exception. The list of

[1] In this report I adopt the practice of enclosing graphemes with /forward slashes/ and pronunications using [square brackets].

exceptions can be thought of as a lexicon, whereas the rewrite rules comprise a strange sort of grammar that is “eerily” similar to the actual phonological grammar of Hungarian.

Several pronunciation dictionaries or phonological lexicons exist for English, including the Hoosier Mental Lexicon (Nusbaum et al., 1984), the Carnegie Mellon Pronouncing Dictionary (CMU, 1993), PRONLEX distributed by the Linguistic Data Consortium, and the CELEX2 database (Baayen et al., 1996). For languages other than English, CELEX2 also contains lexicons for German and Dutch. ELRA (European Language Resources Association) distributes phonetic lexicons based on Spanish and Catalan. 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. However, 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 costs upwards of several thousands of dollars for usage rights.

Additionally, there is distinct a lack of 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 necessarily based exclusively on English. Thus, the development of a pronunciation dictionary for Hungarian, a Finno-Ugric language, offers opportunity to study a lexicon that is not derived from the Indo-European word stock. Hungarian is noteworthy because it is a so-called agglutinative language with a relatively high morpheme-to-word ratio, meaning that most words are likely polymorphemic. Additionally, Hungarian also has a more complex verbal inflection system than Germanic or Romance language families – the language families for which pronunciation dictionaries available to date. Furthermore, several computational tools are already available for Hungarian (Halácsy et al., 2004, Kornai, 1986, Váradi, 2002), meaning that 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. Finally, it must be stressed that due to the relatively close relationship between orthography and pronunciation, Hungarian is conducive to the supervised, automatic creation of a pronunciation dictionary. This will be illustrated in Section 3, where the aspects of Hungarian phonology not already reflected in the writing system are discussed in detail.

While some of the uses for a pronunciation dictionary will be discussed in Section 5, it should be noted at this time that this dictionary is not intended to be used as a definitive guide to the pronunciation of words in Hungarian. This is accomplished by publications directed to the public at large, such as *The Hungarian Pronunciation Dictionary* (Fekete, 1995), which aids L2 Hungarian speakers or L1 speakers living outside Hungarian 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, as these resources only include hard-to-pronounce words or list multiple pronunciations for each word, leaving the reader baffled as to which is the preferred pronunciation or whether the pronunciation variation is across dialects or speakers. These resources are also inadequate for determining the correct vowel length of certain high vowels; a personal motivation for undertaking this work as an L2 Hungarian learner was to sort out length vacillation in instances where the orthography is inconsistent.

2 Goals and design requirements for a 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 comparative work on Hungarian. For approximately 20,000 English words, the HML supplies both written forms and broad phonetic transcriptions in a phonetic alphabet. It also includes additional data – the length of the phonetic form (raw segment count), its consonant-vowel structural makeup, the corpus frequency of the word, word familiarity ratings, and additional information.

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. Due to licensing restrictions in place at this time requiring 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 it is possible as with the freely downloadable Kornai Corpus.

2.1 Contents of the Hungarian pronunciation dictionary

The present version of the pronunciation dictionary includes the following information for each word:

(2) (a) Orthographical form (from Kornai, 1986)
(b) Pronunciation (present work)
(c) CV tier representation (present work[2])
(d) Syllable structure (present work)
(e) Frequency counts (integrated from Halácsy et al., 2004)

This report focuses primarily on the relationship between the items in (2a) and (2b) – the creation of a pronunciation from each orthographic form. My definition of a phonological lexicon distinguishes it from a pronunciation dictionary in that a pronunciation dictionary contains a subset of the richer linguistic data found in a phonological lexicon. Hence the CV tier, syllable structure, and frequency count data would extend the pronunciation dictionary into being a phonological lexicon. As this extension is turns out to be less labor-intensive and more linguistically interesting, these secondary issues are addressed only in section 5.4.

