The hypercorrect key witness

Nóra Wenszky and Attila Novák

1 Introduction

In the course of two projects[1] of the Research Institute for Linguistics of the HAS, morphologically annotated searchable corpora were created from surviving Hungarian texts dating from the 12th c. to the middle of the 18th c. The morphological analyzer software used in both projects is a modified version of the Humor Hungarian analyzer (Novák 2003) created originally to analyze present-day standard written Hungarian. In order that the program could be used to create a morphological annotation of the texts, they had to be normalized, ie the extremely varied orthography of the texts had to be made uniform. In the course of this process, orthographical and dialectal phonetic variation was neutralized, but morphological variation was kept. Normalization and analysis is illustrated in (1).[2]

(1) *Normalization*

original text	hōti	utan	vallia,
normalized text	hite	után	vallja:
stems	hit	után	vall
analysis	N.PxS3	PP	V.S3.Def

Owing to the great variety of texts, normalization proved to be a challenging task. Normalization of words containing inessive *-ban/ben* (bAn) and illative *-ba/be* (bA) suffixes posed a special problem, as these two suf-

[1] Hungarian generative historical syntax [OTKA NK78074], Morphologically analysed corpus of Old and Middle Hungarian texts, representative of informal language use [OTKA 81189].

[2] For a list of abbreviations, see the Appendix.

fixes have been for long neutralized in speech and their orthography also frequently overlapped in the past (Németh 2008). This essay is a case study, which shows how the normalization of words with the suffixes *-bAn* or *-bA* was achieved in a way that made it possible to estimate to what extent different approaches to normalization could possibly have distorted the original data.

The structure of our essay is as follows. First, the distribution and usage of *-bAn* or *-bA* in present-day Hungarian is discussed, with special emphasis on the discrepancy between written and spoken language. Second, the problems raised by these two suffixes in Old and Middle Hungarian texts are described in detail. This is followed by a discussion of what other authors say about the history of the orthography of *-bAn/bA.* Then the results of the analysis of the orthography of *-bAn/bA* in a corpus of Old and Middle Hungarian texts are presented, which if followed by a summary.

2 Present-day Hungarian

In the course of normalization, the original texts were rewritten according to present-day orthographic rules. Therefore, it is indispensable to review the synchronic situation of inessive and illative case marking.

2.1 Locative cases

Similarly to some other Uralic languages, Hungarian has a system of locative cases which exhibit a three-way contrast — a separate class of case endings and postpositions is used for marking location, point of origin or target of motion.[3] The same system applies to temporal location, beginning or endpoint of events. A part of this system refers to inner relations: the suffixes *-bA* 'into', *-bAn* 'in' and *-bÓl* 'out of' represent the illative, inessive and elative case respectively. Their usage is illustrated in (2). Elative case is irrelevant for our further discussion.

(2) *Locative case, inner system*
  a. illative: *-ba/be* 'into'
    Be-ment-em a mozi-ba/kert-be.
    into-go.PAST-1SG the cinema-ILL/garden-ILL
    'I went into the cinema/garden.'

[3] Many Uralic languages also distinguish path of movement as a fourth type of relation. Hungarian uses locative cases (primarily superessive) and locative postpositions for this kind of relation.

b. inessive: *-ban/ben* 'in'
A mozi-ban/kert-ben túl meleg volt.
the cinema-INE/garden-INE too hot be.PAST
'It was too hot in the cinema/garden.'
c. elative: *-ból/ből* 'out of'
Ki-jött-em a mozi-ból/kert-ből.
out-come.PAST-1SG the cinema-EL/garden-EL
'I came out of the cinema/garden.'

In addition of their spatio-local and temporal meaning, these suffixes are also used to mark oblique arguments of verbs in a more or less idiomatic manner:

(3) *Oblique arguments*
a. hisz valami-**ben**
believe something-INE
'believe in something'
b. bízik valaki-**ben**
trust somebody-INE
'trust somebody'
c. kerül valamennyi-**be**
cost some-ILL
'cost some amount (of money)'

In addition to these idiomatic uses of the suffixes, there are also lexicalized adverbs, postpositions and verbal prefixes that etymologically contain one of these suffixes, like *abba(hagy)* 'stop/give up', *elébe* 'in front of—as direction', *hiába* 'in vain', *általában* 'usually, in general', *mostanában* 'nowadays, recently' etc.

2.2 Written and spoken language

In present-day Hungarian, spoken and written language considerably differ regarding the distinction of *-bAn* and *-bA*. While inessive is marked by *-bAn* and illative by *-bA* in writing as shown in (2) above, and failure to do so is stigmatized, both suffixes are generally realized as *-ba/be* in speech, ie the suffix-final *-n* tends to be dropped in inessive. Note that final *n*-dropping is strictly limited to instances of the *-bAn* suffix. Other word final *n*-s are never dropped even if the word happens to end in *-ban/ben*. Eg final *n* of *szebben* 'in a nicer manner' is never dropped even in casual speech—it is part of a different suffix, the essive-modal *-An*. However, lexicalized ad-

verbs etymologically containing the *-bAn* suffix (like, eg *akkoriban* 'at that time', *általában* 'usually') do undergo final *n*-dropping.

