Analogy by frequency: a possible reason for the vowel length neutralisation process?

Katalin Mády & Uwe D. Reichel

1 The Hungarian vowel system

Hungarian has traditionally been described as a language with a phonological quantity distinction for both vowels and consonants. While the relevance of this feature for consonants has been questioned by Siptár (1995) based on the very small number of minimal pairs and their restricted distribution, the quantity distinction for vowels is unquestionably a relevant feature according to both phonetic and phonological descriptions. The vowels *i-í ü-ű u-ú o-ó ö-ő*, where the accent sign marks long quantity, not stress, correspond to the vowel qualities /i-iː y-yː u-uː o-oː ø-øː/ and are accounted for as five pairs distinguished primarily by their length. The remaining vowels, *a-á* and *e-é*, are classified in a different manner in form-oriented (phonetic) and function-oriented (phonological) frameworks. Phonetic descriptions take into account that these vowels do not only differ by quantity, but also by their quality, the latter being the primary cue. Short *a* is sometimes described as a back, mid-low, rounded vowel /ɔ/, while long *á* is produced as central, low, and unrounded /aː/. In Mády (2008) it was argued that vowel height is not necessarily distinctive between *a* and *á*, since the somewhat smaller jaw opening of *a* can be explained by the fact that the vowel is rounded: Hungarian /y/ was also produced with a smaller jaw opening than its unrounded equivalent /i/ in this articulography study. Nevertheless, the *a* vowels /ɒ/ and /aː/ are obviously distinguished by at least one quality feature beside quantity. The same is true for *e* and *é*, of which the short vowel is a mid-low front unrounded /ɛ/, and the long one is a mid-high front unrounded /eː/. The vowel system of Standard Hungarian does not contain any other vowels such as reduced ones or diphthongs.

Phonologists, on the other hand, encounter phonological processes in which /ɒ/ alternates with /aː/ and /ɛ/ with /eː/ in exactly the same way as short and long mid and high vowels do. One such rule is **Final Stem Vowel Shortening**: in certain stems, a long vowel is replaced by its short counterpart if a given suffix is added to the stem, e.g. ***kút*** /kuːt/ 'well' – ***kutak*** /kutɒk/ 'wells', ***kéz*** /keːz/ 'hand' – ***kezel*** /kɛzɛl/ 'to handle', ***sár*** /ʃaːr/ 'mud' – ***sarat*** /ʃɒrɒt/ 'mud + acc.)'. Another example is **Internal Stem Vowel Shortening** that is triggered by certain suffixes such as *-ál*, *-ikus* etc., e.g. ***aktív*** /ɒktiːv/ 'active' – ***aktivál*** /ɒktivaːl/ 'to activate', ***kultúra*** /kultuːrɒ/ 'culture' – ***kulturális*** /kulturaːliʃ/ 'cultural'. (See Siptár & Törkenczy 2000: 58–62 for a detailed description of these rules.) In this paper we will accept the assumption that the vowels /ɒ/-/aː/ and /ɛ/-/eː/ are vowel pairs with quantity as a distinctive feature, regardless of their different qualities, for reasons to be explained below.

Vowel length in Hungarian is encoded by orthography: long vowels are marked by an accent, short vowels by the absence of it (e.g. *ó-o*, *ő-ö* for /o/ and /ø/, the umlaut marking front rounded vowels). This has two consequences: first, native Hungarian speakers are aware of vowel quantity, unlike e.g. German or Swedish speakers, second, the pronunciation norm is conserved by orthography to some extent.

Given this, it might appear surprising that quantity is not consequently realised in the way orthography would suggest. For example, the compound word ***kó**rház* 'hospital' is often produced as /k**o**rhaːz/, i.e. with a short /o/, in Educated Colloquial Hungarian (a variety widely accepted all over the country), although the lexemes *kór* 'disease' and *ház* 'house' are both pronounced with long vowels. This discrepancy between orthography and the usual pronunciation of a word does not involve mid vowels very often, but it is frequent for the high vowels *u*, *y*, and *i*. While there are some examples where an orthographically short vowel is produced as a long one in colloquial speech (such as ***di**csér* 'to praise' that is pronounced as /di**ː**tʃeːr/ by many speakers), the vast majority of the discrepancies involves cases in which long graphemes are produced as short vowels (e.g. *címke* /tsimkɛ/ 'label'). The shortening tendency is even more advanced in unstressed syllables. Siptár & Törkenczy (2000) claim that the quantity distinction for high vowels is missing completely in word-final position, at least in Educated Colloquial Hungarian.