2.2 Dialects and Idiolects

For certain words, more than one pronunciation is possible, and this variation can be across dialects, registers, or individual speakers. I addressed this issue by making a decision to only select one pronunciation for each written form. When possible, the most frequent variant is chosen. I recognize that it is a simplifying assumption or idealization to suppose that a unique pronunciation exists for each word; however, such assumptions

[2] 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.

are typical in other pronunciation dictionaries. If two pronunciations are equally accepted, I choose the one that remains closer to the written form. I have tried to pick 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, known as Educated Colloquial Hungarian (ECH), as opposed to Standard Literary Hungarian (SLH) or one of the various regional dialects. 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 the variant phonological processes in the minority dialects of Hungarian, I can refer the reader to the general overviews provided by Rot (1994) and Kiss (2001).

2.3 Character encodings

Because the Hungarian alphabet uses characters that are not included in the basic ASCII character standard[3], it is often difficult to transport Hungarian computer files between different machines without experiencing problems of encoding – an individual most know the encoding of a file in order to interpret it properly. While the development and adoption of the Unicode standard promises to eliminate these hassles in the future, I expect widespread adoption will not be completed for at least another decade. Hence the default character encoding scheme for this pronunciation dictionary is based entirely on ASCII characters. The alphabet I chose to use is based on Péter Szigetvári's OGOB7, or one-grapheme-one-byte – this encoding schema enforces a principle of one-to-one mapping of sound to symbol. I will refer to my modified version of OGOB7 simply as

[3] 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.

OGOB.[4] While the name of this encoding system might not be 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. The advantage of one-grapheme-one-byte system is perhaps already apparent to the reader – suppose one wants to search for all instances consonant clusters of length exactly equal to two. One would not want to find false positive such as /sz/ (which represents the single sound [s]), nor would it be desirable to fail to find valid cases such as /nsz/ [ns]. The table in (3) gives the digraphs and trigraphs of Hungarian with their single character encoding equivalent.

(3) Encodings of digraphs and trigraphs in OGOB

Hungarian Orthography	cs	ch[5]	dz[6]	dzs	gy	ly	ny	sz	ty	Zs
IPA	tʃ	x	--	ʤ	ɟ	j	ɲ	s	c	ʒ
OGOB	C	H	--	D	G	j[7]	N	S	T	Z

In the case of 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

[4] 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/. Hence I do not consider it particularly desirable to recover the original spelling of a word given the pronounced form.

[5] 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.

[6] The digraph /dz/ is widely considered 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, Péter, and Törkenczy, Miklos. 2000. *The phonology of Hungarian*. Oxford: Oxford University Press.

[7] 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.

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. Szigetvári, 2001, Vago, 1992). I might add also that in the vowel system, there are also 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 not unwarranted. The orthography and corresponding character encoding in OGOB is given for the vowels in (4).

(4) 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

The remaining characters used in OGOB and not appearing in (3) or (4) 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 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 Proszéky encoding, an early system 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). This was an earlier means of working around the ASCII encoding issues.

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 transcriptions are distinct from OGOB transcriptions in that multiple symbols often correspond to a single phoneme, and hence transcriptions utilize whitespace to delimit phones.

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 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, often conforms to what could be called the etymological principle (Keresztes, 1992:31). 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, Kenesei et al., 1998, Kontra, 1995, Nádasdy, 1989a, Nádasdy, 1989b, Nádasdy and Siptár, 1998, Pintzuk et al., 1995). Pronunciation and orthographical mismatches can be broadly grouped into one of three categories: (1) words which retain historical spellings, especially prevalent in place names and person names, (2) orthographic issues relating to the alphabet, i.e. digraphs and trigraphs, and (3) 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 4, while the remaining issues are discussed in this section.

As suggested by my use of slashes and square brackets (// and []) to distinguish graphemes and phonemes, the correspondence between orthography and pronunciation in many ways resembles the correspondence that linguists assert exists between underlying and surface forms. In a sense, creating a pronunciation dictionary for Hungarian is similar to identifying and implementing a rule-ordered system of generative phonology in which the output of one rule serves as the input for the next in a successive chain of alterations (cf. Vago, 1980). With the exception of a few words in which pronunciation cannot be reliably deduced from orthography, in general it appears possible to map a given orthographic form to a phonetic form.