Another process, *n*-insertion is also witnessed in speech. In sporadic cases the *-bAn* form is used instead of the illative *-bA*, which is a case of hypercorrection. This, in contrast to *n*-dropping, is stigmatized in speech as well. The situation is summarized below in (4).

(4) -bAn *and* -bA *in present-day Hungarian*

	written	spoken
inessive	-ban/ben	-ba/be ∼ -ban/ben
illative	-ba/be	-ba/be (∼ -ban/ben)

How wide-spread are *n*-dropping and hypercorrection in actual language use? In a representative survey named Hungarian National Sociolinguistic Survey (HNSS, Kontra 2003), the usage of *-bAn* and *-bA* was examined among other phenomena. The 832 subjects were asked to decide whether sentences read out by the interviewer are correct or not. In another type of question printed sentences had to be corrected by the subjects. It turned out that approximately 60% of subjects regarded sentences with *-bA* for inessive correct, both in the spoken and the written judgment tests. Furthermore, half of the instances of hypercorrect *-bAn* for illative were also deemed correct. People living in Budapest tended to conform to the (written) standard the most in the test, but even in their case approximately half of the non-standard forms were regarded as correct.

Another sociolinguistic survey, the Budapest Sociolinguistic Interview, (BUSZI, Váradi 2003) collected data from people living in Budapest at the end of the 1980's. This survey featured grammaticality judgments, elicited spoken production (eg the completion of sentences), reading aloud and spontaneous speech in directed interviews. Certain test situations were explicitly designed to test *-bAn/bA*. There is remarkable discrepancy between the picture the grammaticality or properness judgments show and the data that we get querying the spontaneous speech data recorded during the interviews that was later painstakingly transcribed and recently made available as a searchable database.

In contrast to what seems to be suggested by the results found in the HNSS, the BUSZI spontaneous speech data show that while informants dropped the *-n* of *-bAn* twice as often as they pronounced it, hypercorrect use of *-bAn* instead of *-bA* is practically nonexistent in the corpus. There were informants who could very reliably judge both *n*-dropping and hypercorrect *n*-insertion as nonstandard usage (indicating that they would

have no difficulty at all producing standard written Hungarian consistently distinguishing inessives and illatives), while performing *n*-dropping in 90% of the cases in spontaneous speech.

Access to the BUSZI data has recently been made possible for any researcher wishing to examine it, and we applied for an account to get real-world data concerning contemporary spoken use of *-bAn/bA.* Due to the lack of necessary categorial distinctions, it is difficult to query the BUSZI data for lexicalized cases of *-bA/bAn* final words, thus we restricted our queries to cases where the word contained an actual illative or inessive case ending. Considering only these cases, we get the following results for the BUSZI-2 spontaneous speech data.

In (5), columns bAn and bA mark uses of the corresponding suffixes conforming to the written standard, ie *-ban/ben* for inessive and *-ba/be* for illative. Columns bA' show results for *n*-dropping, while columns bA'n show occurrences of hypercorrect *-bAn.* Columns INE and ILL show the frequency of inessive vs illative, ie this is the ratio we would find in a version of the texts written in standard orthography.

(5) -bAn *and* -bA *in BUSZI-2 spontaneous speech corpus*

BUSZI-2	bAn	bA'	bA	bA'n	sum	bAn	bA'	bA	bA'n	INE	ILL	BAN	BA	STD	NSTD
all case data	1605	2168	1234	5	5012	0.32	0.43	0.25	0.00	0.75	0.25	0.32	0.68	0.57	0.43
informers	874	1596	881	4	3355	0.26	0.48	0.26	0.00	0.74	0.26	0.26	0.74	0.52	0.48
field workers	731	572	353	1	1657	0.44	0.35	0.21	0.00	0.79	0.21	0.44	0.56	0.65	0.35

The results show that inessive is about three times as frequent as illative. As we shall see looking at the results for historic texts, we find a similar ratio there even when including lexicalized adverbs, postpositions and verbal prefixes in addition to productively case marked words in the analysis. In contrast, as we see in the BAN and BA columns, the *-bA* suffix form is used three times as frequently in casual speech as *-bAn.* The bA'n column shows that hypercorrection is extremely rare in the BUSZI spontaneous speech data. A probable explanation for this is that while *n*-dropping is not stigmatized in speech, hypercorrection is, thus speakers can safely avoid using *-bAn* in situations they are not sure about.

Mátyus and her colleagues have examined how different socio-cultural factors influence *n*-dropping (Mátyus et al. 2010) and found that people without a degree tend to drop more *n*'s than people with a degree. Mátyus (2009) also points out that the presence of *n*-dropping also depends on the exact function of *-bAn*: the final consonant is dropped most often in words used as an adverb of place (ie in the least oblique cases).

