9-1-2012

Loss of Agreement between Hungarian Relative Pronouns and their Antecedents

Daniel Szeredi
New York University, daniel.szeredi@nyu.edu

Loss of Agreement between Hungarian Relative Pronouns and their Antecedents

Abstract

Prescriptive and descriptive grammars of Hungarian frequently discuss variation in the choice of relative pronoun in Hungarian. This paper presents quantitative data about some long standing questions raised by prescriptivists, and a phenomenon that has not been explored: the lack of agreement between the pronoun and the antecedent in Colloquial Hungarian. The study presented here is based on the Budapest Sociolinguistic Interview (Varadi 1998). This corpus consists of 50 sociolinguistic interviews, conducted in 1987 and 1988, totaling approximately 240,000 words. A database was created comprising all nominal relative pronouns from the corpus (N=1714), coded for relevant semantic, syntactic and morphological variables. Two issues, the loss of the relative pronoun *amely* for specific antecedents, and the spread of the selective relative pronoun *amelyik* was confirmed. It is also a clear pattern in the corpus, that the plural form of the most frequent relative pronoun *ami* is avoided, while number agreement is intact everywhere else in Hungarian.

Loss of Agreement between Hungarian Relative Pronouns and their Antecedents

Daniel Szeredi[*]

1 Introduction

The system of relative pronouns in Hungarian is quite complex, with several nominal, adjectival and adverbial pronouns (Kugler and Laczk 2000). This paper investigates some questions regarding the use of nominal relative pronouns using the BUSZI-2 corpus of spoken colloquial Hungarian in Budapest (Vradi 1998). Some of these questions have been raised and discussed widely in prescriptivist literature, but the widespread lack of agreement between the antecedent and the relative pronoun has not been noticed so far. This paper provides an insight on this phenomenon as well.

Section 2 summarizes the questions investigated in this paper about the use of nominal relative pronouns in Hungarian. The corpus used in this study will be presented in Section 3 and the summary of the variables used will be discussed in Section 4. Finally, variable rule analyses on the relevant variables will be presented in Section 5. The case of the loss of agreement between the antecedent and certain relative pronouns found will be investigated in written web corpora as well in Section 6.

2 Questions

There is uncertainty about the use of nominal relative pronouns in Hungarian leading to variation about the choice of pronoun, the presence of number agreement between the pronoun and its antecedent, and the conjugation of the verb. In this paper, data will be presented about the effect of the certain syntactic and semantic features on the selection of relative pronoun and about the apparent lack of agreement between the pronoun and the antecedent.

There are four nominal relative pronouns in Hungarian. The choice between them in Hungarian is determined using three criteria: whether the antecedent is animate, whether the relative clause is *selective* and whether the antecedent is specific:

(1) *aki* – used for [+animate] antecedents
(2) *amelyik* – used for [-animate] selective relative clauses
(3) *amely* – used for [-animate] non-selective relative clauses, when the antecedent is specific
(4) *ami* – used for [-animate] non-selective relative clauses, when the antecedent is not specific

When the antecedent is [+animate], Hungarian uses *aki*. In non-standard Hungarian, groups or institutions of humans can also select this pronoun (Grtsy and Kovalovszky 1980, Peth 2000). Because of the small sample size of such antecedents in this corpus, this issue cannot be investigated in this paper.

The relative pronoun *amelyik* is used when the relative clause is not only restrictive, but selective as well. This means that it brings forward an alternative set of restrictions for the antecedent and selects the one described by the relative clause. This selectivity is similar to focused constituents (compare (5) to (6) and (7) to (8)):

(5) Non-focused constituent
     Láttam   a   magas házat.
     saw-1SG the high   house-ACC
     'I saw the high house'

(6)   Non selective clause
      Láttam   a   házat,   **amely**   magas.
      saw-1SG the house-ACC non-selective rel.pr. high
      'I saw the house, which is high'

(7)   Focused constituent
      A   magas$_{FOC}$ házat   láttam
      The high   house-ACC saw-1SG
      'I saw the high$_{FOC}$ house'

(8)   Selective clause
      Láttam   a   házat,   **amelyik**   magas.
      saw-1SG the house-ACC selective rel.pr. high
      'I saw the house, which is high, and not the others.'