3.1 Simple substitutions

Certain sounds and sound combinations in Hungarian have two possible spelling variants. While most letters used in Hungarian are similar to the IPA or in this case OGOB symbol used in transcription, there are a small handful of letters for which a direct substitution may be made:

(5) 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

Because the OGOB encoding system uses /w/ and /y/ to stand in place of umlauted vowels, it is necessary to run a rule making the replacements suggested by the table in (5) to eliminate the /w/ and /y/ graphemes with the replacements suggested by OGOB *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.[8] The order of presentation of phenomena in this report mirrors the actual order of implementation of the rules.

3.2 Divergent spelling conventions

The possibility of different or multiple spellings of a word represents an expansion of the issue described in the table in (5). 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 (Benkõ and Imre, 1972). As a result of the relatively recent standardization, Hungarian spelling quite 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

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

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 traditions.

(6) Modern orthography as combination of two traditions

Modern orthography	Catholic tradition	Protestant tradition
cs [tʃ]	cs	ts
c [ts]	cz	tz
bántja [ba:ncɔ] 'hurt-3S.DEF'	bántya	bántja
látja [ba:c:ɔ] 'she sees it'	láttya	látja

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. Several additional letters are used in proper names or words of foreign origin (Keresztes, 1992: 30). These letters 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. A summary of these spellings is given in the table in (7).

(7) Non-standard spellings retained in family names

Historical Spelling	Modern Equivalent	Example
aa, aá	á	Gaal, Gaál
eé	é	Veér
eö, eő	ö, ő	Eötvös, Beőthy
ew	ö	Thewrewk [török]
oó	ó	Soós
uu	ú	Kuun
y	i	Ady
ch	cs	Madách
cz	c	Rákóczi
tz	c	Atzél
w	v	Wessenlényi
th	t	Toth, Batthyány

Note that for the present purposes of developing a pronunciation dictionary of Hungarian, these names are mostly peripheral. To my knowledge the Hungarian wordlists I am using do not contain proper names; this omission is not entirely principled, as the pronunciation of proper names usually obey the rules of the phonology of a language, although some phonotactic constraints may be relaxed. For most examples listed in the table above[9] there is no reason not to include rules to rewrite the foreign spellings first in modern orthography and then continue converting this representation to a phonetic transcription.

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

[9] In a few cases, it is not possible to integrate rules relating to some names, as 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/ [császár].

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 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 digraphs with a single ACSII 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 and (4) 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. The following are examples are due to 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.

(8) 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’)[10]

One strategy that was not employed in the present work would be to use probabilistic heuristics to determine whether a potential digraph is a true digraph or simply a segment sequence. For example, for historical reasons the digraph /ly/ (pronounced [j]) is more likely to occur at the end of multisyllabic words than word internally (Szemere, 1987). As a result, the word *muszáj* ‘must’ is incorrectly spelled *muszály* approximately ten percent of the time.[11] 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 (8) only involve compounds or derived words. (In the case of derived words only the stem can be looked 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 whereby the potential grapheme sequence is maximized is always used.

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

[11] This data is based on a Google search of Hungarian web pages in late 2006 that found *múszály* occuring 118,000 times compared to the standard *múszáj* appearing 1,090,000 times.

4 Phonology and morphophonology

In order to survey phonological processes of Hungarian, I consulted a variety of Hungarian grammars, dictionaries, and papers (e.g. Bosch and Daelemans, 1993, Kenesei et al., 1998, Keresztes, 1992, Papp, 1969, Siptár and Törkenczy, 2000, Törkenczy, 1994, Vago, 1980), as well as a number of guides to correct Hungarian writing and spelling styles. Some phonological processes are reflected in Hungarian orthography while others are not. 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 is discussed in detail.

4.1 Assimilation of nasals to place of articulation

A nasal consonant must agree with the specified value of the place of articulation feature of a following consonant. While a backed variant of the nasal appears before velars and palatals, all the examples in (9) 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 I do not use it in the pronunciation dictionary.

(9) Examples of nasal place assimilation

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

Now is an appropriate time to discuss strategies used to create the dictionary. While a linguist may seek to identify the most general statement of a rule – for example stating that a nasal must agree in place of articulation of a following consonant. This general rule is given in (10a), while a concrete instantiation appears in (10b).