In this paper, the relevance of the distribution of short and long vowels is investigated, with a special focus on syllables carrying more or less prominence. Hungarian has fixed lexical stress that is always word-initial. Syllables with lexical stress are potential carriers of sentence-level pitch accents. Thus, the distribution of short and long vowels in these syllables might have a different impact on the preservation of the vowel quantity distinction.

First, a short diachronic overview of quantity variation in dialects is given. Subsequently, frequency of short and long vowels in a type and a token word list is analysed. Furthermore, we investigate a potential interplay between the functional load of a quantity opposition and its preservation.

2 Dialectal variation

There is considerable variation in the distribution of high short and long vowels accross the regional varieties of Hungarian. In large parts of Western Hungary, the vowel system does not contain long high vowels at all (Kálmán 1989). On the other hand, a prevalence of long high vowels can be observed in the Eastern Hungarian dialects, in which many short vowels of the standard variety are lengthened, especially in stressed (i.e. word-initial) syllables. According to Benkő (1957), short vowel lengthening in the Eastern dialects took place from the 16th century on, and the development to shorten long vowels in the Western regions is at least as this old. He explains the unstability of these vowels by the process of the Final Stem Vowel Shortening (see above): while in the Eastern dialects the shortening rule often failed to apply to suffixed stems and resulted in a higher number of long vowels, the Western region shortened vowels in unsuffixed stems analogously to their suffixed forms as opposed to Central Hungarian dialects on which today's standard is based. (E.g. Eastern ***ú**t* 'road', ***ú**tazik* 'to travel', as opposed to Central ***u**tazik* 'to travel, Western *v**i**z* 'water', *vizes* 'wet' as opposed to Central *víz*.)

Another change in vowel quantity was the shortening of word-final long /uː/ and /yː/ that included large regions both in West and East (in the West it also applied to /iː/). Since word-final vowels in polysyllabic words are always unstressed in Hungarian, Benkő suggests that the shortening process is due to the missing prominence in these syllables. Although the same process did not take place in Central Hungary, it is remarkable that vowel shortening in unstressed syllables has become part of today's Colloquial Educated Hungarian that is also spoken in the Central Hungarian capital Budapest.

According to Benkő, the quantity change in stressed syllables was triggered by the coexistence of the stems with long and short word-final vowels in their unsuffixed and suffixed forms. However, as described above, this rule is not restricted to high vowels, on the contrary, it mostly applies to stems with low vowels (for an exhaustive list, see Siptár 2003: 311). On the other hand, the few stems that include word-final mid vowels are special: in course of the shortening process, a /v/ is inserted after the vowel, e.g. ***ló*** 'hourse' vs. ***lovak*** 'hourses'. Thus, the systematic shortening of stem vowels alone cannot account for the variable quantity of high vowels.

3 Distribution of short and long vowels

3.1 Frequency in types

Another potential reason for the shortening of long high vowels could be their lower frequency in the language. According to Gósy (2004): 85ff., long vowels are far less frequent in Hungarian, than short ones. In her statistics, the proportion of long vowels was given with 21% among all vowels, long yː being the most unfrequent one.

A further question is whether the distribution of long and short vowels is identical across word-initial (= stressed and potentially accented) and non-initial syllables. A lower occurrence of long vowels in non-initial syllables could explain the shortening tendency based on analogy.

Type analysis was performed on a word list consisting of 29 245 lemmas based on the lexical entries in the *Magyar értelmező kéziszótár*. The list word stems (such as *asztal* 'table') and derived forms (such as *asztalos* 'carpenter'), but no words with inflectional suffixes such as plural forms, since the latter were not given as lexicon entries.

Frequency counts for vowels are given in Table 1. Since there is no general agreement on the corresponding IPA symbols in the literature, orthographic symbols are used.

Figures 1 and 2 show the frequency of short and long vowels in the dataset, consisting of 85 198 vowels altogether, 29 212 of which occurred in the first syllable of the word (34%). The amount of short vowels was 62 776 (74%).