3 A normalization problem

Although present-day spoken and written language marks inessive and illative in a different way, the standard marking of these case endings in writing only causes problems in the first years of schooling for most Hungarians. However, as manuscripts form the Old (13th c.–1526) and Middle Hungarian (1526–1772) era show, the orthography of these suffixes has not been uniform for several hundreds of years and seems to have been a problem for many authors.

In order to make automatic morphological annotation of the corpora of Old and Middle Hungarian texts possible, all texts were manually normalized to present-day orthography. In the course of this process, orthographic and dialectal variation was neutralized, but the identity of morphemes was kept.

The suffix pair inessive *-bAn* and illative *-bA* posed a special problem, as the shortened form of the inessive suffix is identical to the illative suffix, and the hypercorrect form of the illative suffix is identical to the inessive suffix. So the two distinct suffixes have identical allomorphs. As discussed above in §2, this *-bAn*∼*-bA* alternation is extensively present in present-day spoken Hungarian. According to Németh (2008), this alternation has been present in the language ever since the 14th c., ie since the time these suffixes had acquired their present form. He claims that orthography played a leading role in preserving *-ban* forms.

It is a crucial question from the point of view of normalization how to deal with these alternating suffixes. How could one decide whether a specific *-ba* word ending is an orthographic variant of *-ban* or it is a different morpheme and vice versa? Below we outline three possible solutions for this problem.

3.1 Method 1

One solution, argued for by some of our colleagues, is to keep the original orthography and thus suppose that *-bAn* always corresponds to inessive, while *-bA* marks illative, as altering these forms would amount to changing a morpheme into another in our representations. We will examine in §5, whether this solution is a feasible one.

3.2 Method 2

Another solution is to consider any discrepancy of the *-bAn/bA* marking in the original texts a mere orthographic deviation and normalize all instances of *-bAn* and *-bA* according to the present-day orthographic rules

without any further marking. It can be argued for that if this is done without care, the current norm would be projected onto the historical data, which could result in (unintentional) data falsification.

This issue is most likely to affect idiomatic oblique arguments that may have changed in the course of time. The corpus contains some examples of clear and unambiguous discrepancy between historical and contemporary argument structures. One such example is illustrated in the following example, where the verb *megy* 'to go' is used in a similar fashion to one use of the verb *jár valahol* 'to visit a place', where the verb has a locative argument: 'Going to (or visiting?) Rozsnyó a second time, Mrs Beke said to her husband.'

(6) megy + *superessive*

Masodban	ugyan	Rozsnyon	menven	Bekene	mondotta	az	Uranak
Másodban	ugyan	Rozsnyón	menvén	Bekéné	mondta	az	urának:
másod	ugyan	Rozsnyó	megy	Bekéné	mond	az	úr
Adj. Ine	Adv	N.Sup	V. PartAdv	N	V. Past. S3. Def	Det	N. PxS3. Dat

Here the *-n* ending of the word *Rozsnyón* is an unambiguous locative suffix: the superessive *-On* and its directional counterpart, sublative *-rA,* can never neutralize like *-bAn* and *-bA.* This sense and argument structure of the verb *jár* exists in contemporary standard Hungarian, *megy,* on the other hand, lacks this pattern today. To be fair, of the 782 occurrences of the lemma *megy* in the Middle Hungarian corpus, we find only two undebatable occurrences of the exceptional argument structure above. Both instances contain the superessive. We find no instances containing any other locative suffix (ie the adessive) or postposition (excluding words ending in *-bA* or *-bAn*). And there are 380 instances which are clear counterexamples: the verb *megy* without any verbal prefix and with a directional argument not containing a *-bA/bAn* suffix.

It is of course not at all evident that this ratio of 2 to 380 (0,5%) is representative of all cases of potential data falsification by normalization, but this negligible number seems to indicate that we are not bound to perform extremely massive data corruption if we choose Method 2.

3.3 Method 3

Nevertheless, there is a third option: to normalize the texts in a way that allows for modification but also keeps the original encoding. In this case, all instances of *-bAn* that should be written as *-bA* according to present-day

rules get a special symbol, and similarly, all *-bA'*s that should be *-bAn* today are assigned another special symbol.

This encoding ensures that the morphological analyzer can assign an analysis to the data that corresponds to what is a presumably correct interpretation of the intended meaning of the text, while the normalized form itself explicitly indicates that the original form was altered in a specific way. This makes it possible to detect and correct mistakes—if later a class of instances turn out to have been modified or left unmodified in error, they can easily be located and fixed. For example, in view of the above example, it may well be the case that *Malomban* 'in . . . mill' in example (7) below was normalized to an illative in error. Nonetheless, we can locate such suspicious cases in the corpus easily. Moreover, this approach renders it possible to evaluate whether the first approach mentioned above would be feasible. The meaning of the following example is: 'But once the witness went to (or visited?) the mill in Babót with Andor Bóna'.