(10) a. $N \rightarrow [\alpha \text{ place}] / \_\_ C_{[\alpha \text{ place}]}$

b. n → m / __ {b, p, f, v}

The formulation in (10b) is more specific, 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 (10a) requires detailed feature data for each phoneme, while using the option in (10b) is less cumbersome and has the advantage of being very specific.

4.2 Voicing assimilation

Hungarian consonant clusters must agree in voicing, and the assimilation process is anticipatory (also termed regressive assimilation). The voicing feature all consonants in a cluster must agree with the voicing feature; in instances of triconsonantal clusters across morpheme boundaries, this rule must apply iteratively. However, some consonants are exceptional: /h/, /j/, /m/, /n/, /ny/, and /r/ do not undergo assimilation. Most of these exceptional cases are due to the segment not having a voiced or voiceless counterpart. Furthermore, [v] does not seem to cause assimilation. Consonant clusters appearing in native stems already agree in their voicing features. Examples of illicit clusters resolved through morphophonology either involve loanwords (11a), affixed forms (11b), or (11c) compound words.

(11) 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	[olvɔʒd ɛl]	‘read it’
kútban	[ku:dbɔn]	‘in the well’
c. népdal	[ne:bdɔl]	‘folksong’
húsdarálo	[hu:ʒdɔra:lo]	‘meatgrinder’
kerékgyártó	[kɛrɛɟ:a:rto:]	‘wheelmaker’

As stated above, the rule in (12) 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].

(12) $C \rightarrow C_{[\alpha \text{ voice}]}$ / ____ $C_{[\alpha \text{ voice}]}$

4.3 Coronal palatalization

A coronal stop is palatalized before the imperative morpheme or third person singular verbal suffix /j/. The result is coalescence, but the moraic timing of the component segments is preserved – in other words, the resulting palatal is a geminate.

(13) Palatalization of coronal stops involving [j] imperative/subjunctive 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’

4.4 Alveolar plosive affrication

When morphology creates 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 sibilant place of articulation. The resulting affricate is usually a long consonant unless reduced due to being adjacent to another consonant (a consonant cluster reduction rule is discussed in Section 4.8). A rule giving the relevant segments involved is given in (14) and examples appear in (15).

(14) t, t^y → tʃ: / __ ʃ
t, t^y → ts: / __ s

(15) Examples of alveolar plosive affrication

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

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

4.5 Hiatus resolution

A glide consonant [j] is inserted to interrupt hiatus between two vowels whenever one of the vowels is *i* or í. The pattern is less clear for vowel sequences involving *é*, 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.

(16) 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 í.

(17) $\varnothing \rightarrow$ j / __ {i, í}
$\varnothing \rightarrow$ j / {i, í}__

4.6 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 low functional load of high vowels. 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 (18) to apply to foreign borrowings such as *unió* [u:nijo:] ‘(European) union’.

(18) $V_{[+high]} \rightarrow [+long] / \#C_0\_\_\,]_\sigma$

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), except for in some compound words, where lexical similarity overrides the syllabification preference. In the case of intervocalic consonant clusters, VC.CV is standard, preferred syllabification. Síptar and Törkenczy (2000) claim that Hungarian syllables do not allow onsets, 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.

For the actual implementation of this, 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).

4.7 Phonotactics and syllable structure constraints

There is a phonotactic constraint stating that round, mid vowels are long in word final position. The only exceptions are two function words: *no* ‘well [interjection]’and *ö* ‘ahh’. This phonotactic constraint is almost always marked in the orthography. Even most foreign borrowings such as *unió* indicate the proper vowel length, although words such as *euro* or *Brno* do not. 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.

(19)

$$\overset{\text{V}}{\begin{bmatrix} - & \textit{lo} \\ - & \textit{hi} \\ + & \textit{rnd} \end{bmatrix}} \rightarrow [+\text{long}] \;/\; \underline{\hspace{2em}}\,\#$$

4.8 Consonant Shortening

Underlying geminate consonants are always short when appearing as part of a consonant cluster. However the stem typically complies with this constraint, as the spelling of the stem itself usually reflects a phonetic pronunciation; instead it is through examination of compound words (20a), derived stems (20b), and loan words (20c) where the consonant shortening process is evident.

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

The rule in (21) formalizes the generalization stated above.

