Corpus Building from Old Hungarian Codices

Eszter Simon

1 Introduction

The availability of annotated language resources is becoming an increasingly important factor in more and more domains of linguistic research, since high-quality linguistic databases can provide a fertile ground for theoretical investigations. Historical corpora represent a rich source of data, but only if the relevant information is specified in a computationally interpretable and retrievable way. Digitization should not be confined to scanning old manuscripts as images but should extend to making the primary data available in digital form. After linguistic enrichment, data sources ensure the possibility to access the data in a much more sophisticated way. Computers can provide support in ensuring consistency, completeness, and reliability of the metadata.

Several databases of historical texts enriched with some kind of linguistic information and metadata have recently been created for various Indo-European languages, such as the Penn-Helsinki Parsed Corpus of Middle English (Kroch and Taylor, 2000), the Tycho Brahe Parsed Corpus of Historical Portuguese (Galves and Britto, 2002), or the Welsh Prose

corpus (Thomas et al., 2007).

One of the major aims of our project was to produce such an annotated corpus for specific stages of the history of the Hungarian language, similar in purpose to the initiatives of the database building projects mentioned above.

The collection of Old Hungarian texts and presenting them in a computationally retrievable way started without any Hungarian predecessors, thus building the Old Hungarian Corpus was a pioneering effort.

Building databases of historical texts and developing language processing tools for the cultural heritage domains is a highly interdisciplinary endeavour, which requires close collaboration across disciplines. General corpus building attempts tend to process texts which have been already digitized or originally created in an electronic format; but this is not the case with historical documents. Building corpora from the time before electronic formats is more costly and time-consuming and needs more laborious methods.

The goal of this appendix is to describe the full workflow of text processing from scanning the codices to submitting queries through an online search service. Section 2 presents the acquisition of source data and the digitization process of the original linguistic material. It discusses the heterogeneity of the Old Hungarian orthographic system and several challenges during Optical Character Recognition (OCR) and text encoding. Section 3 gives an overview of corpus annotation, the normalization of tokens, the morphological analysis, and the morphosyntactic

disambiguation. Section 4 presents the structure of the corpus, the text processing levels, and how linguistic annotation and several metadata are represented in the corpus. Section 5 describes the corpus query tool, which facilitates the linguistic analysis of large amounts of linguistic data.

2 Collecting the Original Linguistic Material

A corpus is a well-organized collection of data, "collected within the boundaries of a sampling frame designed to allow the exploration of certain linguistic feature (or set of features) via the data collected" (McEnery, 2004). The corpus should aim for balance and representativeness within a specific sampling frame, in order to allow a particular variety of a language to be studied. However, if the object of study is a highly restricted sublanguage or a dead language, identifying the texts to be included in the corpus is straightforward. This is the case with Old Hungarian texts: when constructing the Old Hungarian Corpus, we acquired all available sources from the Old Hungarian period, creating a corpus of a fixed size (more than 2.2 million tokens).

The project aimed to collect and process only the continuous texts: codices and several minor texts, thus Hungarian fragments found in foreign texts were not considered. As a result, 47 codices have been made available digitally in their original orthographic form, eleven of them have also been normalized, and four of them have been morphologically analyzed and morphosyntactically disambiguated. Furthermore, the original and normalized versions of several minor texts have also been produced.

Work in the first phase started with the acquisition of source data, part of which has already been converted into some electronic text format. Documents coming from various sources (publishing companies or historical linguists) were converted into uniform, UTF-8 encoded simple text files (see Section 2.1).

Another source was the Computational Database for Historical Linguistics (Jakab and Kiss, 1994, 2001; Jakab, 2002). The database contains the stems of the words of a few Old Hungarian codices in modern transcription, in alphabetical order. The corresponding tokens in their original form are presented with locus markers (page and row numbers), and orthographic, etymological, phonological, morphological, and semantic information. Since the information about the order of tokens in a row is not provided, recovering of the original word order was needed. Afterwards, the documents were converted into a uniform, UTF-8 encoded simple text. Normalized word forms were reconstructed from the combination of stems in modern transcription and the corresponding morphological information. The latter was also used to supply the part-of-speech (POS) tags and the full morphological analysis for each token. After a manual proof-reading and correction, we obtained three codices digitally available in their original orthographic form, normalized, and morphologically analyzed.