The following observations have been made by the prescriptivist and descriptive literature regarding the use of relative pronouns in the vernacular, which will be discussed in this paper:

(9)   The pronoun *ami* is spreading after specific antecedents in place of *amely* (Grtsy and Kovalovszky 1980, Fogarasi 1993, Ttfalusi 2004)

(10) There is a hypercorrective use of *amely* in place of *ami* as a backlash to the phenomenon above (Grtsy and Kovalovszky 1980)

(11) The pronoun *amelyik* appears in non-selective clauses in place of *ami* or *amely* (Kenesei 1992)

An issue, which has not been discussed in the literature, is the lack of agreement between the antecedent and the pronoun. In standard Hungarian, the agreement between the semantic and formal plurality of the antecedent and the relative pronoun is compulsory:

(12) találkoztam egy fiúval,   **aki**...
      I met   a   boy.SG-INSTR who...
      'I met a boy, who ...'

(13) találkoztam pár   fiúval,   **akik**...
      I met   some boy.SG-INSTR who-PL...
      'I met some boys, who ...'

(14) találkoztam fiúkkal,   **akik**...
      I met   boy-PL-INSTR who-PL...
      'I met boys, who ...'

There is no discussion about the loss of this agreement in the literature, as number agreement in other places like subject-verb agreement is intact in the vernacular and in dialects as well. While investigating the corpus, it became obvious, however, that agreement fails quite frequently, and not only after semantically plural antecedents like in (13), but after formally plural antecedents like in (14) as well.

3 Corpus

The tokens used in this study are from the BUSZI-2 corpus of the Budapest Sociolinguistic Interview (Vradi 1998), which consists of 50 sociolinguistic interviews, with approximately 240,000 words in total. These interviews were conducted in 1987 and 1988, and the exact source that was used for this paper is the transcription of the interaction between the interviewer and the subject.

The 50 subjects for these interviews were selected to represent five different social classes, who would speak different sociolects of Colloquial Hungarian in Budapest, the capital and the most populated city in Hungary. The stratification probably reflects a different view of social status: in then-socialist Eastern Europe, education and cultural role was much more important in high social status than wealth or success. Therefore high school teachers and intellectuals had (and still have)

a bigger effect on the standard and used (and still use) a much more conservative register than wealthier businesspeople or politicians. The five professions representing social classes were coded as follows:

Social class	Profession	Age range
1	high school teachers (language and literature)	53–65
2	university students	21–25
3	salesclerks	17–44
4	skilled workers	20–49
5	technical school students	15–16

Table 1: Social classes in the BUSZI-2 corpus.

As it can be seen from Table 1, a major problem with this corpus is that age groups cannot be investigated independently of social status.

The interviews are divided and marked for discussion modules of different formality. The high number of modules and the low number of tokens in each module leads to the result that the use of these modules is problematic in the scale of this paper. The modules could arbitrarily group by differing formality, but these module groupings have not turned out to be a significant factor for the problems discussed here.

4 Variables

A database has been built from every occurrence of the four nominal relative pronouns in the corpus. For every token, several variables were coded, but only those which will be analyzed in this paper are listed below.

- Speaker's social status: **1-5** as in Table 1, **6** for the interviewer

- Specificity: **1** if the antecedent is specific, **0** if it is not

- Restrictivity: **1** if the relative clause is restrictive, **0** if it is not

- Selectivity: **1** if the relative clause is selective, **0** if it is not

- Antecedent form: **p** if the antecedent is plural in form, **s** if it is singular in form

- Antecedent meaning: **p** if the antecedent is plural or collective semantically, **s** if it is singular in meaning

- Pronoun plurality: **p** if the relative pronoun has the plural marker, **s** if it does not

- Pronoun: **i** for *ami*, **y** for *amely*, **k** for amelyik, **a** for *aki*

Examples for how the plurality of the antecedent and the pronoun were marked are provided in Table 2:

Antecedent form	meaning	Pronoun form	Example				
s	s	s	*az* the	*első* first	*olyan* such	*kongresszus,* congress	*ami* that
			'the first such congress, that'				
s	s	p	*arra* that.SG-SUPL	*tudok* I can	*támaszkodni* rely on	*amiket* that-PL-ACC	*hallottam* I heard
			'I can rely on what I heard'				

Antecedent		Pronoun	Example
form	meaning	form	
s	p	s	*tehát sok van olyan ami érdekelne*
			so much is such that.SG I would be interested
			'so there are many in which I would be interested'
s	p	p	*egy csomó szót … amiket*
			a bunch of word.SG (!) … that-PL-ACC
			'a bunch of words, that'
p	s	s	N/A
p	s	p	N/A
p	p	s	*azokkal az árukkal foglalkoztam ami*
			those the good-PL I worked with that.SG
			'I only worked with goods, which'
p	p	p	*a nagy romantikus barátságok amikrl*
			the big romantic friendship-PL that-PL-DELAT
			'the big romantic friendships, about which'

Table 2: Coding plurality of antecedents and relative pronouns.