(21) C → [-long] / { __ C, C __ }

4.9 /l/-assimilation

The liquid /l/ assimilates to a following /r/ or /j/. The assimilation is typically only word-internal but can happen across word boundaries in fluid speech (Kenesei et al., 1998:438). Representative examples appear in (22) and a rule in (23).

(22)	tol-juk [tojjuk]	‘push-DEF.1PL’
	gól-ja [go:jja]	‘goal-POSS.3SG’
	bal-ra [barra]	‘to the left’
	el-rejt [errejt]	‘conceal’

As an L2 learner of Hungarian with careful speech, I was originally found this rule to be dubious. However the literature regards it as fairly uncontroversial.

(23) l → r / __ r
l → j / __ j

4.10 Exceptions

In order to account for words with irregular pronunciations that be described by the above phonological patterns, I have been working to compile a list of exceptions. One subpattern of exceptions in (24) is where consonants are written short but pronounced long. Apparently all cases of dzs [ʤ] that occur intervocalically or word final are long, although there are fewer that ten such instantiations of this in the language.

(24) Consonant length exceptions[12] (examples from Nádasdy and Siptár, 1989)

/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.

Another subpattern of exceptions in (25) involves back round vowels in loan words that are written short but pronounced long. The lengthening is always in an open syllable.

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

Exceptional words are the final words assigned pronunciations, and hence the rewrite rules pertaining to the exception list override any rule outputs that may have previously applied to these exceptional words.

[12] 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.

4.11 Notes on rule ordering

In this report I have been forced to be somewhat vague about some exact details, as I have used orthography, IPA, and OGOB systems to indicate pronunciations. To be clear what is actually taking place, two sample derivations for ‘freedom’ and ‘once’ are given in (26). First the forms are converted to OGOB and then the phonological constraints apply.

(26) szabadság → SabadsAg → SabatsAg → sabaCCAg
egyszer → eGSer → eTSer → eccer

From OGOB it is possible to convert to and from all the other encodings listed in Appendix.

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 that is always already clearly marked in the orthography.

5 Possible future developments to the dictionary

Certain phenomena were necessarily overlooked in the creation of the pronunciation dictionary. For each case in this section, I briefly describe both the phenomenon itself and why it was not possible or desirable to implement it. Reasons for not implementing a particular rule range from it representing the wrong dialect or register to 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. There do now appear to be open source morphological parsers for Hungarian (Trón et al., 2005), but I have not yet integrated this into the dictionary creation process.

5.1 Long vowel reduction before consonant clusters

Extra heavy syllables (i.e. >2 moras) are not well-tolerated in Hungarian except across morpheme boundaries. A long vowel in an extra heavy syllable will reduce in certain instances, and this constraint is often abbreviated *VVCC. 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 together into the onset and there is no vowel shortening. It is reported that this shortened vowel is not necessarily always short, but certainly "shorter" than a long vowel would appear in typical environments. I would assume that this shortened vowel is a true short vowel and make no allowances for a third gradation in vowel length. Judgments about the vowel reduction may vary from person to person (Rebrus, pc).

The examples in (27) give forms where shortening is permitted to take place, while shortening is not necessary in the cases in (27) because the syllable is not extra heavy due to syllabification of consonant cluster in the following syllable.

(27) Vowel reduction according to sonority

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

5.2 Rapid speech processes

There is an optional process of consonant deletion in triconsonantal clusters (Dressler and Siptar, 1989; Siptar, 1991). Because this process is optional, I elect to not implement it and I consider it a function of only rapid speech. The process elides the middle consonant of a tri-consonantal sequence, and it seems to frequently act on coronals, as in /mindnyájan/ being pronounced [minnyájan] and kezdhetjük as [keszhettyük] or [keszthettyük]. It 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, as in the optional pronunciation of [zöld] as [ződ] 'green'. This pronunciation is more common in non-standard dialects.

5.3 Non-standard spellings

I do not treat with informal spellings styles even if they do tend to reflect relaxed pronunciations. Such spellings do not appear in 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.