(7) *A possible normalization mistake*

ha nem	egykor	az	Baboti	**Malomban**	**ment**	volt	az	fatens,	Bona	Andorral,
hanem	egykor	a	babóti	**malomba'n**	**ment**	volt	a	fatens	Bóna	Andorral,
hanem	egykor	a	babóti	**malom**	**megy**	van	a	fatens	Bóna	Andor
C	Adv	Det	Adj	**N.Ill**	**V. Past.3S**	V. Past	Det	N	N	N.Ins

In the normalization process we opted for Method 3. Specifically, we marked *-ba/be* final words that we assumed should read *-ban/ben* (ie cases of *n*-dropping) as *-ba'/be'* and *-ban/ben* final words that should read *-ba/be* (ie cases of hypercorrection) as *-ba'n/be'n*. This latter, somewhat counterintuitive notation was motivated by the fact that this way it is extremely easy to formulate a regular expression to search for all of these modified cases in the corpus. We tried our best not to change the words where we thought that the interpretation suggested by the original spelling was sufficiently feasible.

During the analysis, the morphological analyzer software was set to interpret all *-bA'* cases as inessive, and all *-bA'n* cases as illative. This, in essence, is identical to Method 2 (§3.2), ie the data were interpreted according to the present-day norm. However, the unique encoding of modified data made it possible to analyze texts from several respects, which are discussed in §5 below.

4 Causes of variation

Before discussing what our normalization revealed, it is useful to look at the possible causes of the great variation in spelling of *-bA* and *-bAn.* The array of texts on which the two analyzed corpora are based is rather versatile. The earliest documents are mainly codices with translated religious texts. Although most of Middle Hungarian sources are printed documents (Dömötör 2006), in the corpus being discussed only originally handwritten texts are present, namely personal correspondence and records taken at witch trials and other court hearings. The long time span and the great variety of text types and authors naturally leads to a diversity of orthography in the lack of a widely accepted orthographic code. Furthermore, the writers spoke different dialects, which could be another source of variation.

4.1 Alternation in the past

As Németh (2008) points out, *n*-dropping at the end of the inessive suffix *-bAn* had already happened by the time the Csángó people left the Hungarian speaking community, ie the 14th c. What is more, he claims the *-bA* variant was probably the only one present in speech until schooling became general and thus more and more people were exposed to the written norm.

4.2 Orthography in the past

The first orthographic code for Hungarian was accepted only in 1832 (Szemere 1974), ie well after the time in which these texts were created. As Kniezsa (1952) points out, Hungarian orthography was formed relatively slowly and the lack of norm lead to the chaotic spelling of Old Hungarian and Middle Hungarian texts. He claims that the setting up of a permanent chancellery in the middle of the 13th c. was the first step towards the formation of a spelling norm. Németh (2008) adds that the writing traditions of offices were in several respects different from that of everyday correspondence.

As for the orthography of inessive *-ban/ben* and illative *-ba/be, -bAn* is already present in its current form as a suffix in the first extant texts, although it is not yet a harmonizing suffix. In the first documents, the function of illative is realized by a postposition *bele* 'into', although it is written in one word with the preceding noun. The ending *-bA* is first recorded as a harmonizing suffix in the earliest extant Hungarian codex, the Jókai

Codex, the surviving copy of which was probably created around 1448 as a copy of a codex written after 1370 (Korompay 1991, 1992).

Németh (2008) claims that the latent orthographical norm of offices in the 17th–18th centuries was that all instances of the inessive and illative suffixes were written as *-bAn*, while the spoken language had only *-bA* for both suffixes. His survey of official documents and private letters show that while official documents stick to *-bAn* in both inessive and illative case, both *-bAn* and *-bA* appear in private letters for each of the cases. A third tradition is discussed by Sinkovics (2011). He claims that beside the spelling norm of authorities, who used *-bAn* in both cases, another tradition has developed by the 17th–18th centuries in Hungary. Printers applied *-bAn* in inessive and *-bA* in illative case, which is the orthographic norm up to the present day.

Printers, authorities and intellectuals were also influenced by the grammars published for Hungarian. Szathmári (1968), in his study on early Hungarian grammars, found that already Mátyás Dévai Bíró's *Orthographia Vngarica* (1549) distinguishes inessive and illative case, and, from Albert Szenci Molnár (1610) on, all grammars claim that inessive case is marked by *-ban/ben,* while illative case is marked by *-ba/be.* Although it seems grammarians suggested the clear distinction of the two cases already in the 1600s, it was not until 1832 that the first official orthographic code for Hungarian made this a rule.

In sum, the orthography of inessive and illative case markers was influenced by the following conflicting facts and demands.

(8) a. *-bAn* ∼ *-bA* alternation from the 14th c. up to the present day in speech, preferring *-bA*
b. orthographic norm for authorities: always *-bAn*
c. orthographic norm for printers: inessive *-bAn,* illative *-bA*
d. grammars from the 17th c. on: inessive *-bAn,* illative *-bA*

5 Results

Having normalized the texts as outlined in §3.3, we could compare texts in the corpora with respect to the orthography of word forms containing *-bAn/bA* suffixes contrasting it with the present-day orthography of the corresponding words. Doing so we found a varied picture, which is, however, to a great extent in accordance with what Németh (2008) claims. In order to get meaningful statistics, we needed to ensure that the subcorpora we examine contain texts that were written by the same people, unless a sub-

corpus exhibits a homogenous behavior in spite of having a heterogeneous set of authors. Fortunately, the majority of texts have been examined by paleographers. Consequently, in many of the publications that we based our research on, it is marked which parts of the text were written by the same author.