2.1 The Digitization Process

A significant part of the linguistic material was only available in print. In this case, digitization was carried out by manual typesetting or by scanning

followed by a conversion process from the scanned images into regular text files aided by an OCR software.

2.1.1 Optical Character Recognition and the Orthographic System of Old Hungarian

Old Hungarian texts are heterogeneous mainly because of the absence of a spelling norm. The adaptation of the Latin alphabet to Hungarian posed several problems. The main challenge was that there are Hungarian phonemes which do not exist in Latin, so new characters were needed to represent them. The orthography in the 14-16th centuries was far from uniform, in addition, one codex could be written by more than one author, which causes even more heterogeneity in the texts.

Typically, sound–letter correspondences vary a lot even within a single text sample. One sound is often written with various characters, e.g. *v̲yragnac u̲iraga* [v̲irágnak v̲irága] 'flower of flower' (Old Hungarian Lamentations of Mary). In addition, one letter can stand for multiple sounds, e.g. *z̲erz̲ete zerent* [sz̲erz̲ete szerint] 'after his order' (Jókai C. 124). Moreover, some letters can refer to vowels and consonants as well, e.g. the letter *v* was used to represent the sounds [v, u, u:, y, y:] for centuries.

According to Kniezsa's classification (Kniezsa, 1952), Hungarian phonemes not existing in Latin are represented in three ways:

1. In the first type, scribes work without diacritics: they combine more letters to represent a sound, e.g. $[t\int] \rightarrow ch \sim cz \sim chy \sim chi \sim cy$

2. The second type works with diacritics: letters with diacritical marks

are used for representing Hungarian sounds, e.g. *[tʃ]* → *ć* ∼ ʟ ∼ ʟ*'* (ʟ is the so called Hussite [tʃ], see Section 2.1.2)

3. The third type is a kind of mixture of the first and the second types. In this case, the scribe applies letter combinations and diacritical marks as well, e.g. *[tʃ]* → *ch* ∼ *chy* ∼ *cyh* ∼ *c* ∼ *chi* ∼ *ch'* ∼ *cz* ∼ *ts* ∼ *ć* ∼ ʟ ∼ ʟ*'* ∼ ʟ*h* ∼ ʟ*z*

As can be seen, Old Hungarian texts contain a large number of special characters, so a key aspect of an OCR software was its ability to be trained. This means that the software does not work with a closed set of characters, but has a training system built in enabling it to deal with characters different from basic Latin ones. For this purpose, we used Abbyy FineReader 9.0 Professional edition[1], which can be trained in an interactive way and produces a fairly good quality result.

The performance of the OCR system was evaluated by counting word accuracy, which is the rate of the number of correctly recognized words and the number of all words in a document. As expected, the results show that accuracy highly depends on the orthographic system used in a codex. We chose three codices representing the three orthographic types mentioned above for evaluation. A Modern Hungarian text was also processed as a baseline.

As can be seen in Table 1, the best result was produced on the first orthographic type which does not use diacritical marks: this is similar to the result on Modern Hungarian text. The large number of special characters used in codices of the second and third types decreased the

Table 1: Word accuracy of the OCR system for orthographic types.

codex	type	word count	correct	WAcc (%)
Kulcsár	no diacritics	36,321	35,258	97.07
Munich	diacritics	74,657	50,790	68.03
Czech	mixed	11,478	7,910	68.91
–	modern	5,121	5,068	98.97

performance by approx. 30%. In the case of codices using characters without diacritical marks, the task of the OCR system is recognizing the basic Latin characters, so it produces a fairly good quality result. However, recognizing complex, combined characters, which can be very similar to each other, causes difficulties. This is due to the fact that the OCR system cannot handle the diacritics properly, as it is also mentioned in reports of similar projects, e.g. (Volk et al., 2010).