(28) Slang spellings reflecting reductions of unstressed syllables

Standard	Non-standard	Gloss
azt hiszem	asszem	'I think (that)'
nem tudom	nemtom	'I dunno'
valoszinuleg	valszeg	'probably'
tetszik	teccik	'I like'[13]

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

5.4 Possible future developments

Possible future developments include continuing to find data about lexical exceptions to the grapheme-phoneme correspondences I have noted in Section 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 psychological task. Data from confusion matrices indicating the likelihood the word is mistaken for another lexical item in the language would also be useful; a similar addition would be to note the number of phonetic "neighbors" a given word has in order to indicate it density in the phonetic lexicon.

At some point support for divergent dialects would be very interesting, although I admit this would require a more detailed understanding of the dialect variation; ultimately such fine phonetic detail might be more appropriately handled by lexicographers. More within the realm of consideration would be encoding certain suprasegmental information such as secondary stress placement or for syllable weight to allow for further exploration of these patterns in Hungarian. None of these additions described in this section are planned at this time, although I believe a need for this data would drive development.

6 Evaluation

If this pronunciation dictionary is to be a valuable resource to other researchers, it is quite important to be able to assess the accuracy of the pronunciations. I have identified three possible ways in which this could be accomplished.

First of all, consulting with linguists and Hungarian researchers would yield useful advice about the accuracy of the rules identified in Sections 3 and 4. In many ways this has already been done, as the output of this work is based on a body of phonetic of phonological research compiled over several decades by other researchers. I have received useful feedback by presenting this paper at two conferences and exchanging emails with various scholars. Another more informal way of examining the dictionary is to follow several sample derivations of the more complex forms to ensure the output of the rule system gives the intended result; such exploration was naturally a result of the debugging process used for creating the dictionary.

A more precise assessment could be achieved by presenting a random sample of words from the dictionary and asking a Hungarian native speaker to evaluate whether the pronunciation is correct or not. Based on the size of the sample, a confidence estimate could be inferred for the dictionary as a whole. In this way a percentage correct score could be assigned to the dictionary. However, this undertaking would require giving significant training to an informant concerning the symbols used to represent sounds, and this training would likely destroy the unbiased nature of the informant. For example, an informant would first need to become acquainted with the encoding scheme. It is unclear to me how assessment would proceed from this point, however, as the informant would now be essentially an amateur linguist, having neither the advantage of an impartial informant nor the skill training of a linguist. I also worry about the orthography influencing judgments – recall many Hungarians do believe the orthography is phonetic.

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. An open source, free speech synthesizer called Festival contains a Hungarian voice that uses Hungarian biphones as part of Mbrola (Dutoit et al., 1996). However the speech synthesizer does not really work because it 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 the present 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.

7 Applications of a pronunciation dictionary

This section briefly surveys potential applications of a pronunciation dictionary to phonological research. To be sure, there are more applications than can possibly be listed here, and it beyond the scope of this paper to address any one application in great detail.

7.1 Phonological neighborhoods and structure of the mental lexicon

Linguists and psychologists have been especially interested in identifying what constitutes a phonological neighborhood and how a phonological neighborhood is influenced by word frequency (cf. Barlow, 2000, Gruenenfelder and Pisoni, 2006, Luce, 1986, Luce and Pisoni, 1998, Metsala, 1997). String edit distance is typically used as a measure of phonological similarity, but new measurements are being proposed (cf. Kapatsinski, 2006). Because 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, it is not clear whether the conclusions drawn for English can be generalized. Hence this work would be used to address the development of an alternative resource for the Hungarian language, an agglutinative language with several unique typological properties. Due to the high amount of inflectional and derivational morphology in Hungarian, we expect lexical neighbors to be more heavily influenced by morphological considerations in Hungarian than in English. Additionally, because Hungarian words are significantly longer than English words, the notion of a phonological neighborhood may also need to be redefined.

7.2 Phonotactic learning

In a recent paper Hayes and Wilson (to appear) attempt to learn the phonotactic constraints of a language using principles of maximum entropy. Their algorithm operates by comparing phonetic forms of language to find patterns of phoneme occurrence and develop constraints based on the likelihood of these patterns. With a pronunciation dictionary, Hungarian could be studied to confirm known phonotactic patterns or discover new phonotactic constraints.