Since the proportion of non-initial syllable positions was substantially higher, the ratios given in the figures were calculated based on the overall count of vowels in word-initial vs. non-initial syllables: e.g. the absolute frequency of short /o/

Table 1: Frequency of vowels in lexicon entries in the Magyar értelmező Kéziszótár.

vowel	word-inital syllable	non-initial syllable	sum
a	5804	10130	15934
á	2121	6387	8508
e	6016	9990	16006
é	1823	3554	5377
i	3926	8420	12346
í	532	1188	1720
o	3673	6766	10439
ó	686	2938	3624
ö	1295	1143	2438
ő	339	1382	1721
u	1749	2742	4491
ú	417	484	901
ü	610	512	1122
ű	221	350	571

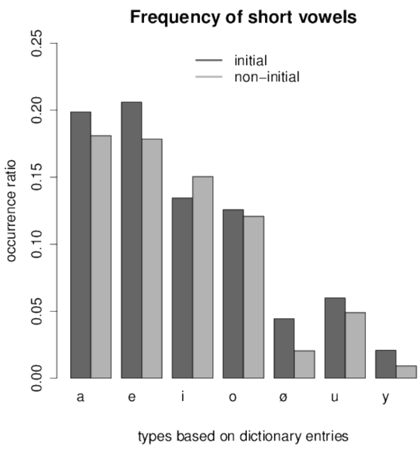

Figure 1: Frequency of short vowels in lemmata. Dark grey: vowels in word-initial syllables, light grey: vowels in non-initial syllables.

in the word-initial syllable was divided by the number of all vowels in the same position.

The amount of short vowels in word-initial syllable position was higher than in non-initial syllables, except for /i/. At the same time, long vowels occurred more often in non-initial syllables in most vowel pairs. A possible reason for this could be the fact that certain derivational suffixes with high frequency contain long vowels, such as *-ás/-és*, *-ság/-ség*.

Interestingly, vowel frequency does pattern along with the three categories high, mid, and low. Short /ɛ/ is the most frequent vowel closely followed by

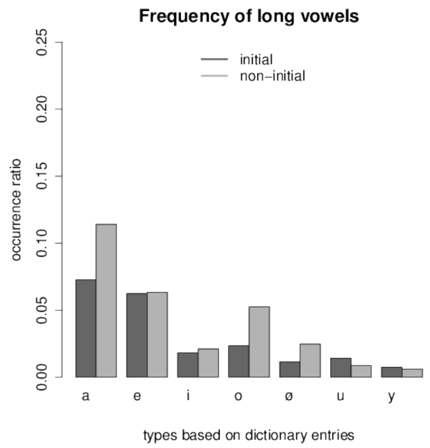

Figure 2: Frequency of long vowels in lemmata. Dark grey: vowels in word-initial syllables, light grey: vowels in non-initial syllables.

short /ɒ/, and also their long equivalents are more frequent than the other five vowels. The low frequency of the other high vowels /u y/ is in line with the analogy hypothesis. On the other hand, the frequency of long /iː/ is almost identical with that of long /øː/, while only /i/ is involved in the shortening process described in the literature. Thus, the tendency observed in the type-based word list does not support the analogy hypothesis.

3.2 Frequency in tokens

Type frequency by itself is not necessarily informative on the actual occurrence of long vowels in spoken language for various reasons. First, token frequency is not taken into account. Second, inflexional suffixes that do not occur in the list of lexical entries discussed above contain more often short vowels in Hungarian, than long vowels. Third, the frequency of certain types in colloquial speech can substantially differ from the word list discussed in the previous section.

For this reason, spoken language data from a maptask corpus was used to analyse the distribution of short and long vowels in spontaneous speech. The corpus contains data from 27 speakers between 18 and 63 years, including 13 female and 14 male speakers. Dialectal background and social status of speakers differed (see Mády 2010*a* for more detail).

The database was created for the acoustic analysis of short and long vowels in identical consonantal environments. On the map of the first speaker, a path objects is marked. The first speaker is supposed to guide the second speaker along this path that is not visible for the second speaker. However, as can bee seen in Figure 3, the two maps were not completely identical. The maps used for this specific task are shown in Figure 3.