The OCR step was completed by extensive manual proof-reading and correction to ensure good quality initial resources as input to further processing steps.

2.1.2 Text Encoding

Character encoding is rarely an issue for languages like English, which typically use basic Latin characters. However, for languages which use a large number of special characters, encoding is an important issue if one wants to build a consistent corpus to be searched reliably and displayed properly. Consistency is a basic requirement so that one can ask a query on

the whole corpus. One of the great advantages of corpora is that they provide not only separate examples but all instances of the searched term, so analyses based on frequency become available. This important property of corpora can be ensured only if one follows the principle of consistency, and always uses the same appropriate character for representing the same letter and different characters for representing different letters.

For this purpose, we use UTF-8 encoded standard Unicode characters in the entire corpus. The Unicode Standard[2] is a multilingual coding system which provides a consistent encoding for most of the world's writing systems. Recently, it became an international standard, which supports the worldwide interchange, processing and display of written texts of diverse languages. One of the great advantages of Unicode is that it properly handles various accented and multi-accented characters, since basic characters and combining diacritical marks are represented by their own codes. For example, the character *ÿ*, which frequently appears in Old Hungarian texts, can be easily compiled from a *y* and a combining diaeresis. Combining diacritical marks can also be accumulated, so that most of the special Old Hungarian characters can be represented by standard Unicode characters.

However, it is a hard task to ensure consistency when dealing with such an extremely diverse language material that we have. There is still an Old Hungarian character which is not present in Unicode charts: this is the so called Hussite [t$\int$]. It is widely used in the Hussite Bible, the orthography of which was influenced by early 15th century Czech spelling. This orthography later spread among Hungarian scribes and had a great

influence on the spelling of later 16th century Hungarian codices. Later it became extinct and is not used in the Modern Hungarian alphabet. It looks like a small capital L (ʟ) and is similar to some Unicode characters. However, one of the Unicode design principles is that characters have well-defined semantics, thus if we want to be consistent, we cannot use the characters which only look alike, but are not the same in their semantics. So we decided to follow Volf (1874) and replace this character by *č*, which is used if and only if the Hussite [t$\int$] is used in the original codex.

Codices from the Old Hungarian period are hand-written texts, which already have transcribed editions. We opted for using editions as the basis of corpus compilation. However, the editions were prepared in different periods, following different scientific requirements, and restricted by varying typographical possibilities. Thus, the same character is often displayed in different ways in different editions. To ensure consistency, we eliminated this kind of randomness by using the same standard Unicode character for characters with the same semantics.

When constructing the texts in their original orthographic form, we kept the punctuation marks, the hyphenation (or the lack thereof), and the upper- and lower-case letters as they are in the codices. However, we did not record the colours, boldface markings, and other kinds of emphasis applied in the codices. We did not aim for strict paleographic adherence, but our goal was to build a consistent database for linguistic research purposes.

3 Corpus Annotation

In the second phase of the corpus building workflow, linguistic annotation was developed. The development of an annotation requires a number of standard computational language processing tasks:

- tokenization and sentence segmentation;
- normalization of tokens;
- morphological analysis and morphosyntactic disambiguation.

However, processing of texts from the time before electronic formats is far from trivial. Since spelling and punctuation rules in this early period of the Hungarian language were not regularized, this step requires manual work, which can be aided by automatic pre-processing tools.

3.1 Tokenization and Sentence Segmentation

In the case of codices which have not been normalized, but are digitized only in their original orthographic form, tokenization means that we simply separate words from each other and concatenate hyphenated word parts. In the case of normalized codices, tokenization was done manually during the normalization step. We followed the Modern Hungarian spelling rules, so some words had to be split up, while others had to be joined together. When a word in the original text belongs to different constituents (as defined by our normalization guidelines), the word is split into the relevant parts. It is marked by double equals signs at the end of the first word and

at the beginning of the next word. In (1), we present a typical case: verbal particles appear postverbally in imperative sentences and are spelled apart from the verb according to the Modern Hungarian spelling rules. However, they are often spelled together in Old Hungarian texts.