This work is related to research on information theory and phonological complexity (Goldsmith, 2002). Frequent segment combinations may be stored as units and retrieved faster than infrequent or novel segment combinations. Recent hypotheses suggest phoneme frequency could be a factor in determining the quality of an epenthetic vowel in some languages. The pronunciation dictionary could be used to confirm or deny any of these hypotheses for Hungarian.

7.3 Functional load

Despite the implicit generative view on segments in which all segments are created equal, 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 (alternatively confusability) than place of articulation or manner. It would be interesting to see what patterns could be established for Hungarian.

Another line of exploration would involve investigating the effect of morpheme frequency on the makeup of the lexicon. Frequent suffixes for Hungarian nouns are *-t* ‘accusative’ and *-k* ‘plural’. My suspicion is that the Hungarian lexicon has evolved so that nominative singular stems tend not to end in these sounds in order to avoid confusion with plural or accusative endings; words previously ending in these sounds may have been subsequently reanalyzed. 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.

7.4 Applications specific to Hungarian research

The use of this dictionary could also be to inform Hungarian-specific research and not simply cross-linguistic comparisons. A distribution, 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. The user of the pronunciation dictionary must be aware of how it was created – because assumptions about vowel length and assimilation were programmed into the dictionary based on linguistic research, subsequent researchers must be careful to avoid circularities in reasoning by drawing conclusions from these assumptions.

8 Conclusion

In summary, this report has detailed the relevant considerations used to create a pronunciation dictionary for Hungarian. Using a relatively small number of rewrite rules, a pronunciation dictionary was generated that is more representative of actual spoken Hungarian than Hungarian orthography. I hope that this resource can be put to use in some of the applications described in Section 7.

9 References

Baayen, R.H., Piepenbrock, R., and Gulikers, L. 1996. CELEX2: Linguistic Data Consortium, Philadelphia.

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

Benkõ, Loránd, and Imre, Samu. 1972. *The Hungarian Language*. Budapest: Mouton, Akadémiai Kiadó.

Bosch, Antal van Den, and Daelemans, Walter. 1993. Data-oriented methods for grapheme-to-phoneme conversion. Paper presented at *European Chapter of ACL*, Utrecht.

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

Deme, László. 1950. Kiejtésünk néhány kérdésről [A few questions on Hungarian pronunciation]. *Magyar Nyelv* 46.

Dutoit, T., Pagel, V., Pierret, F., Bataille, O., and van der Vrecken, O. 1996. The MBROLA project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. Paper presented at *ICSLP96*, Philadelphia.

Fekete, László. 1995. *Magyar Kiejtési Szótár [Hungarian pronunciation dictionary]*. Budapest: Gondolat.

Goldsmith, John. 2002. Probabilistic models of grammar: Phonology as information minimization. *Phonological Studies* 5:21-46.

Grimes, Stephen. 2007. Word final consonant extrametricality in Hungarian: Indiana University ms.

Gruenenfelder, T., and Pisoni, D. B. 2006. Modeling the Mental Lexicon as a Complex System: Some Preliminary Results Using Graph Theoretic Measures. In *Speech Research Laboratory Progress Report*: Indiana University.

Halácsy, Péter, Kornai, András, Németh, László, Rung, András, Szakadát, István, and Trón, Viktor 2004. Creating open language resources for Hungarian Paper presented at *4th International Conference on Language Resources and Evaluation (LREC2004)*.

Hayes, Bruce, and Wilson, Colin. to appear. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*.

Kapatsinski, V. M. 2006. Sound similarity relations in the mental lexicon: Modeling the lexicon as a complex network. In *Speech Research Lab Progress Report*, 133-152: Indiana University.

Kassai, Ilona. 1989. On vowel length variability in Hungarian. Paper presented at *Speech Research '89*, Budapest.

Kenesei, István, Vago, Robert M., and Fenyvesi, Anna. 1998. *Hungarian*: Descriptive Grammars. New York: Routledge.

Keresztes, László. 1992. *A practical Hungarian grammar*. Debrecen: Debreceni Nyári Egyetem.

Kiss, Jenő. 2001. *Magyar dialektológia [Hungarian dialectology]*. Budapest: Osiris Kiadó.

Kontra, Miklós. 1995. On current research into spoken Hungarian. *International Journal of the Society of Language* 111:9-20.