The standard prescriptive literature would only accept the *sss* and *ppp* agreement patterns. The examples above are all direct citations from the corpus itself, thus it can already be seen that non-standard agreement patterns do exist in spoken Hungarian. The high frequency of these patterns will be discussed in Section 5.4.

5 Analysis

In this section, a summary of the database obtained from the corpus will be presented, and three phenomena that involve variation will be analyzed: the choice of pronoun after specific antecedents, the choice of selective relative pronouns in non-selective environments and the lack of agreement between the pronoun and its antecedent. Each issue will be presented quantitatively and a variable rule analysis (Cedergren and Sankoff 1974) will be provided using the GoldvarbX program.

5.1 Summary of Data

There are 1714 nominal relative pronouns in the database collected from the BUSZI-2 corpus. Their distribution is not balanced, though:

Pronoun	Subjects	Interviewers	Sum
ami	574	289	863
amely	13	29	42
amelyik	57	32	89
aki	583	137	720

Table 3: Summary of tokens in the corpus.

Despite the large number of tokens, there are still very few instances of the more infrequent *amely* and *amelyik* pronouns. This means that a certain amount of restraint is probably needed for some of the analysis below.

5.2 Specificity

The hypothesis that *ami* is spreading in the place of *amely*, when the antecedent is specific, is largely verified in the corpus. The following table shows the percentage of the use of either pronoun when the antecedent is specific (and the selective *amelyik* is not used):

	amely	*ami*
Subjects	12 (4.2%)	259 (95.8%)
Interviewers	28 (15.6%)	123 (84.4%)
Sum	40 (8.7%)	382 (91.3%)

Social class	ami_{spec}	$amely_{spec}$	Percentage of *ami*
1	53	10	84%
2	64	2	97%
3	60	0	100%
4	51	0	100%
5	43	0	100%
Sum	271	12	95.8%

Table 4: Distribution of relative pronouns after specific antecedents.

A variable rule analysis could not be made quite simply because of the very small number of *amely* tokens. There were some serious knockout factors: the social classes 3, 4, and 5 did not use *amely* at all, and there was no hypercorrection anywhere in the whole corpus: there were no cases where *amely* was used after a non-specific antecedent. To handle this problem, specificity was removed as a factor, and only *ami* and *amely* pronouns after specific antecedents have been examined. To handle the lack of *amely* occurrences in the lower social classes, these classes were excluded from the analysis as well. This makes sense: the pronoun *amely* is probably not present as a productively used lexical item with these speakers.

The results are still extreme: the low number of *amely* tokens presents a situation where this pronoun is probably in the process of falling into disuse. The binomial up and down analysis of Goldvarb found the social class and the restrictivity of the clause as significant factors with the form of the pronoun as independent variable. The log likelihood of the null model was -170.2, and the log likelihood of the best fitting model was -159.281, which means that the latter one predicts the data significantly better ($\chi^2 = 21.84$, d.f. = 3, $p < 0.001$). The factor weights are as follows (application is true if the pronoun is *amely*):

(15)   input probability: 0.06
       social class: **6 (interviewer)**: 0.652, **1**: 0.552, **2**: 0.175
       restrictivity of clause: **0**: 0.351, **1**: 0.551

The overall frequency of *amely* is low indeed throughout the corpus. The older speakers (interviewers and high school teachers) used this form more often and it was used more frequently in front of restrictive clauses. The conclusion is quite clear: this relative pronoun is very close to falling out from the spoken colloquial language.