(1)

de	**säbädicz==**	**==mk**	mikët	a	gonostwl
de	szabadít-s	meg	mink-et	a	gonosz-tól
but	deliver-IMP	PRT	we-ACC	the	evil-ABL

'but deliver us from evil' (Munich Language Record 114v)

Words which are spelled apart in the original text, but constitute one word in Modern Hungarian, are joined, e.g. the noun phrase and the adverbial suffix in (2). When the original word is broken apart by a line or page break, it is marked by double *at* signs, as can be seen in (3).

(2)

harmal	napon	**halottay bool**	felthamata
harmad	nap-on	halott-a-i-ból	fel-támad-a
third	day-SUP	dead-POSS-PL-ELA	up-rise-PST.3SG

'on the third day he is risen from the dead' (Munich Language Record 114v)

(3)

egmen-@@denic	ǫ	at't'afiat	nē	zorongat't'a
egymindenik	ő	atyjafiá-t	nem	szorongat-ja
either	he	brother.POSS.3SG-ACC	not	thrust-DEF.3SG

'neither shall one thrust another' (Vienna C. 205)

Since modern punctuation rules were created only in the 17th century, we cannot split the text into sentences based on the punctuation marks used in

the original texts. For this reason, sentence splitting was made manually during the normalization step. In the case of non-normalized texts, we applied a quasi-sentence splitting, i.e. the text was split into 10-token sequences.

3.2 Normalization

Because of the heterogeneity of the Old Hungarian orthographic system, a normalization step is required, in which the original tokens are transcribed into their modern form. This is a common step applied in most of the projects aiming at processing historical linguistic material, e.g. McEnery and Hardie (2003). Normalization is inevitable and is obviously of critical importance: without normalization the performance of automatic annotation in later stages will suffer a dramatic decrease (Rayson et al., 2007).

One of the principal criteria of the normalization step is adherence to the original text – at least at the level of the morphosyntactic representation. Thus, we aimed for preserving all words and morphemes, even those which do not exist in Modern Hungarian. In (4), the word *ýsa* is an adverb which is known from the Funeral Sermon and Prayer. The word means 'certainly' and soon disappeared from the Hungarian language. Since we wanted to preserve all words, we normalized it as 'isa', not as 'bizony', its most appropriate translation into Modern Hungarian.

(4)

ýsa	pur	es	chomuv	uogmuc
isa	por	és	hamu	vagyunk
sure	dust	and	ash	be.1PL

'sure, we are dust and ashes' (Funeral Sermon and Prayer)

In (5), the word form *fekette* preserves a morphological construction which does not exist in Modern Hungarian. It is an adverbial participle which is used to modify the verb phrase or the whole sentence, thus plays a role similar to that of an adverb. Its speciality is that it agrees with the subject of the construction in number and person. There are adverbial participles even in Modern Hungarian, but they have only one form and do not agree with the subject. Following the principle of adherence to the original text, this and similar morphological constructions are preserved in the normalization process.

(5)

lata	ǫ	napat	**fèkette**
lát-á	ő	napá-t	fek-ett-e
see-PST.DEF.3SG	he	mother.in.law.POSS.3SG-ACC	lie-PART-3SG

'he saw his wife's mother laid' (Munich C. 14rb)

The second principle of normalization is consistency, thus orthographic variants of the same lexical item must be neutralized and converted into the same normalized version, e.g. *mēden* ∼ *menden* ∼ *minden* ∼ *mėnden* ∼ *mēndėn* → *minden* 'all'. We always followed the Modern Hungarian spelling rules during the normalization process.

