Gradual expansion in the use of the definite article – Checking a theory against the Old Hungarian Corpus

Barbara Egedi, Eszter Simon

Research Institute for Linguistics, Hungarian Academy of Sciences

University of Oslo,
14th June 2012

Outline

1. Introduction
2. The definite article in Old Hungarian
3. The absence of article in definite contexts
4. The definite article in Old Hungarian – spreading
5. Acquisition of source data
6. Corpus building workflow
7. Corpus architecture
8. Corpus query tool

The project

Hungarian **G**enerative **D**iachronic **S**yntax (HGDS)

supported by the Hungarian Scientific Research Fund
(OTKA No. 78074)
April 2009 – March 2013
Research Institute for Linguistics, Hungarian Academy of Sciences

The aims of the project

- to digitise all the Old Hungarian records and some selected texts from the Middle Hungarian Period, and to build an online searchable historical language corpus
- to reconstruct the syntax of different synchronic systems, and to examine and model the syntactic changes

The aim of the present talk

- to explore and present the possibilities and the limits of checking a linguistic hypothesis against a larger amount of data

Data:
Five (at least) normalised codices from the Old Hungarian Corpus

The hypothesis to be checked:
the expansion in the use of the article

Historical language stages of Hungarian

Proto-Hungarian		1000 BC – 896 AD
Old Hungarian	Early Old Hungarian	896 – 1370
	Late Old Hungarian	1370 – 1526
Middle Hungarian		1526 – 1772
Modern Hungarian		1772 – present day

The definite article in Old Hungarian – the data

Modern Hungarian

(1)

a	szerzetes-nek	***a***	könyv-e
the	monk-DAT	*the*	book-POSS

'the book of the monk'

(2)

a	te	könyv-ed
the	your	book-POSS.2SG

'your book'

Old Hungarian

(3)

a(z)	szerzetes-nek	∅	könyv-e
the	monk-DAT		book-POSS

'the book of the monk'

(4)

∅	te	könyv-ed
	your	book-POSS.2SG

'your book'

The definite article in Old Hungarian – the data

Modern Hungarian

(5)	ez/az	***a***	könyv
	this/that	*the*	book
	'this/that book'		

Old Hungarian

(6)	e(z)/a(z)	∅	könyv
	this/that		book
	'this/that book'		

(7)	a.	az	kapu	b.	az	kapu
		the	gate		*that*	gate

The definite article in Old Hungarian – the hypothesis

Regular appearance of this early determiner in semantic and pragmatic contexts where only an article can appear (e.g. *larger situational use, associative-anaphoric use*) ⇒ true article

Investigation by a manual search on a closed uniform text:
the Gospel of Matthew in the *Munich Codex*
(the first half of the Late Old Hungarian period)

- to classify the articleless noun phrases with a definite interpretation
- to understand why the article is still absent

The definite article in Old Hungarian – the hypothesis

The early definite article appears only in the constructions where the referent of the noun phrase is not anchored in another way, thus absent:

- with proper names and with a group of lexemes that describe entities with a prototypically unique referent
- with demonstratives
- with possessor expressions
- in case of a generic reading of the noun phrase

The absence of article in definite contexts (Munich Codex)

Proper names and nouns with unique referent

nouns with special lexical properties → inherently referential

a group of lexemes describing entities with a prototypically unique referent: *god, lord, father* (referring to God), *devil, king, queen, heaven* etc.

(8)	és	Ø	*atyá-t*	senki	sem	esmerte	hanemcsak	Ø	*fiú-t*
	and		father-ACC	nobody	not	knew	but		son-ACC

'neither knoweth any men the Father, save the Son' [Matt 11:27]

The absence of article in definite contexts (Munich Codex)

Modified by a demonstrative

directly accessible reference → necessarily definite

(11)	*Az*	*napok-ban*	jövő	Jánus	baptista
	that	days-INE	came	John	Baptiste

'In those days John the Baptist came' [Matt 3:1]
in Latin: *in diebus illis*

New pattern: *determiner doubling* (only from Middle Hungarian)

(5')	az	***a***	könyv
	that	*the*	book

'that book'

The absence of article in definite contexts (Munich Codex)

Possessive structures: pronominal possessors

the referent of the possessed noun is identified via its relation to the referent of the possessor → prototypically definite