The table in (9) below[4] summarizes our findings for a variety of sources in the corpora. The first four pieces of text are codices containing religious texts from the Old Hungarian era: Jókai Codex, Könyvecse az szent apostoloknak méltóságokról [Booklet on the dignity of saint apostles], Festetics Codex and Guary Codex.

The witch trials court records subcorpus consists of the minutes of over a hundred witch trials. This subcorpus covers a time span over a century. The rest are selected parts of the Middle Hungarian correspondence corpus. Poppel-Batthyány is the already processed part of the correspondence of Éva Lobkowitz-Poppel mainly with members of her family (containing letters addressed to Éva Poppel as well). This corpus consists of letters written by several people. Nevertheless, this subcorpus is uniform with regard to the orthography of *-bAn/bA*. In addition to this, the table contains data on the autographic correspondence of three members of the Nádasdy family and letters written by Pál Telegdy and Sándor Károlyi.

(9) *Inessive and illative case marking*

Corpus	**Date**	**Size**	**bAn**	**bA′**	**bA**	**bA′n**	**sum**	**bAn**	**bA′**	**bA**	**bA′n**	**Type**
Jókai Codex	1370–1440	21945	414	38	153	3	608	0.68	0.06	0.25	0.00	**STD**
Könyvecse	1521	8743	170	5	32	3	210	0.81	0.02	0.15	0.01	**STD**
Festetics Codex	1492–1494	19358	395	143	64	43	645	**0.61**	**0.22**	**0.10**	**0.07**	**HYB**
Guary Codex	1490–1508	20239	25	356	157	2	540	0.05	0.66	0.29	0.00	**BA**
Tamás Nádasdy ag.	1544–1559	4535	4	72	30	0	106	0.04	0.68	0.28	0.00	**BA**
Tamás Nádasdy ??	1559	96	4	1	1	0	6	0.67	0.17	0.17	0.00	**∼STD**
Anna Nádasdy ag.	1548–1558	2237	38	6	6	0	50	0.76	0.12	0.12	0.00	**∼STD**
Ferenc Nádasdy ag.	1568–1569	2870	28	1	1	7	37	0.76	0.03	0.03	0.19	**BAN**
Pál Telegdy	1592–1594	3799	71	0	4	22	97	0.73	0.00	0.04	0.23	**BAN**
Poppel–Batthyány	1625–1641	17493	283	11	10	48	352	0.80	0.03	0.03	0.14	**BAN**
Witch trials	1653–1767	132706	2399	42	62	638	3141	0.76	0.01	0.02	0.20	**BAN**
Sándor Károlyi	1704–1722	14314	237	3	72	6	318	0.75	0.01	0.23	0.02	**STD**

[4] For typographical reasons the table in (9) has been broken in two parts, the second part is located on the next page.

Corpus	Date	INE	ILL	BAN	BA	STD	NST	Type
Jókai Codex	1370–1440	0.74	0.26	0.69	0.31	**0.93**	**0.07**	**STD**
Könyvecse	1521	0.83	0.17	0.82	0.18	**0.96**	**0.04**	**STD**
Festetics Codex	1492–1494	0.83	0.17	0.68	0.32	0.71	0.29	**HYB**
Guary Codex	1490–1508	0.71	0.29	**0.05**	**0.95**	0.34	0.66	**BA**
Tamás Nádasdy ag.	1544–1559	0.72	0.28	**0.04**	**0.96**	0.32	0.68	**BA**
Tamás Nádasdy ??	1559	0.83	0.17	0.67	0.33	**0.83**	**0.17**	**∼STD**
Anna Nádasdy ag.	1548–1558	0.88	0.12	0.76	0.24	**0.88**	**0.12**	**∼STD**
Ferenc Nádasdy ag.	1568–1569	0.78	0.22	**0.95**	**0.05**	0.78	0.22	**BAN**
Pál Telegdy	1592–1594	0.73	0.27	**0.96**	**0.04**	0.77	0.23	**BAN**
Poppel–Batthyány	1625–1641	0.84	0.16	**0.94**	**0.06**	0.83	0.17	**BAN**
Witch trials	1653–1767	0.78	0.22	**0.97**	**0.03**	0.78	0.22	**BAN**
Sándor Károlyi	1704–1722	0.75	0.25	0.76	0.24	**0.97**	**0.03**	**STD**

The table contains the date of creation of the texts, their size in words and *-bAn/bA* statistics. In the case of the Jókai Codex, which is a copy of an older, now lost codex, the two dates are the presumable date of creation of the original and that of the extant copy.

The bAn columns contain the number and ratio of occurrences of standard usage of the *-bAn* suffix (corresponding to inessive, or lexicalized *-bAn* final adverbs etc written in standard orthography). The bA' columns contain the same data on the non standard, *n*-dropped *-bA* forms that should be written *-bAn* according to the present orthographic standard. The bA column contains data on standard illative *-bA* or lexicalized words. bA'n is the count and percentage of hypercorrect *-bAn* usage.