Since the Old Hungarian linguistic material mostly consists of Bible translations and religious texts, there is a large amount of biblical names in

it. Following the principle of consistency, proper names written in several diverse forms were also normalized. For this purpose, we used a modern Bible translation[3], and all names were transcribed to the form used in this translation.

3.3 Morphological Analysis and Disambiguation

The normalization step has two main purposes: on the one hand, it makes it possible to find all instances of a word, irrespectively of how they are originally written; on the other hand, the normalized word form is the input to the morphological analysis. Since the original word forms are converted into Modern Hungarian spelling, technology developed for the morphological analysis of Modern Hungarian can be applied to the historical texts.

We used the morphological analyzer engine called Humor (High speed Unification MORphology) (Prószéky and Kis, 1999), which was originally developed for Modern Hungarian and has been adapted to Old Hungarian. First, Old Hungarian morphological constructions which are extinct now have been formalized and added to the grammar of the analyzer. Second, its lexicon has been expanded by adding special old words which are not used in the modern language.

Since the analyzer generates all potential morphological analyses for each token, a disambiguation step is required to select the most appropriate analysis. We used HunPos (Halácsy et al., 2007), a statistical POS tagger, which requires a large amount of manually disambiguated Old Hungarian

texts as a training corpus. For this purpose, morphologically analyzed and disambiguated texts produced from the Computational Database for Historical Linguistics (see Section 2) were used. For getting a corpus as error-free as possible, we manually validated and corrected the output. Since the theoretical aim of the project was to investigate syntactic changes in the history of the Hungarian language, we do not aim at full morphophonological representation, thus we do not mirror the whole morphemic structure of tokens in the analysis. The inflectional suffixes are fully encoded, while the derivational suffixes are not. Since Hungarian is a highly inflectional language, it expresses grammatical elements in a single word form using affixes for expressing grammatical phenomena. Suffixes, often multiple ones, must be attached to the word stem in strict order, and the last one always provides information about the syntactic role of the word form in the sentence. For this reason, several syntactic phenomena can be explored even at the morphological annotation level.

4 The Structure of the Corpus

The structure of the corpus, i.e. the annotation levels are parallel with the text processing steps, which are presented in Table 2. Based on this, six levels and five tasks can be distinguished throughout the text processing workflow.

The sophisticated, linguistically relevant query often refers to different levels of language information contained in the corpus. For making all pieces of information available, the corpus contains all kinds of textual data

Table 2: Text processing levels.

(1)	scanned codex
	→ *automatic* OCR
(2)	raw OCR output
	→ *manual* correction
(3)	original orthographic form
	→ *manual* normalization
(4)	normalized form
	→ *automatic* morphological analysis
(5)	lemmatized and morphologically analyzed form
	→ *semi-automatic* disambiguation
(6)	disambiguated form

corresponding to the text processing levels. Thus, for each token, the corpus provides the following pieces of linguistic information:

- original orthographic form (3): *adÿad*
- normalized form (4): *adjad*
- lemma (6): *ad*
- morphological analysis (6): *V.Sub.S2.Def*

The example is a definite subjunctive/imperative form of the Hungarian verb *ad* 'give' in 2nd person singular. The numbers in parentheses are the numbers of the text processing levels (see Table 2) from which the information comes.

The basic data format is the so called multitag format, i.e. a tab separated simple text file which contains one token in every row and additional information corresponding to text processing levels in columns, as can be seen in Table 3. Sentence boundaries are marked by empty lines.

Table 3: The multitag format.

page	original	normalized	lemma	analysis
1	Vram	Uram	Úr	N:P.PxS1
1	engem	engem	én	N:Pro.S1.Acc
1	segeýtheny	segíteni	segít	V.Inf
1	syees	siess	siet	V.Subj.S2

The morphological analyzer adapted to Old Hungarian has been originally developed for Modern Hungarian, for which it is widely used in the Hungarian language technology community. However, its linguistic formalism does not fit into any international annotation schemes, therefore we plan to convert it into one of the widely used international standard formalisms. For clarification, here we provide the linguistic glosses of the example in Table 3.