The overall length of the speech material was 115 minutes. Recordings were transcribed into their orthographic form, i.e. according to the grapheme system of Hungarian (not IPA or other phonetic symbols). The material was segmented into word forms. Unfinished forms due to interruption were removed from the

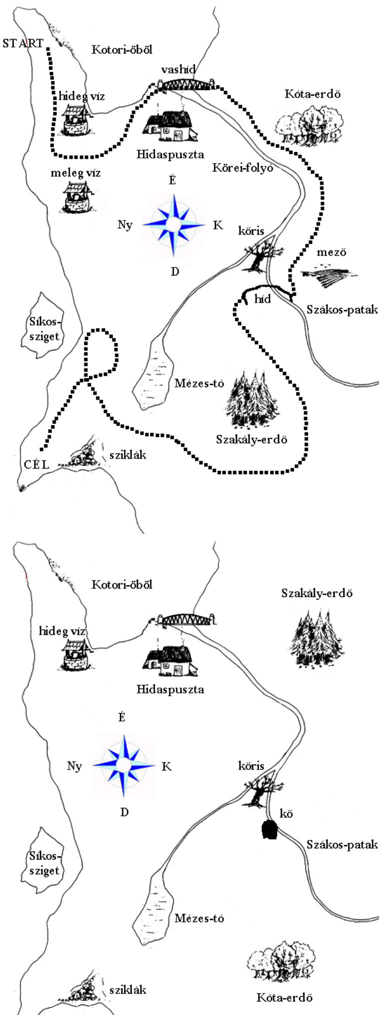

Figure 3: *Maps used for the task. Top: first speaker's map, bottom: second speaker's map.*

list. Since target words were partly invented geographical names such as *Szákos-patak* referring to a stream that are not part of everyday language usage, proper names were not taken into account for further analysis.

One of the questions of this paper is whether the frequency of short and long vowels located in potentially prominent (i.e. pitch-accented) syllables is identical. However, not every word can carry a pitch accent. Therefore definit and indefinit articles such as *a* 'the', non-accentable conjunctions such as *és, ha* 'and', 'if' and modal particles such as *hát* 'well' were excluded from the analysis. Admittedly, this procedure is blind for the presence of pitch accents on these words. For example, the indefinit article *egy* 'a' is homophonous with the numeral *egy* 'one', and the latter usually carries a pitch accent. Since manual checking of the accent patterns was not possible in this case, all tokens of this type were disregarded. Other words that can function both as a content or a function word, e.g. *fog* 'grab' or a future auxilary, were regarded as potential prominence carrier units and were included in the analysis.

The dataset contained 17 916 vowels, 9024 were located in word-initial syllables (50%), and 13 475 were short ones (75%). The high proportion of vowels in word-initial syllables is due to the overall high occurrence of monosyllabic words such as verbal prefixes.

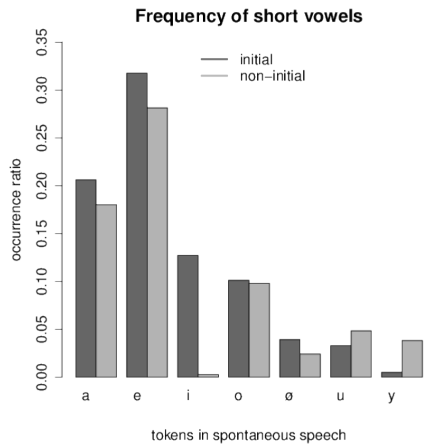

Figure 4: Frequency of short vowels in spontaneous speech. Dark grey: vowels in word-initial syllables, light grey: vowels in non-initial syllables.

The distribution of short vowels is partly different from frequencies in the word list. Here, ɛ is by large the most frequent vowel. This is in line with the wide-spread assumption that this sound is the most frequent one in Hungarian, but differs from frequency data based on the lexicon entries where /ɒ/ was only slightly less frequent than /ɛ/. It is interesting that /i/ is extremely unfrequent in non-initial syllables.

The relative frequencies of long vowels show a similar pattern to type frequencies, with the exception of /uː/.

The distributional data do not favour the hypothesis that the small number of long high vowels could be responsible for the preference for short high vowels. First, /iː/ is not less frequent than /øː/. Second, not only high, but all long

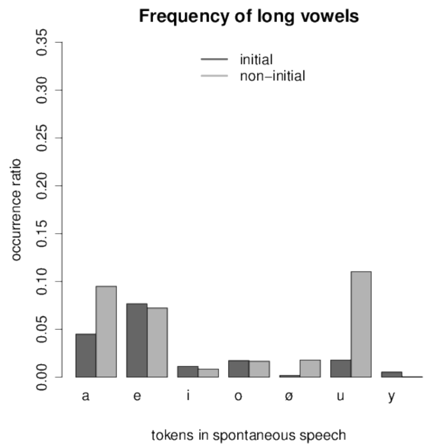

Figure 5: Frequency of long vowels in spontaneous speech. Dark grey: vowels in word-initial syllables, light grey: vowels in non-initial syllables.

vowels are less frequent both in stressed and unstressed position than short ones.