(14) És elhozaték egy tálnyér-on ∅ *ő feje*
and was.brought a platter-SUP his head-POSS.3SG
'And his head was brought on a platter' [Matt 14:11]

The absence of article in definite contexts (Munich Codex)

Possessive structures: nominal possessors

no determiner on the head noun

(15)	az	gyermek-nek	∅	lelk-é-t
	the	child-DAT		soul-POSS-ACC

'the soul of the child' [Matt 2:20]

(16)	az	papok	∅	fejedelm-i-hez
	the	priests		chief-POSS.PL-ALL

'to the chiefs of priests' [Matt 26:57]

The absence of article in definite contexts (Munich Codex)

Generic reading: without article ↔ individual reading: with article

(21)
Tahát felkelvén parancsola ***az*** *szelek-nek* és ***az*** *tenger-nek* (...)
so up.getting commanded *the* winds-DAT and *the* sea-DAT
'*So he got up and commanded the winds and the sea (...)*

Bizony az emberek csudálkodnak vala, mondván: Minemő ez,
verily the men were.amazed AUX saying what.kind this
The men were amazed, saying: "What kind (of man) is this,

mert ∅ *szelek* és ∅ *tenger* engednek neki?
that winds and sea obey.they to.him
that the winds and the sea obey him!"' [Matt 8:26-27]

The definite article in Old Hungarian – spreading

In the Middle Hungarian period (from the 16th century onward) the definite article appears in new contexts:

- co-occuring with demonstratives
- preceding a possessed noun with dative-marked possessor

Manual checking of the NT *loci* in a later Old Hungarian ms. Codex Jordánszky (1516-1519) → expansion in article use with generic NPs and before possessive pronouns

Aims:

- to demonstrate the proportional increase in the use of the article *within* the Old Hungarian period
- to find out in which context(s) it took place earlier

The definite article in Old Hungarian – spreading

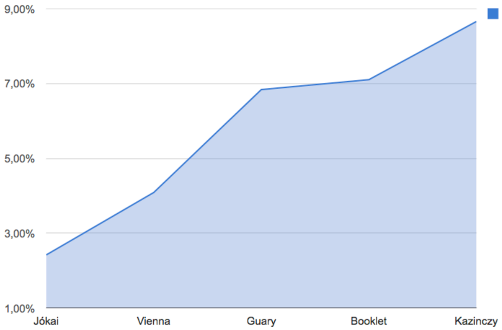


Figure: The proportion of definite articles in five Old Hungarian codices

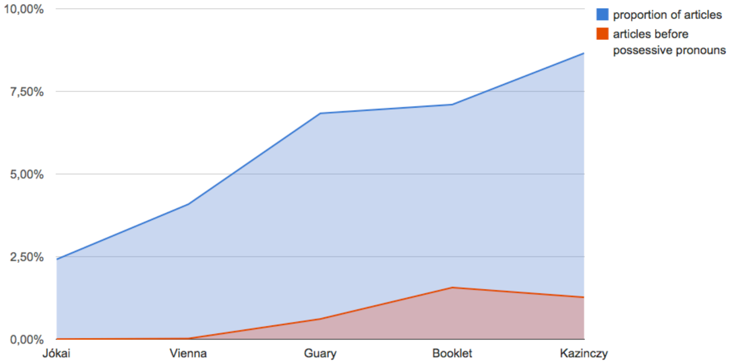
In a modern Bible translation: 11,12%

The definite article in Old Hungarian – spreading

Table: Increase in article use in specific contexts

Context	Method	Increase in article use
Nouns with unique referent	One by one checking of lexical items in the corpus	slightly; inconsistent results
Demonstratives	Automatic query	none
Possessives (pron.)	Automatic query	√
Possessives (nom.)	Automatic query	none/minimal
Generics	Not possible automatically	???