(6)

Vram	engem	segeýtheny	syees
Ur-am	engem	segít-eni	sies-s
Lord-POSS.1SG	I.ACC	help-INF	hurry-IMP.2SG

'my Lord, hurry to help me' (Festetics C. 1)

Besides the linguistic annotations, the corpus is also enriched with several kinds of metadata. The primary metadata are locus markers, which provide

information about the place of the token in the original document (page, line, etc.). In texts containing Bible translations, biblical markers (book, chapter, verse) are also provided in a standard way, according to the modern Bible translation of the Szent István Társulat (St. Stephen Association), which provides the possibility to find the given part in other Bible translations.
The multitag format files also contain several other metadata in the form of the following codes:

- If a title or subtitle is part of the original text, it will have the `TITLE` code. Otherwise, it functions as a locus marker.

- Old Hungarian codices often contain parts in another language (mostly Latin). If the foreign word is provided with some kind of Hungarian inflection, i.e. it functions as a standard part of the Old Hungarian language, it will be normalized and morphologically analyzed as usual. However, if it is not inflected, it is only a foreign word wedged between Hungarian words, it will have the `LANG{latin}` code and will not be normalized and analyzed.

- The scribe's corrections in the original text material are also marked by codes: supplementary addition (`ADD`), cancellation (`STRIKE`), failed, but not cancelled word (`FAIL`), fragmentary word (`FRAG`).

5 The Corpus Query Tool

The effective retrieval of relevant information is fundamental for linguistic research. For this purpose, we have constructed a publicly available query interface (`http://ohc.nytud.hu`), which offers the user several features that greatly facilitate the linguistic analysis of large amounts of linguistic data.[4]

Text files in multitag format are converted into XML files, which are then validated, thereby checking the consistency of the database. These validated XML files are the suitable input for the Emdros corpus query engine (Petersen, 2004), on which we have built the query interface.

A good corpus query tool has to be able to formalize sophisticated linguistically relevant queries. Such queries often refer to different levels of language information contained in the corpus. Therefore, our corpus contains all of the linguistically relevant levels of language data (see Table 2), and the query interface allows the user to refer to these levels even simultaneously. The presentation of corpus results is independent of the query, in the sense that text processing levels different from the query can also be displayed.

The corpus query interface allows the user to specify a query with the help of easy-to-use buttons and pop-up menus. The query is then formalized in the query language of Emdros, which can be edited for enabling more sophisticated queries.

Figure 1 shows a sample part of the result page of a query in concordance format. We submitted a search for the normalized version of the Hungarian

Figure 1: A sample part of the result page in concordance format.

word *föld* 'ground, earth'. In corcondance format, one result is one sentence, but the context can be broadened in a 5-sentence window. Above each result, there is a marker containing the name of the document in which the sample is found, the locus marker inside the document, and the unique identifier of the token. Each result is displayed in a tabular form: the original orthographic form in green, the normalized form in black, and the morphological information (lemma and analysis) in grey. The requested word is always highlighted with boldface setting.

There is another feature of the corpus query tool in concordance format: the user can add more queries to the already submitted one in an n-word window, where n can be specified by the user. This provides the possibility to search several morphosyntactic patterns in sentences. To illustrate the usefulness of this feature, we present a real research question which is relevant for Hungarian in a diachronic perspective.

Modern Hungarian makes extensive use of the definite article, but in Old Hungarian, the definite article appears only in constructions where the referent of the noun phrase is not anchored in another way. Possessor expressions represent one instance of such constructions, where the definite article is an obligatory element of the possessive construction with a dative-marked possessor in Modern Hungarian, e.g. in (7), but is absent in Old Hungarian, e.g. in (8).