4 Functional load of quantity oppositions

Next to the vowel frequency analyses we investigated whether the functional load of a quantity opposition could be accounted for its preservation. We hypothesise that oppositions with a high load are more stable than oppositions for which the load is low. The importance of a quantity opposition is quantified by two measures described in the following.

4.1 Functional load

The functional load FL of a phonological opposition of the phonemes a and b is related to the number of contrasts this opposition is responsible for in a language L. The information theoretic definition adopted here was first introduced in (Hocket 1967):

$$\mathrm{FL}(a,b) \quad = \quad \frac{H(L) - H(L_{a=b})}{H(L)}$$

$H(L)$ is the entropy of a language L. $L_{a=b}$ denotes a language lacking an opposition of a and b. $\mathrm{FL}(a,b)$ thus stands for the relative amount of information loss resulting from such a merging, reflecting the increase of homophones.

L and $L_{a=b}$ are the sets of word types w of the *Magyar értelmező kéziszótár* before and after vowel merging, respectively, which was done separately for each vowel quantity pair. From the word type frequencies contained in this dictionary maximum likelihood probabilities were calculated in order to derive the entropies for L and $L_{a=b}$ as follows:

$$H(L) = -\sum_{w \in L} p(w) \log_2 p(w)$$

$$H(L_{a=b}) = \sum_{w \in L_{a=b}} p(w) \log_2 p(w)$$

The frequencies for merged types in $L_{a=b}$ were simply obtained by summing up the frequencies of all types of L undergoing this merging after the loss of the opposition of a and b.

4.2 Type number ratio

Since the FL is calculated over the entire lexicon it does not normalise for vowel-related frequencies, that in turn determine the number of resulting homophones after quantity merging. FL is positively correlated with vowel frequency, since the merging of frequent vowels results in more homophones so that their quantity opposition receives a high functional load.

In order to reduce this frequency bias, we additionally calculated the ratio of those types only, that are affected by a vowel quantity merging. As an example, for the merging of /uː/ we considered only those words that contain the letter 'u' and/or 'ú'. For these types we derived the ratio N_a/N_b, where N_b denotes the number of types before merging and N_a denotes the number of types after merging.

4.3 Word stress

In order to test the impact of word stress on quantity merging, we applied the two measures for three different vowel merging scenarios: (1) in all syllables, (2) in the word-stressed (initial) syllable only, and (3) in all non-stressed (non-initial) syllables only.

4.4 Results

The functional loads and type number ratios for each vowel pairing are shown in Figures 6 and 7, respectively. Important oppositions are indicated by a high functional load and a low type number ratio.

Based on Figure 6 for the impact of functional load on quantity preservation the following conclusions can be drawn:

- The quantity oppositions for /e, a/ have the highest functional loads, which – in line with our hypothesis – prevents them to undergo quantity merging. This merging would result in a significant decrease in lexical contrasts and therefore an increase in ambiguity.
- Over the entire word /i, u, y/ quantity oppositions have the lowest functional loads. Thus it is not this crucial to maintain these oppositions, and indeed, these vowels are least stable in preserving them.

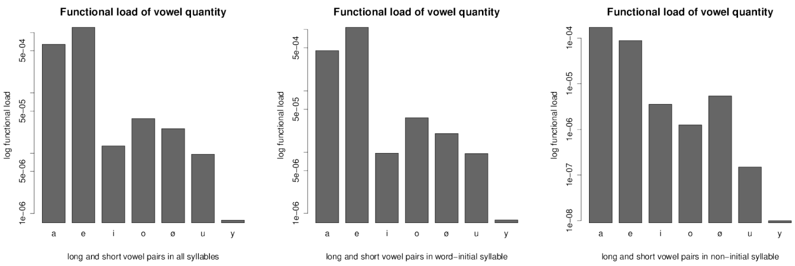

Figure 6: Functional loads of vowel quantity oppositions over entire word (left), in word-initial stressed position (mid), and in non-initial unstressed position (right).