5.3 Spread of *amelyik*

Although the number of *amelyik* tokens is quite low in the corpus as well, it is worthwhile to examine if this pronoun indeed spreads to clauses which are not selective. An approach to investigating this question is to count the proportion of *amelyik* occurrences among other (non-animate) relative pronouns in non-selective restrictive clauses. The contingency table can be seen below:

Social class	ami	amely	amelyik	% of *amelyik*
1	88	6	3	3.1%
2	95	1	4	4%
3	89	0	3	3.3%
4	71	0	3	4.1%
5	73	0	5	6.4%
Interviewer	194	26	6	2.7%
Sum	610	33	24	3.6%

Table 5: Non-selective restrictive pronouns.

The first variable rule analysis was run with the independent variable as the form of the pronoun, and the application as the occurence of *amelyik*. Social class, specificity of the antecedent, restrictivity and selectivity were used as dependent variables. The result of Goldvarb shows that the lower social status leads to preference of this pronoun (and that restrictivity is not a significant factor):

(16)   input probability: 0.024
       social class: **6 (interviewer)**: 0.406, **1**: 0.483, **2**: 0.403, **3**: 0.508, **4**: 0.698, **5**: 0.729
       specificity: **1**: 0.781, **0**: 0.167
       selectivity of clause: **1**: 0.935, **0**: 0.421

In the next step, it was tested what happens if non-selective cases are excluded to see if lower classes really extend the use of *amelyik* in this direction. It is not clear that this is the case. The Goldvarb analysis does not even choose to include social class as a relevant factor here ($\chi^2 = 4.224$, d.f. $= 5$, $p = 0.52$), but if it is included in the analysis nevertheless, it does show the effect seen above, although not that clearly:

(17)   input probability: 0.012
       social class: **6 (interviewer)**: 0.446, **1**: 0.481, **2**: 0.479, **3**: 0.428, **4**: 0.631, **5**: 0.67
       specificity: **1**: 0.781, **0**: 0.167

The other approach is to check the proportion of non-selective occurrences of *amelyik* among all tokens of this pronoun. In the variable rule analysis this approach means that selectivity is the independent variable (with non-selective cases as application) and social class, restrictivity and specificity of the antecedent are the dependent variables. Of these, non-restrictivity entails non-selectivity, so there is no application in non-restrictive cases, and non-specific antecedents cannot act as a basis of the selection. These interactions lead to the exclusion of these factors as well, so only social class remains as dependent variable. The resulting model is not significantly better than the null model ($\chi^2 = 4.592$, d.f. $= 5$, $p = 0.47$), but the values show that the non-selective use of *amelyik* is more frequent for **higher** social classes:

(18)   input probability: 0.356
       social class: **6 (interviewer)**: 0.415, **1**: 0.707, **2**: 0.783, **3**: 0.475, **4**: 0.497, **5**: 0.475

The final conclusion of these two approaches is that lower classes use more selective relative clauses, and use *amelyik* in the place of *ami* or *amely* more probably than speakers with higher social status, but the non-selective use of this pronoun is not stigmatized at all and higher social classes even use this construction more frequently.

5.4 Agreement

The following table shows that the plurals of *ami* and *amelyik* are very infrequent:

Pronoun	sing	plur	plur%
ami	827	36	4.17%
amely	16	26	61.9%
amelyik	89	0	0%
aki	486	234	32.5%

Table 6: Distribution of the plural form of relative pronouns.

The lack of plural forms for *amelyik* is not surprising as its plural **amelyikek* is not grammatical in the standard language. The plural of *ami* is present in every register, though, and no claim has been made so far that it would be a form which is avoided in the language. The data from this corpus give strong evidence for this, however, and the speakers of higher social status fail to have the plural form of the pronoun after plural antecedents even more:

Social class	sing	plur	sing%
1	9	2	81.8%
2	10	5	66.7%
3	2	17	10.5%
4	4	7	36.4%
5	3	3	50%
Interviewer	9	6	40%

Table 7: Form of relative pronoun after plural antecedent.

It seems reasonable to write up a variable rule for the number agreement of the relative pronoun with its antecedent:

(19) $\text{RelPn} \rightarrow [+\text{plur}] / \begin{matrix} DP \\ \langle[+plur]\rangle \\ \langle[+collective]\rangle \\ \langle[+specific]\rangle \end{matrix} \left\langle \begin{matrix} \overline{} \\ aki \\ ami \\ amely \\ amelyik \end{matrix} \right\rangle \begin{matrix} RelClause \\ \langle[+restrictive]\rangle \\ \langle[+selective]\rangle \end{matrix}$

In the variable rule analysis above the plurality of the relative pronoun form is taken as independent variable and the formal plurality, the semantic plurality (collectivity) and the specificity of the antecedent, the choice of the nominal pronoun stem, and the restrictivity and selectivity of the relative clause as dependent variables. Unfortunately, specificity and the semantic properties of the relative clause for *aki* had not been coded in the database, so only whether these factors have a significant effect on the plural agreement with *ami* and *amely* can be tested. The pronoun *amelyik* provides a knockout factor as it has no plural, therefore its tokens were not taken into account either.