The INE column contains the ratio of *-bA/bAn* occurrences that should be *-bAn* according to current standard orthography and ILL contains the same data for *-bA*. BAN and BA contain the ratio of actual *-bAn* and *-bA* occurrences. Columns STD and NST contain the ratio of standard and non-standard *-bAn/bA* orthography in the subcorpus.

5.1 Orthographies of *-bAn/bA*

Concerning the orthography of *-bAn/bA,* the texts in the corpora can be divided into four or five clusters with regard to the extent to which they differ from present-day standard written usage, depending on whether we distinguish one or two degrees of slight deviation form today's standard. The different groups are as follows.

1a. Some of the texts clearly distinguish the two suffixes in a manner that to a great extent corresponds to our present-day grammatical intuition. We marked this class of documents as STD (standard) in the table. An example of this is Jókai Codex, which was written in the Old Hun-

garian period. Another is Könyvecse (another codex containing religious texts) from the beginning of the 16th century. Part of the Middle Hungarian correspondence, eg that of Sándor Károlyi from the beginning of the 18th century also belongs to this group.

1b. Some documents mostly use these suffixes as the present-day standard, but sometimes (a little more often than in cluster 1a) *-bAn* is replaced by *-bA.* Such a distribution of the suffixes resembles present-day careful speech. Autographic letters written by Anna Nádasdy belong to this group. We marked this class of documents as ∼STD in the table.

2. Other sources mostly neutralize the two suffixes either as *-bAn* or as *-bA,* although sporadic occurrences of the other suffix form are generally present in most of the neutralizing type of documents as well. The documents that tend to use *-bA* in all places are the Guary Codex and the autographic part of Tamás Nádasdy's correspondence. We marked this class of documents as BA in the last column of the table.

3. Completely hypercorrect usage, ie *-bAn* in all (most) places, primarily occurs in Middle Hungarian texts. This type of orthography seems to have emerged in the second half of the 16th century and characterizes almost all of the official court records and legal documents of the era, and much of personal correspondence as well. We marked this class of documents as BAN in the table.

4. The last category is made up of texts that use both forms of the suffix, generally in the way the present standard would require, but hypercorrect *-bAn* and shortened *-bA'* also occur in considerable numbers. An example for such a hybrid text is the Festetics Codex. We marked this class of documents as HYB in the table. A plausible explanation of this distribution of suffixes could be that a text was written by several hands. However, while this is true for Festetics Codex, 98% of the text is the work of a single hand according to paleographers, thus the explanation mentioned above is hardly tenable. A more fine-grained analysis of the Festetics Codex data would be needed to come up with another explanation.

5.2 Analysis

As shown in §4.2 above, several factors must have influenced the orthography of the inessive and illative suffix in Hungarian. In the lack of Old and Middle Hungarian speech recordings, however, the distribution of *n*-dropping and hypercorrect *-bAn* use in actual speech can only be hypothesized based on written records. Németh (2008) claims that *n*-dropping has been wide-spread ever since the 14th c. in the whole Hungarian speaking community, what is more, in certain periods the form *-bAn* was only

present in writing. This view suggests that the orthography of extant texts does not reflect actual language use in this respect.

Our data confirm that the orthography of surviving texts was primarily influenced by factors other than the actual pronunciation of the inessive and illative suffixes. As the chart in (9) above shows, texts conforming to the present standard orthography appear in the whole examined range of time, ie from the 14th c. to the 18th c. These texts clearly mark the semantic distinction of inessive and illative case.

This distinction disappears in texts dated from the first half of the 16th c. Both cases are marked by the suffix *-bA,* which probably reflects actual language use. However, in the second half of the 16th c. a new tendency emerges: both cases are marked by *-bAn,* both in official documents and private letters. Could this mean such a rapid change of pronunciation?

This orthography used as a norm for official legal documents of the Middle Hungarian era seems to be rather counterintuitive from our present-day perspective, given that hypercorrect *-bAn* usage is extremely rare and stigmatized today both in speech and in writing. Note, however, that this type of neutralizing orthography deviates from current standard written language to a much lesser extent when looking at the frequency of mismatches than the more intuitive BA-type orthography. This is due to the fact that inessive is at least three times as frequent as illative in any sizable body of Hungarian text (even when counting lexicalized items containing one of the *-bA/bAn* suffixes as well). This is true of present-day Hungarian as well as Old and Middle Hungarian.

However, it is hard to believe that this orthography corresponded to actual spoken usage, although it may have influenced spoken performance of those using it. It is also interesting to note that although *-bA* forms do sporadically occur in court records, these are mostly restricted to records of actual statements of witnesses and defendants. They practically never occur in the formulas describing circumstances of the trial or the sentence.