The definite article in Old Hungarian – spreading



Corpus building: the texts

no antecedent → building a corpus of Old Hungarian codices is a pioneering effort

the amount of texts to process:

- 48 codices
- 27 shorter texts

all together approx. 2 million tokens, from which about 1.5 million are already searchable

The acquisition of source data

a part of source data has already been converted into some electronic format

a significant part was only available in print:

```
for codex in codices:
    if codex is short:
         typewriting
    else:
         scanOCRmanualcheck
```

Scan&OCR

1. scanning in high resolution – high enough to OCR
2. key aspect of an OCR software: training system to make it able to deal with other languages → Abby FineReader 9.0 Professional edition

special characters cause problems

Heterogeneous orthography

- the Hungarian writing system evolved from the need to translate Latin religious texts, but the adaptation of the Latin alphabet to Hungarian posed several problems
- there are Hungarian phonemes which do not exist in Latin, so new characters were needed to represent them
- sound-letter correspondences vary a lot even within a single text sample, e.g. *Vylag uilaga* [világ világa]
- one letter could stand for multiple sounds, e.g. *zerzete zerent* [szerzete szerint]
- some letters can refer to vowels and consonants as well, e.g. the letters *u,v,w* were used to denote the *u,ú,ü,ű,v* sounds

Representing Hungarian sounds

representing the Hungarian sounds which do not exist in Latin:

1. without diacritics: combination of more letters, e.g. *cs [tʃ]* $\longrightarrow$ *ch* $\sim$ *cz* $\sim$ *chy* $\sim$ *chi* $\sim$ *cy*
2. with diacritics: a similar letter with diacritics, e.g. *cs* $\longrightarrow$ *ć* $\sim$ L $\sim$ L’
3. mixture, e.g. *cs* $\longrightarrow$ *ch* $\sim$ *chy* $\sim$ *cyh* $\sim$ *c* $\sim$ *chi* $\sim$ *ch’* $\sim$ *cz* $\sim$ *ts* $\sim$ *ć* $\sim$ L $\sim$ L’ $\sim$ L*h* $\sim$ L*z*

Maintaining and representing the original spelling: Unicode

- international standard
- consistent encoding for the most of the world's writing systems
- the various accented and multi-accented characters are properly handled, e.g. $e + \text{‘} + \bar{\ } = \bar{\grave{e}}$

standard Unicode characters in UTF-8 encoding

uniformity is a basic requirement of asking queries in the whole corpus
BUT: there are some Old Hungarian characters which are not present in Unicode → replacement character, e.g. $\mathrm{L} \longrightarrow \check{c}$

Normalisation

original	normalised
a varoſba	a városba
ahazi	a házi
annėphėz	a néphez
az arpak	az árpák
ānėp	a nép
a” tew	a tű
a’ nyaar	a nyár
a · mendenhato	a Mindenható

orthographic variants of the same lexical items must be neutralised and converted into Modern Hungarian spelling

Morphological analysis and disambiguation

normalised wordform → automatic morphological analysis → automatic disambiguation → manual check

Humor ('**H**igh speed **U**nification **MOR**phology'): originally developed for Modern Hungarian → extended version of the lexicon and the morphological rules for Old Hungarian has been created

hunpos: statistical POS-tagger → requires a quite large amount of manually disambiguated Old Hungarian texts as a training corpus

Text processing levels

six levels and five tasks can be distinguished throughout the processing of the texts

(1) scanned codex
→ OCR
(2) raw OCR result
→ *manual* correction
(3) original spelling
→ *manual* normalisation
(4) normalised form
→ *automatic* morphological analysis
(5) lemmatised and POS-tagged words
→ *semi-automatic* disambiguation
(6) disambiguated corpus

Corpus architecture

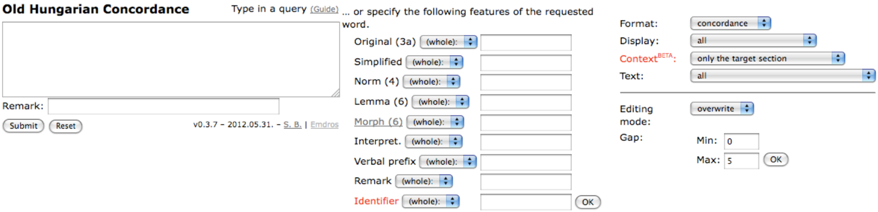
page	line	orig	norm	lemma	analysis
001	01	Mÿ	mi	mi	Pro.Nom_Gen
001	01	vronknac	urunknak	úr	N.PxP1.Dat_Gen
001	01	iesus	Jézus	Jézus	N:P.Nom
001	01	cristusnac	Krisztusnak	Krisztus	N:P.Dat_Gen
001	01	gyczeretÿre	dicséretire	dicséret	N.PxS3=i.Sub
001	02	es	és	és	C
001	02	gyczewsegere	dicsőségére	dicsőség	N.PxS3.Sub
001	02	es	és	és	C
001	02	my	mi	mi	Pro.Nom_Gen
001	02	atyancnak	atyánknak	atya	N.PxP1.Dat_Gen
001	03	bodog	boldog	boldog	Adj
001	03	ferencznek	Ferencnek	Ferenc	N:P.Dat_Gen

Searching on every level

- our corpus contains all the text processing levels, and the query interface allows the user to refer to these levels even simultaneously
- the presentation of corpus results is independent of the query: different levels can be used in display and in the query

http://ohc.nytud.hu/

User interface



Old Hungarian Concordance

Type in a query (Guide) ... or specify the following features of the requested word.