(7)
az ember-ek-nek **a** fia-i-val
the man-PL-DAT the son.POSS-PL-INS
'with the sons of the people'

(8)
embereknek fÿaÿual
∅ ember-ek-nek **∅** fia-i-val
man-PL-DAT son.POSS-PL-INS
'with the sons of the people' (Könyvecse 4v)

How can we check such a linguistic hypothesis against a corpus, that is, how can we search for something which is absent? In this case, we can submit a query on the morphologically analyzed part of the corpus by using the morphosyntactic annotation and the feature of context addition. Here we want to find a dative-marked possessor directly followed by a possessive-marked noun. We can formalize this query in the following way:

```
[W FOCUS w_6e ~ 'Dat_gen\)\)$']
[W FOCUS w_6e ~ '^6e\(\(N.Px']
```

In this query, we do not allow any other elements to stand between the possessor and the possessive. This query resulted in more than 2.000 hits.

We can also formalize the case when one additional element is allowed to stand between them, either a determiner or an adjective:

```
[W FOCUS w_6e ~ 'Dat_gen\)\)$']
.. BETWEEN 1 AND 1
[W FOCUS w_6e ~ '^6e\(\(N.Px']
```

After investigating the results, we found that most of the results contain an adjective before the noun, not an article. To reduce the number of the results to the relevant hits where a possessor is directly followed by a definite article and a possessive-marked noun, we have to use the normalized level of the corpus. Since definite and non-definite articles are not distinguished on the morphosyntactic annotation level, but the definite article has only two forms (*a* and *az*), the formalization of the word between the possessor and the possessive can be specified on the normalization level in the following way:

```
[W FOCUS w_6e ~ 'Dat_gen\)\)$']
[W FOCUS w_4 ~ '^4\(\(az?\)\)$']
[W FOCUS w_6e ~ '^6e\(\(N.Px']
```

Since this query resulted in only one hit, this is a good indicator of the fact that the definite article was not used in dative-marked possessor constructions in Old Hungarian.
Besides the concordance format, the corpus query tool also allows the user to ask for a frequency list. This service is not only for listing all possible variants of each word in the texts, but it calculates the total amount of

each original orthographic form and the normalized word form itself. As can be seen in Figure 2, the corpus contains 596 occurrences of the word *föld* 'ground, earth'. (Recall that only part of the entire linguistic material has been normalized, i.e. the result can only be interpreted on the normalized subcorpus.)

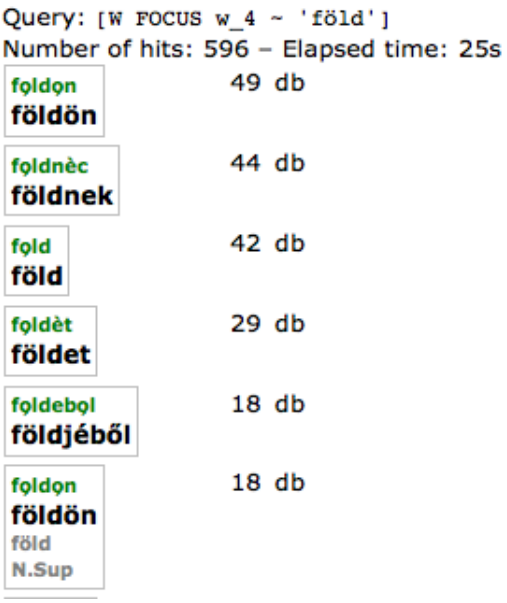

Figure 2: A sample part of the result page in frequency list format.

6 Final Remarks

The method of gaining empirical linguistic data from searchable historical corpora in order to describe and reconstruct diachronic changes has recently become extremely popular and can be considered as one of the mainstream linguistic trends. In view of that, our efforts made so far for developing an open access, digitized historical corpus for Hungarian are

state-of-the-art efforts. Moreover, since Hungarian is the longest documented language of the Uralic language family, the searchable corpus and the theoretical results based on it can be useful not only for the study of Hungarian but also for comparative Uralic studies. Additionally, both the corpus and the theoretical findings are extremely useful for researchers who are interested in approaching language change from a perspective broader than the well-established Indo-European one.