Kornai, Ándrás. 1986. Szótári adatbázis az akadémiai nagyszámitógépen [A dictionary database of Hungarian]. In *Working Papers*, 65-79. Budapest: Hungarian Academy of Sciences Institute of Linguistics

Kornai, Ándrás. 1990. The sonority hierarchy in Hungarian. *Nyelvtudományi Közlemények* 91:139-146.

Luce, Paul A. 1986. Neighborhoods of words in the mental lexicon In *Research on Speech Perception*. Bloomington, IN: Speech Research Laboratory, Indiana University.

Luce, Paul. A., and Pisoni, David. B. 1998. Recognizing spoken words: the neighborhood activation model. *Ear and Hearing* 19:1-36.

Metsala, J. L. 1997. An examination of word frequency and neighbourhood density in the development of spoken-word recognition. *Memory Cognition* 25.

Nádasdy, Ádám. 1989a. The exact domain of consonant degemination in Hungarian [June 1-3, 1989]. *Proceedings of the Speech Research '89 International Conference [Hungarian Papers in Phonetics]* 20:104-107.

Nádasdy, Ádám. 1989b. Consonant length in recent borrowings into Hungarian. *Acta Linguistica Hungarica* 39:195-213.

Nádasdy, Ádám, and Siptár, Péter. 1989. Issues in Hungarian phonology: Preliminary queries to a new project. *Acta Linguistica Hungarica* 39.

Nádasdy, Ádám, and Siptár, Péter. 1998. Vowel length in present-day Hungarian. *The Even Yearbook* 3:149-172.

Nusbaum, H.C., Pisoni, D.B., and 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.

Papp, Ferenc. 1969. *A Magyar Nyelv Szóvégmutató Szótára [Reverse-Alphabetized Dictionary of the Hungarian Language]*. Budapest: Akadémiai Kiadó.

Pintzuk, Susan, Kontra, Miklós, Sándor, Klára, and Borbély, Anna. 1995. The effect of the typewriter on Hungarian reading style. *Working Papers in Hungarian Sociolinguistics*.

Rot, Sándor. 1994. *Hungarian. Its Origins and Originality*. Budapest: Korona Publishing House.

Siptár, Péter, and Törkenczy, Miklos. 2000. *The phonology of Hungarian*. Oxford: Oxford University Press.

Szemere, Gyula. 1987. *Hogy is írjuk? [How should we write?]*. Budapest: Gondolat.

Szigetvári, Péter. 2001. Dismantling syllable structure. *Acta Linguistica Hungarica* 48:155-181.

Törkenczy, Miklós. 1994. A szótag [The syllable]. In *Strukturális magyar nyelvtan, 2. kötet: Fonológia [Structural Hungarian Grammar volume 2: Phonology]*, ed. Ferenc Kiefer, 272-392. Budapest: Akadémiai Kiadó.

Törkenczy, Miklós, and Siptár, Péter. 1999. Hungarian syllable structure: Arguments for/against complex constituents. In *The Syllable: Views and Facts*, eds. Harry van der Hulst and Nancy A. Ritter, 249-284. Berlin: Mouton de Gruyter.

Tótfalusi, István 2006. *Kiejtési szótár: Idegen nevek, szavak helyes kiejtése [Pronunciation dictionary: The correct pronunciation of foreign names and words]*. Budapest: Tinta Kiadó.

Trón, Viktor, Németh, László, Halácsy, Péter, Kornai, András, Gyepesi, György, and Varga, Dániel. 2005. Hunmorph: open source word analysis. Paper presented at *ACL*.

Vago, Robert M. 1980. *The Sound Pattern of Hungarian*. Washington, D.C.: Georgetown University Press.

Vago, Robert M. 1992. The root analysis of geminates in the moraic phonology of Hungarian. In *Approaches to Hungarian*, eds. István Kenesei and Csaba Pléh. Szeged: JATE.

Váradi, Tamás. 2002. The Hungarian National Corpus. Paper presented at *Third International Conference on Language Resources and Evaluation*, Las Palmas, Spain.

Appendix

ORTHOGRAPHY	IPA	OBOG	SAMPA	PROSZÉKY
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