On the creation of a pronunciation dictionary for Hungarian

Recent research on the phonological structure of the mental lexicon has almost exclusively been based on the English 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, 2005; cf. 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, in press). However, 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 presentation addresses 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 sound similarity to be more heavily influenced by morphology in Hungarian than in English. Additionally, because Hungarian words are significantly longer than English words, new definitions for what constitutes a phonological neighborhood may also need to be defined.

The pronunciation dictionary of Hungarian under consideration here is based on the Hoosier Mental Lexicon developed in the Psychology Department at Indiana University (Nusbaum et al., 1984). The target is to have a text file with columns representing orthography, pronunciation, and corpus frequency for each word (the Hoosier Mental Lexicon additionally has data on word familiarity ratings). The initial input was a word list of orthographic Hungarian developed at the Research Institute for Linguistics in Budapest during the 1980's (Kornai, 1986).

In creating a pronunciation dictionary, there were several phonological, morphological, and historical factors to consider. Standards for spelling in modern Hungarian (called helyesírás) were developed and standardized in the late 19th and early 20th centuries (Benkõ and Imre, 1972), and as a result the output of many morphophonological processes are reflected in the orthography. In fact, Hungarian linguists are constantly reminding native Hungarian speakers that the Hungarian alphabet is in fact not phonetic. In this research, several sources were used to determine standards for the Budapest dialect described (Deme, 1950; Kassai, 1989; Kontra, 1995; Nádasdy, 1989a; Nádasdy, 1989b; Nádasy and Síptár, 1998; Pintzuk et al., 1995). Deviations of pronunciation from orthography that remained to be accounted for were historical spelling variants in (1), segment degemination in superheavy syllables depending on sonority sequencing principles in (2), consonant cluster voicing assimilation in (3), variable high vowel lengthening in (4), final consonant lengthening in monosyllabic words in (5), and the use of digraphs and trigraphs to represent single sounds digraphs and trigraphs in (6). The correspondence between

orthography and pronunciation in many ways resembles the correspondence linguists assert between underlying and surface forms. In a sense, creating a pronunciation dictionary for Hungarian is parallel to implementing a generative phonology rule ordered system in which the output of one rule serves as the input for the next in a successive chain of alterations, as was developed by Vago in "The Sound Pattern of Hungarian" (Vago, 1980). With the exception of a few words in which pronunciation cannot be reliably deduced from orthography, I conclude that it is possible to map orthographic form to phonetic form.

(1)

Historical	Modern
ly, j	[j]
ts, cs	[tš]

(2) The segment sequence VVCC is not permitted within a Hungarian syllable, and in such sequences the long vowel (indicated by VV) reduces in length. However, this is dependent on the sonority of the consonants involved. If the sonority is falling, but consonants are syllabified in the syllable and the vowel must reduce:

öörs	[örs]	'sentry'
gyüüjt	[gyüjt]	'collect'

But in cases of rising sonority, the second consonant serves as the onset to the following syllable and there is no vowel reduction.

ródli	[ródli]	'sled'
csúzli	[csúzli]	'catapult'

(3) Anticipatory (regressive) consonant cluster voicing assimilation

abszolút	[apsolút]	'absolute'
joghurt	[jokhurt]	'jogurt'

(4) One example of high vowel lengthening occurs only in content words, not in function words such as personal pronouns.

áru	[árú]	'goods'
menü	[menüü]	'menu'
ti	[ti]	'you (pl).'

(5) Word-final consonant lengthening in monosyllabic syllables

egy [ed^{y}] ~ [ed^{y}d^{y}]	'one'
nagy [nad^{y}] ~ [nad^{y}d^{y}]	'big'

It has been suggested that this lengthening can be attributed a the minimal word condition in Hungarian (Grimes, 2005).

(6) The digraphs sz [s] and zs [ž], along with monograph s [š] and z [z] create orthographic sequences that are phonetically ambiguous.

egészség [egéšség] ~ [egéššég] 'health'

The rule rewrite system was implemented in Perl using that language's regular expression functionality. Just as in a generative phonology rule system, care was taken to correctly order rules, monitoring feeding and bleeding relationships. It is the hope that development of a pronunciation dictionary for Hungarian will encourage development of similar dictionaries for other languages. With the prevalence of speech recognition systems, pronunciation dictionaries are often part of the acoustic model of such systems, but these are often proprietary and not available to the public. Pronunciation dictionaries allow for the quantitative investigation of syllable structure,

phonotactics, and cluster frequencies; development of new types of corpora may additionally spawn unforeseen areas of linguistic and cognitive research.

References

BARLOW, JESSICA A. 2000. A preliminary typology of word-initial clusters with an explanation for asymmetries in acquisition. Papers in Experimental and Theoretical Linguistics: Proceedings of the Workshop on the Lexicon in Phonetics and Phonology, ed. by 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.

DEME, LÁSZLÓ. 1950. Kiejtésünk néhány kérdésrol [A few questions on Hungarian pronunciation]. Magyar Nyelv 46.

GRIMES, STEPHEN. 2005. Moraic weight, extraprosodic word-final consonants, and morphophonological length alterations in Hungarian. International Conference on the Structure of Hungarian. Veszprém

GRUENENFELDER, T. and PISONI, D. B. 2005. Modeling the mental lexicon as a complex graph: Indiana University

KAPATSINSKI, V. M. in press. Phonological similarity relations: Network organization of the mental lexicon. Paper presented at VIII Encuentro Internacional de Linguistica en el Noroeste.

KASSAI, ILONA. 1989. On vowel length variability in Hungarian. Paper presented at Speech Research '89, Budapest.

KONTRA, MIKLOS. 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]. Working Papers. 65-79. Budapest: Hungarian Academy of Sciences Institute of Linguistics

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. and PISONI, D. B. 1998. Recognizing spoken words: the neighbourhood activation model. Ear and Hearing, 19.

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. Consonant length in recent borrowings into Hungarian. Acta Linguistica Hungarica, 39.195-213.

—. 1989b. The exact domain of consonant degemination in Hungarian. Paper presented at Speech Research '89, Budapest.

NÁDASY, ÁDÁM and SÍPTÁR, PÉTER. 1998. Vowel length in present-day Hungarian. The Even Yearbook, 3.149-72.

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, 10.357-76.
PINTZUK, SUSAN, KONTRA, MIKLÓS, SÁNDOR, KLÁRA and BORBÉLY, ANNA. 1995. The Effect of the Typewriter on Hungarian Reading Style. Working Papers in Hungarian Sociolinguistics, 1.
VAGO, ROBERT M. 1980. The Sound Pattern of Hungarian. Washington, D.C.: Georgetown University Press.