Remark:

v0.3.7 – 2012.05.31. – S. B. | Emdros

Original (3a) (whole):

Simplified (whole):

Norm (4) (whole):

Lemma (6) (whole):

Morph (6) (whole):

Interpret. (whole):

Verbal prefix (whole):

Remark (whole):

Identifier (whole):

Format: concordance

Display: all

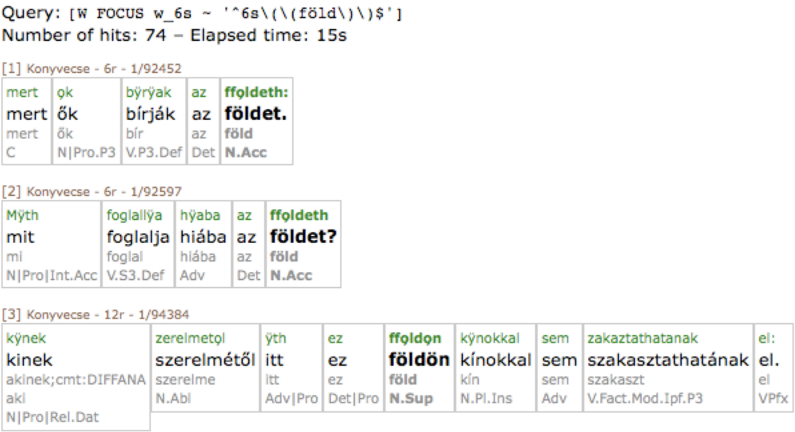
Context^{BETA}: only the target section

Text: all

Editing mode: overwrite

Gap: Min: 0 Max: 5

Corpus query result



Query: `[W FOCUS w_6s ~ '^6s\(\(föld\)\)$']`
Number of hits: 74 – Elapsed time: 15s

[1] Konyvecse - 6r - 1/92452

mert	ǫk	bÿrÿak	az	**ffǫldeth:**
mert	ők	bírják	az	**földet.**
mert	ők	bír	az	**föld**
C	N|Pro.P3	V.P3.Def	Det	**N.Acc**

[2] Konyvecse - 6r - 1/92597

Mÿth	foglallÿa	hÿaba	az	**ffǫldeth**
mit	foglalja	hiába	az	**földet?**
mi	foglal	hiába	az	**föld**
N|Pro|Int.Acc	V.S3.Def	Adv	Det	**N.Acc**

[3] Konyvecse - 12r - 1/94384

kÿnek	zerelmetǫl	ÿth	ez	**ffǫldǫn**	kÿnokkal	sem	zakaztathatanak	el:
kinek	szerelmétől	itt	ez	**földön**	kínokkal	sem	szakaszthatának	el.
akinek;cmt:DIFFANA	szerelme	itt	ez	**föld**	kín	sem	szakaszt	el
aki	N.Abl	Adv|Pro	Det|Pro	**N.Sup**	N.Pl.Ins	Adv	V.Fact.Mod.Ipf.P3	VPfx
N|Pro|Rel.Dat								

Contributors

Júlia Bácskai-Atkári
Sylvia Blaho
Judit Farkas
Mátyás Gerőcs
Veronika Hegedűs
Andrea Kacskovics-Reményi
Gergely Kántor
Eszter Mihály
Iván Mittelholcz

Attila Novák
Csaba Oravecz
Márta Peredy
Katalin Pólya
Bálint Sass
Dániel Szeredi
Johanna Szőke
Orsolya Tánczos
Ildikó Tóth

Thank you for your attention!

emails: `egedib@yahoo.com`
`simon.eszter@nytud.mta.hu`

corpus query tool: `http://ohc.nytud.hu`
downloadable publications:
`http://www.nytud.hu/oszt/elmnyelv/mgtsz.html`