1. Natural language processing

2. ELTE Poetry Corpus – automatic annotation of poems in TEI XML

Péter Horváth
ELTE BTK, Centre for Digital Humanities

Natural language processing

- NLP = Natural language processing
- The computational, automatic linguistic analysis of written and spoken texts.
- The computational generating of written and spoken texts.

The main steps of natural language processing

1. Tokenization
2. Lemmatization, part of speech tagging, morphosyntactic tagging
3. Syntactic tagging

- The order of the steps is always the same.

Tokenization

- Token: a word or a punctuation mark.
- Tokenization: segmentation of texts into sentences and tokens.
- Tokenizer: a program tokenizing texts automatically.
- Token is the basic unit of every further automatic linguistic analysis.

Output of tokenization

- *The youngest girl went to the supermarket.*

tokens:

1. The
2. youngest
3. girl
4. went
5. to
6. the
7. supermarket
8. .

Second step

- Lemmatization
- Part of speech tagging
- Morphosyntactic tagging

Lemmatization

- Lemma: the dictionary form of a word.
- e.g. went, go, going, gone, goes → go
- Lemmatization: reducing of tokens to their dictionary form.
- Lemmatizer: a program lemmatizing tokens automatically.

Output of lemmatization

- *The youngest girl went to the supermarket.*

lemmas:
1. the
2. young
3. girl
4. go
5. to
6. the
7. supermarket
8. .

Why is lemmatization useful?

- Queries are simple and more efficient.
  - You do not have to type every word form into the query field.
  - e.g. input lemma: go → output tokens: go, goes, going, gone, went
  - e.g. input lemma: ad → output tokens: adok, adom, adtam, adsz, adod, adtál, adtad, ad, adja, adott, adta, adunk, adjuk, adtunk, adtuk, adtok, adjátok, adtatok, adtátok, adnak, adják, adtak, adták, adjak, adjam, adjál, add, adjad, adjon …
- It is possible to measure the size of vocablulary.

Part of speech tagging

- Every word has a part of speech (e.g. noun, verb, adjective)
- POS tagging: classification of tokens into parts of speech.
- POS tagger: a program assigning part of speech to tokens automatically.

Output of part of speech tagging

- *The youngest girl went to the supermarket.*

POS:
1. DET
2. ADJ
3. NOUN
4. VERB
5. ADP
6. DET
7. NOUN
8. PUNCT

Morphosyntactic tagging

- Words have other grammatical properties besides part of speech: morphosyntactic features and syntactic function.
- Morphosyntactic features:
  - noun: number, case
  - verb: number, person, tense, mood, definitness
- Morphosyntactic tagging: the automatic analysis of the morphosyntactic features of words.
- Output of morphosyntactic taggers: lists of feature-value pairs.

Output of morphosyntactic tagging I.

- *The youngest girl went to the supermarket.*

  feature-value pairs:

1. The: Definite=Def|PronType=Art
2. youngest: Degree=Sup
3. girl Number=Sing
4. went Mood=Ind|Tense=Past|VerbForm=Fin
5. to _
6. the Definite=Def|PronType=Art
7. supermarket Number=Sing
8. . _

Output of morphosyntactic tagging II.

- *A legfiatalabb lány elment a boltba.*

Feature-value pairs:

1. A Definite=Def|PronType=Art
2. legfiatalabb Case=Nom|Degree=Sup|Number=Sing|
   Number[psed]=None|Number[psor]=None|
   Person[psor]=None
3. lány Case=Nom|Number=Sing|Number[psed]=None
   None|Number[psor]=None|Person[psor]=None
4. elment Definite=Ind|Mood=Ind|Number=Sing|Person=3|
   Tense=Past|VerbForm=Fin|Voice=Act
5. a Definite=Def|PronType=Art
6. boltba Case=Ill|Number=Sing|Number[psed]=None|
   Number[psor]=None|Person[psor]=None