5.3 A family of key witnesses

Having a closer look at the texts of the Middle Hungarian corpus, it is worth noting that a considerable amount of letters come from the members of the renowned Nádasdy family. Baron Tamás Nádasdy (1498–1562) made a spectacular career in the 16th century, he was the governor of Croatia and Slavonia, and the palatine of Hungary. He also set up a printing press in Sárvár, Hungary. In his autographic letters, he almost exclusively uses *-bA* for both inessive and illative case. There is a letter in Nádasdy's legacy that paleographers could not categorize as autographic for sure.

This text (marked as "Tamás Nádasdy ??" in chart (9)) clearly falls into the ∼STD category, ie inessive is mostly marked by *-bAn* with a single exception, while illative is marked by *-bA*. Based on this single variable, we could bet that the debated letter was probably written by another hand, as Tamás Nádasdy's own writing definitely falls into the BA category.

Other letters in the Nádasdy family come from Anna Nádasdy, who was the sister of Tamás Nádasdy. She, contrarily to her brother, uses an orthography which is near the present standard (∼STD). Tamás Nádasdy's son, Ferenc (1555–1604), however, almost exclusively used *-bAn* for both inessive and illative, ie his writings fall into the BAN group. Could these three members of one family living in the same period of time speak three different dialects?

This is rather unlikely. A much more feasible explanation is that they learnt and used different orthographic norms. The rise of the hypercorrect *-bAn* norm for official documents happened in the second half of the 16th c., and in the 17th and 18th centuries this practice seems to have been almost exclusive for official written language (Németh 2008 : 100). Ferenc Nádasdy, the youngest of the three members of the family discussed here, apparently learnt this emerging latent norm of hypercorrection. The three Nádasdys could thus be summoned as key witnesses in our debate on the issue of normalization of *-bAn/bA* final words.

5.4 Normalization revisited

At the beginning of this discussion, three ways of normalization were sketched in §3 above. According to Method 1 (§3.1), all *-bA/bAn* suffixes in extant texts should be considered to have the present grammatical function, ie all written *-bA*'s should be considered illative, while all written *-bAn*'s an instance of inessive case. According to Method 2 (§3.2), the suffixes should be corrected according to the present norm, taking into account the context of each occurrence. These two methods were discarded in favour of a third solution (§3.3): all endings not conforming to the present-day standard got a special symbol: *-bA'* marks supposed cases of *n*-dropping, while *-bA'n* stands for hypercorrect instances of *-bAn.*

On the one hand, Method 1 would have suggested that in the Nádasdy family three contemporary relatives spoke three different dialects: one lacking inessive case marking, the other having both inessive and illative, while the third member having no illative marking. Method 2, on the other hand, would have hidden the different orthographic traditions used side by side in the 16th c.

Thus the choice of the Method 3 was justified, and it helped to reveal facts about the history of spelling norms of Old and Middle Hungarian. Method 3 made it possible to calculate how much a certain body of text deviates from present-day orthography and in what ways, and revealed that *-bAn/bA* data in the corpus does in fact deviate from current usage due to merely orthographic factors. Furthermore, it made it possible for us to estimate how much choosing Method 1 could have distorted the data.

It must be emphasized, however, that the actual analysis of data by the morphological analyzer software can follow either Method 1—respecting original orthography —, or Method 2 — taking into consideration the normalization for the present-day norm. Our analyzer program was set to Method 2, ie all instances of *-bA'* endings (*n*-dropping) were analyzed as inessive, and all instances of *-bA'n* endings (hypercorrection) were regarded as illative case.

The table in (10) summarizes *-bAn/bA* statistics for the Middle Hungarian corpus, the four Old Hungarian codices mentioned above, for the aggregate data of both corpora and of subcorpora of BA and BAN type orthographies.

(10) *Summary*

	bAn	bA'	bA	bA'n	sum	bAn	bA'	bA	bA'n	INE	ILL	BAN	BA	STD	NSTD
Mid. Hung.	3251	231	239	750	4471	0.73	0.05	0.05	0.17	0.78	0.22	0.89	0.11	**0.78**	**0.22**
Old Hung.	834	537	374	48	1793	0.47	0.30	0.21	0.03	0.76	0.24	0.49	0.51	**0.67**	**0.33**
All	4255	773	645	801	6474	0.66	0.12	0.10	0.12	0.78	0.22	0.78	0.22	**0.76**	**0.24**
BA	29	428	187	2	646	0.04	0.66	0.29	0.00	0.71	0.29	0.05	0.95	**0.33**	**0.67**
BAN	2781	54	77	715	3627	0.77	0.01	0.02	0.20	0.78	0.22	0.96	0.04	**0.79**	**0.21**

The data show that we would have misanalyzed 1574 tokens, about one quarter of all *-bAn/bA* final words in the whole corpus had we applied normalization Method 1 (column NSTD). For subcorpora using the BA type orthography, the error rate would have been as high as 67%, while for those using the BAN type hypercorrect orthography, it is about 21%. On the other hand, our (very rough) estimate for errors introduced by possible incorrect projection of our present-day intuitions about *-bAn/bA* to the historical data using Method 2 for analysis is about 0.5%.