The results of this Goldvarb analysis confirm that these semantic and syntactic factors do not improve the model significantly as the following model is provided as best:

(20)   input probability: 0.001 (!)
        social class: **6 (interviewer)**: 0.447, **1**: 0.825, **2**: 0.733, **3**: 0.115, **4**: 0.383, **5**: 0.517
        form of antecedent: **sing**: 0.385, **plur**: 0.966
        semantical number of antecedent: **sing**: 0.333, **plur**: 0.972
        pronoun: **amely**: 0.981, **ami**: 0.444

These factor weight values are seriously skewed because of the small number of plural forms in this subset of the data. Having established that syntactic and semantic factors do not play a role in this variable rule, *aki* can now be included in the analysis. The rewritten variable rule is much simpler:

(21)   RelPn→[+plur] / $\begin{array}{c} DP \\ \langle[+plur]\rangle \\ \langle[+collective]\rangle \end{array}$ $\left\langle \begin{array}{c} aki \\ ami \\ amely \\ amelyik \end{array} \right\rangle$ RelClause

The remaining factors are still significant:

(22)   input probability: 0.009
       social class: **6 (interviewer)**: 0.577, **1**: 0.640, **2**: 0.709, **3**: 0.167, **4**: 0.410, **5**: 0.462
       form of antecedent: **sing**: 0.309, **plur**: 0.960
       semantical number of antecedent: **sing**: 0.247, **plur**: 0.962
       pronoun: **amely**: 0.929, **ami**: 0.147, **aki**: 0.872

It is visible from these factor weights above, that speakers with higher social status prefer number agreement more than those of lower status (but they still use it less than 50% of the time). The formal and semantic number of the antecedent are obviously significant: these values seem to be equally important for agreement. The form of pronoun shows the most important conclusion: while the animate *aki* and the obsolete *amely* prefer agreement, the default and most frequent nonanimate *ami* disprefers it. The reason for the avoidance of the plural *amik* is puzzling: this form is present in the grammar, there is no stigmatization of its use or the lack of its use. Still, this form is dispreferred in spoken colloquial Hungarian, based on the data from the BUSZI-2 corpus.

6 Agreement in Other Corpora

In this section, much bigger corpora of written texts will be compared to the pattern seen in the BUSZI-2 corpus, which is based on the casual informal spoken language. The results in Section 5.4 were compared to two web corpora. As finding the antecedent of every relative pronoun in these large corpora is impossible, the only quantifiable figure which can indicate the dispreference for plurality of a given relative pronoun is the proportion of plural form for the pronoun, as in Table 6 above.

The first corpus used for this purpose was the open Szószablya web corpus (Halácsy et al. 2004), which includes phonological and morphological analysis, type and token frequencies of every word form. The N of nominal relative pronouns in this corpus was 4,481,305 tokens. The other corpus used was the Google search engine by finding the frequencies of all forms in the paradigms of the nominal relative pronouns which appeared in the Szószablya corpus, and constraining the search on hits in Hungarian only. The N of nominal pronouns in Google was 1,091,666,000.

The results can be seen in Figure 1. The percentage of the plural forms shows the same pattern in the three corpora: the plural of *amelyik* is practically nonexistent, and the plural forms of *ami* are notably rare when compared to its animate counterpart *aki*. The high percentage of plural forms of *amely* in the BUSZI-2 corpus might be accidental due to small sample size, but it might also reflect the less formal register of the spoken corpus.

It can be seen in this comparison, that except for the unusually high percentage of the plural *amelyek* in the BUSZI-2 corpus, the main claim of the avoidance of the plural of *ami* stands in very different, written corpora as well.