However, one can never say that a corpus is ready; it can be extended both in a horizontal and in a vertical dimension to be the source for wider and deeper theoretical investigations in the future. As for the horizontal dimension, we plan to expand the database by adding Middle Hungarian sources, mainly Bible translations. These texts can then be compared to the Old Hungarian records of similar content in a particularly efficient way to observe and track the gradual changes in the Hungarian language. Developing the database in a vertical way includes the linguistic annotation (normalization, morphological analysis and disambiguation) of the Old Hungarian texts which have not been normalized yet.

The more support linguists can get from the historical corpus, the larger quantity of results can be expected. One of the prospective extensions is making the corpus bilingual. The addition of an English vocabulary to the already normalized parts of the corpus, would make the historical corpus accessible to non-Hungarian users as well in the future.

Notes

[1] http://finereader.abbyy.com/

[2] http://www.unicode.org

[3] St. Stephen Association Bible translation (http://szentiras.hu/SZIT)

[4] The query interface has been built in our project web site which is available via the URL http://oldhungariancorpus.nytud.hu. The Old Hungarian texts in their original orthographic form and their normalized versions are also available on this web site.

References

Galves, C. and Britto, H. (2002). The Tycho Brahe Corpus of Historical Portuguese. Online publication.

Halácsy, P., Kornai, A., and Oravecz, Cs. (2007). HunPos – an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 209–212, Prague, Czech Republic. Association for Computational Linguistics.

Jakab, L. (2002). *A Jókai-kódex mint nyelvi emlék szótárszerű feldolgozásban [The Jókai Codex as a language record in a dictionary format].* Számítógépes Nyelvtörténeti Adattár 10. [Computational Database for Historical Linguistics 10.]. Department of Hungarian Linguistics, University of Debrecen, Debrecen.

Jakab, L. and Kiss, A. (1994). *A Guary-kódex ábécérendes adattára [The alphabetical database of the Guary Codex].* Számítógépes Nyelvtörténeti Adattár 6. [Computational Database for Historical Linguistics 6.]. Department of Hungarian Linguistics, University of Debrecen, Debrecen.

Jakab, L. and Kiss, A. (2001). *A Festetics-kódex ábécérendes adattára [The alphabetical database of the Festetics Codex].* Számítógépes Nyelvtörténeti Adattár 9. [Computational Database for Historical Linguistics 9.]. Department of Hungarian Linguistics, University of Debrecen, Debrecen.

Kniezsa, I. (1952). *Helyesírásunk története a könyvnyomtatás koráig [The*

history of our orthography until the age of the printing press]. Akadémiai Kiadó, Budapest.

Kroch, A. and Taylor, A. (2000). The Penn-Helsinki Parsed Corpus of Middle English (PPCME2). CD-ROM.

McEnery, T. (2004). Corpus Linguistics. In Mitkov, R., editor, *The Oxford Handbook of Computational Linguistics*, page 449. Oxford University Press, New York.

McEnery, T. and Hardie, A. (2003). *Lancester Newsbooks Corpus*.

Petersen, U. (2004). Emdros – a text database engine for analyzed or annotated text. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1190–1193, Geneva, Switzerland.

Prószéky, G. and Kis, B. (1999). A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 261–268, College Park, Maryland, USA.

Rayson, P., Archer, D., Baron, A., Culpeper, J., and Smith, N. (2007). Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In *Proceedings of the Corpus Linguistics Conference (CL2007)*, UK. University of Birmingham.

Thomas, P. W., Smith, D. M., and Luft, D. (2007). Rhyddiaith Gymraeg 1350-1425.

Volf, G. (1874). *Nyelvemléktár I. [Repository of language records I.]*. The Publishing House of the Hungarian Academy of Sciences, Budapest.

Volk, M., Marek, T., and Sennrich, R. (2010). Reducing OCR Errors by Combining Two OCR Systems. In *Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010)*, Lisbon, Portugal. Faculty of Science, University of Lisbon.