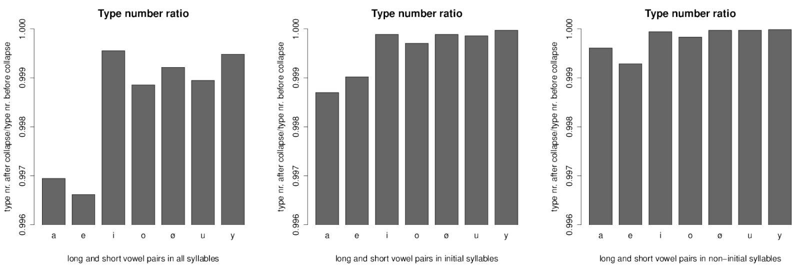

Figure 7: Type number ratios of vowel quantity oppositions over entire word (left), in word-initial stressed position (mid), and in non-initial unstressed position (right).

- Also word-initially the functional loads of /i, u, y/ quantity oppositions are the lowest, so that the absence of these oppositions e.g. in Western Hungarian is not harmful.
- However, in non-initial syllables also the quantity opposition for /ø/ has a low functional load, but against the expectation for this vowel the quantity opposition is maintained.

For the type number ratios shown in Figure 7 we obtained the same tendencies.

- For /a, e/ the lowest ratios have been measured, again well explaining the stability of their quantity contrasts in order not to drastically increase the number of homophones.
- /i, u, y/ show high ratios, especially in non-initial syllables, indicating only a negligible increase of ambiguity in case of quantity merging.
- However, high ratios are given also for /o, ø/.

5 Discussion and conclusions

Vowel statistics did not turn out to play a crucial role in explaining varying degrees of stability of quantity oppositions. The functional load of an opposition, however, has been found to have an impact on maintaining quantity contrasts. A high functional load is a sufficient motivation to maintain such contrasts. The reverse, that a low functional load leads to quantity merging, holds for high, but not yet for mid vowels.

The quantity opposition in mid vowels might be subject to an ongoing sound change process at present. Perception experiments in Mády (2010*b*) and Mády (2012) show that the perceptual boundary between long and short /o/ in word-final position is shifted towards the short vowel in young participants compared to the older group. They heard both shorter and more centralised /o/ segments as long vowels, whereas a segment had to be longer and more centralised to be idenfified as a long /oː/ by listeners above 50 years. Thus, the quantity distinction might become less stable also for mid vowels at some point. This development would again be well explainable by the low functional load of mid-vowel quantity oppositions.

6 Acknowledgements

Data for the maptask corpus were recorded at the Laboratory of Speech Acoustics at the Technical University of Budapest. We highly appreciate the help of Klára Vicsi and her colleagues.

References

Benkő, L. (1957), *Magyar nyelvjárástörténet*, Egyetemi magyar nyelvészeti füzetek, Tankönyvkiadó Vállalat.

Gósy, M. (2004), *Fonetika, a beszéd tudománya*, Osiris Kiadó, Budapest.

Hocket, C. (1967), 'The quantification of functional load', *Word* **23**, 320–339.

Kálmán, B. (1989), *Nyelvjárásaink, 5. kiadás*, Tankönyvkiadó, Budapest.

Mády, K. (2008), 'Magyar magánhangzók vizsgálata elektromágneses artikulográffal gyors és lassú beszédben', *Beszédkutatás* pp. 52–66. [An EMA investigation of Hungarian vowels in normal and fast speech].

Mády, K. (2010*a*), 'Hungarian vowel quantity neutralisation as a potential social marker', *Acta Linguistica Hungarica* **57**(2–3), 167–188.

Mády, K. (2010*b*), Shortening of long high vowels in Hungarian: a perceptual loss?, *in* 'Proc. Sociophonetics at the crossroads of speech variation, processing and communication', Pisa, Italy, pp. 41–44.

Mády, K. (2012), Implicit and explicit language attitude in a sound change process, *in* 'Proc. 2nd Sound Change Conference', Kloster Seeon, Germany, p. 87.

Siptár, P. (1995), *A magyar mássalhangzók fonológiája*, Linguistica, Series A: Studia et Dissertationes 18, MTA Nyelvtudományi Intézet, Budapest. [The phonology of Hungarian consonants].

Siptár, P. (2003), Hangtan [phonology], *in* 'Új magyar nyelvtan [New Hungarian grammar', Osiris, Budapest.

Siptár, P. & Törkenczy, M. (2000), *The Phonology of Hungarian*, University Press, Oxford.