7 Summary

This paper presented a quantitative analysis of the use of nominal relative pronouns in Hungarian. It has been shown that the pronoun *amely*, which is to be used after specific antecedents according to prescriptivists, is falling into disuse in Hungarian, but speakers with higher social status preserve it sometimes as an archaism. On the other hand, the selective pronoun *amelyik* is spreading in lower classes, and it is clearly used in non-selective environments in the vernacular of every social class.
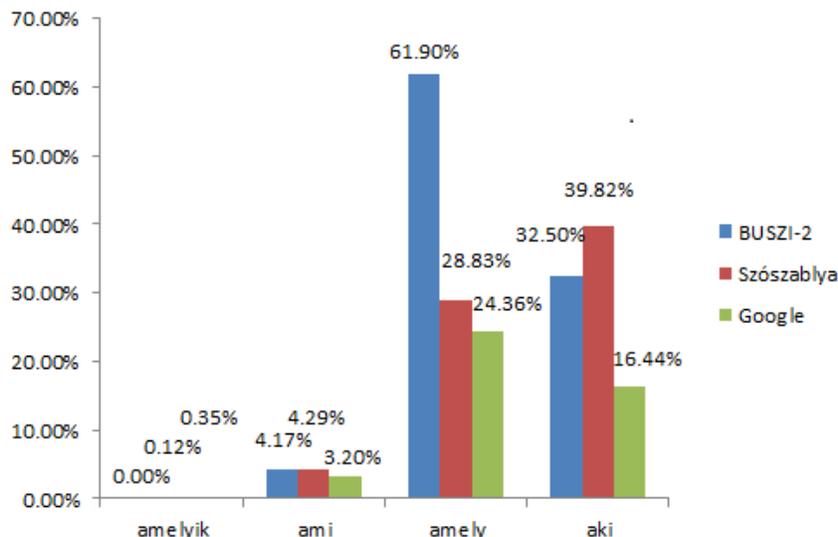
Figure 1: Percentage of the plural form of nominal relative pronouns in three corpora.

It has also been shown that while number agreement between antecedent and the relative pronoun is preserved in every social class (for the animate *aki*, for example), its status is very weak for the default relative pronoun *ami*, for reasons yet not very well understood. Further research could provide arguments for the loss of agreement in the case of this pronoun only.

References

Cedergren, Henrietta J., and David Sankoff. 1974. Variable rules: Performance as a statistical reflection of competence. *Language* 50:333–355.

Fogarasi, Mikls. 1993. Bcs az *amely*-tl [Farewell from *amely*]. In *Nyelvmvel kis tkr*, ed. Imre P. Kovcs, 113–114. Szombathely: Pedaggiai Intzet.

Grtsy, Lszl, and Mikls Kovalovszky, ed. 1980. *Nyelvmvel kziknyv. Els ktet A–K [Prescriptivist handbook. First A-K]*. Budapest: Akadmiai Kiad.

Halácsy, Péter, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón. 2004. Creating open language resources for Hungarian. In *Proceedings of Language Resources and Evaluation Conference (LREC04). European Language Resources Association*.

Kenesei, Istvn. 1992. Az alrendelt mondatok szerkezete [The structure of subordinate clauses]. In *Strukturlis magyar nyelvtan 1. Mondattan [Structural Hungarian Grammar 1. Syntax]*, ed. Ferenc Kiefer, 529–713. Budapest: Akadmiai Kiad.

Kugler, Nra, and Krisztina Laczk. 2000. A nvmsok [Pronouns]. In *Magyar grammatika [Hungarian Grammatica]*, ed. Borbla Keszler, 152–174. Budapest: Nemzeti Tanknyvkiad.

Peth, Jzsef. 2000. „Az ifjsg, aki?" ["The youth, who?"]. In *ltet anyanyelvnk. rsok Grtsy Lszl 70. szletsnapjra [Our vital mother language. Writings for the 70th birthday of László Grétsy]*, ed. Gza Balzs, Anna A. Jsz, and dm Kolti, 387–392. Budapest: Tinta Knyvkiad.

Ttfalusi, Istvn. 2004. *Kis magyar nyelvklinika [Small Hungarian language clinic]*. Budapest: Anno Kiad.

Vradi, Tams. 1998. *Manual of the Budapest Sociolinguistic Interview Data*. (Working Papers in Hungarian Sociolinguistics No. 4, February 1998). Budapest: Linguistics Institute, Hungarian Acedemy of Sciences.

Department of Linguistics
New York University
10 Washington Place
New York, NY 10003
daniel.szeredi@nyu.edu