6 Summary

This study shows how inessive *-bAn* and illative *-bA* endings in Old Hungarian and Middle Hungarian extant texts were normalized for the purposes of automatic morphological analysis. These endings have shown al-

ternation in speech in the past 600 years and their orthography was not at all uniform. In the course of normalization, special symbols were assigned to endings with *n*-dropping (non-standard inessive case) and hypercorrect *-bAn* suffixes (non-standard illative case).

Based on the distribution of the standard and non-standard suffixes, texts fell into four major categories: near standard (STD), extensive *n*-dropping (BA), extensive hypercorrection (BAN) and mixed (HYB). As the first three types existed in the 16th c. even in sources from one family (namely the Nádasdy family), it is plausible to suggest that three orthographic norms were simultaneously present at that time. This confirms the findings of Németh (2008).

REFERENCES

Dömötör Adrienn. 2006. *Régi magyar nyelvemlékek. A kezdetektől a XVI. század végéig.* Budapest: Akadémiai Kiadó.

Kniezsa István. 1952. *Helyesírásunk története a könyvnyomtatás koráig.* Budapest: Akadémiai Kiadó.

Kontra Miklós (ed.). 2003. *Nyelv és társadalom a rendszerváltáskori Magyarországon.* Budapest: Osiris.

Korompay Klára. 1991. *A névszóragozás.* In: Benkő Loránd (ed.), *A magyar nyelv történeti nyelvtana I.* Budapest: Akadémiai Kiadó. 284–318.

Korompay Klára. 1992. *A névszóragozás.* In: Benkő Loránd (ed.), *A magyar nyelv történeti nyelvtana II/1.* Budapest: Akadémiai Kiadó. 355–410.

Mátyus Kinga, Bokor Julianna, and Takács Szabolcs. 2010. „Abban a farmerba nem mehetsz színházba." A (bAn) variabilitásának vizsgálata a BUSZI tesztfeladataiban. In: Váradi Tamás (ed.), *IV. Alkalmazott Nyelvészeti Doktoranduszkonferencia.* Szeged: SZTE. 85–99. (www.nytud.hu/alknyelvdok10/proceedings10.pdf)

Mátyus Kinga. 2009. Az inessivusi (bAn) nyelvtani szerepei. In: Váradi Tamás (ed.), *III. Alkalmazott Nyelvészeti Doktoranduszkonferencia.* Szeged: SZTE. 69–86. (www.nytud.hu/alknyelvdok09/proceedings.pdf)

Németh Miklós. 2008. *Nyelvi változás és váltakozás társadalmi és műveltségi tényezők tükrében. Nyelvi változók a XVIII. században.* Szeged: SZTE, Juhász Gyula Felsőoktatási Kiadó.

Novák Attila. 2003. Milyen a jó humor? In: Alexin Zoltán and Csendes Dóra (ed.), *Az 1. Magyar Számítógépes Nyelvészeti Konferencia előadásai.* Szeged: SZTE. 138–145.

Sinkovits Balázs. 2011. *Nyelvi változók, nyelvi változások és normatív szabályozás.* PhD dissertation, University of Szeged.

Szathmári István. 1968. *Régi nyelvtanaink és egységesülő irodalmi nyelvünk.* Budapest: Akadémiai Kiadó.

Szemere Gyula. 1974. *Az akadémiai helyesírás története (1832–1954).* Budapest: Akadémiai Kiadó.

Váradi Tamás. 2003. A Budapesti Szocilingvisztikai Interjú. In: Kiefer Ferenc and Siptár Péter (ed.), *A magyar nyelv kézikönyve.* Budapest: Akadémiai Kiadó. 339–359. (www.nytud.hu/oszt/elonyelv/adat/buszi.pdf)

Appendix: List of abbreviations

Adj	adjective	**HYB**	hybrid
Adv	adverb, adverbial	**ILL**	illative
ag	autographic	**INE, Ine**	inessive
An	*an* ∼ *en* (inessive-modal suffix)	**Ins**	instrumental
bA	*ba* ∼ *be* ∼ *ban* ∼ *ben* (illative suffix)	**N**	noun
bA'	*n*-dropping in inessive	**NSTD**	non-standard
bAn	*ban* ∼ *ben* ∼ *ba* ∼ *be* (inessive suffix)	**On**	*on* ∼ *en* ∼ *ön* (superessive suffix)
bA'n	hypercorrect bAn in illative	**Part**	participle
bÓl	*ból* ∼ *ből* (elative suffix)	**PP**	postposition
BUSZI	Budapest Sociolinguistic Interview	**Px**	personal suffix
C	conjunction	**rA**	*ra* ∼ *re* (sublative suffix)
Dat	dative	**S, SG**	singular
Def	definite	**STD**	standard
Det	determiner	**Sup**	superessive
EL	elative	**V**	verb
HNSS	Hungarian National Sociolinguistic Survey		

Nóra Wenszky and Attila Novák
wenszkynora@gmail.com
novak.attila@itk.ppke.hu
Hungarian Language Technology Research Group
Hungarian Academy of Sciences
Pázmány Péter Catholic University