Syntactic tagging

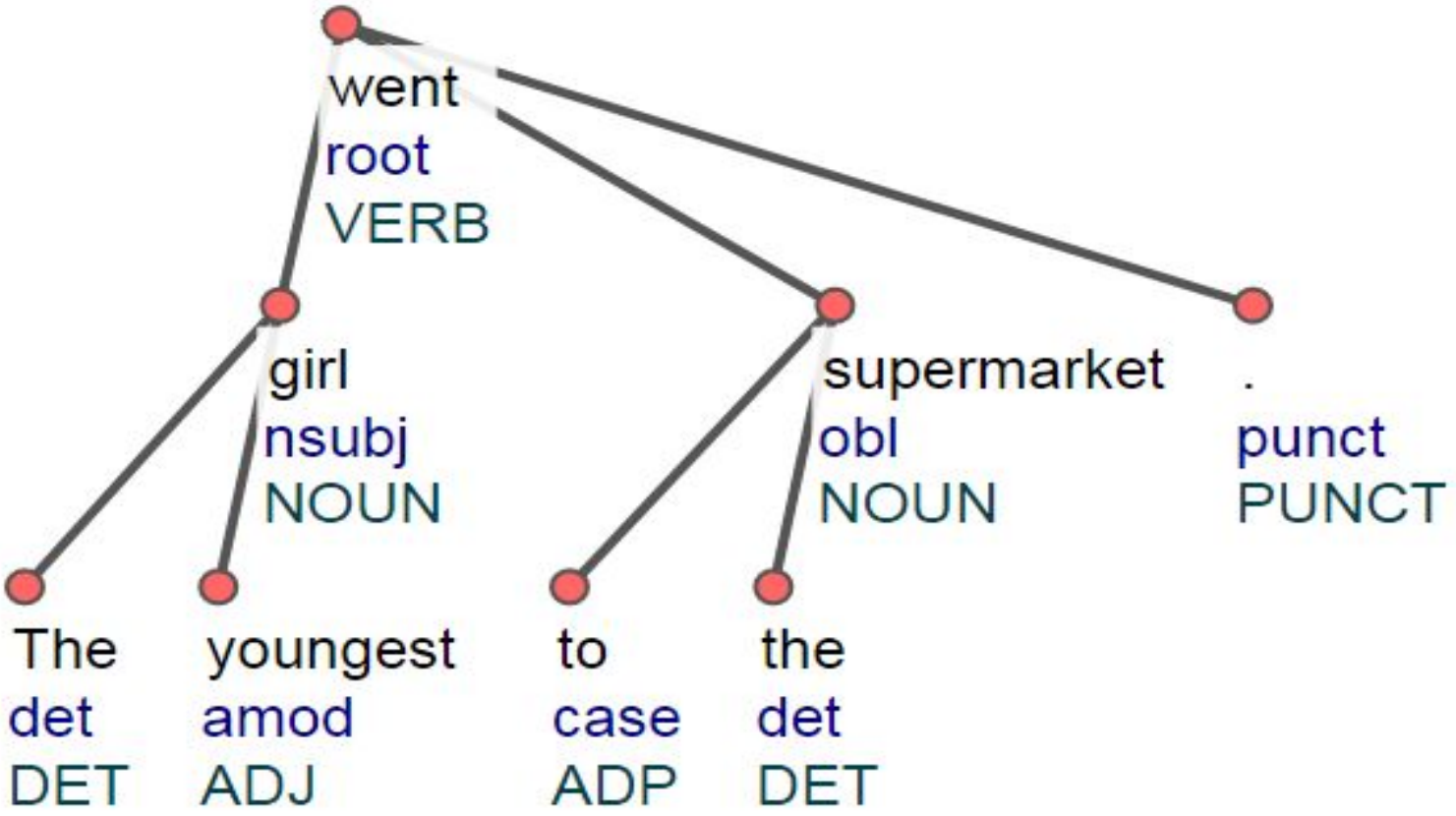
- The analysis of the syntactic structure of sentences and the syntactic function of words.
- Output of syntactic taggers:
  - The labels of syntactic functions (e.g. subject, object)
  - The representations of the syntactic structures of sentences:
    - Dependency trees
    - Constituency trees

Dependency trees

- Graph: nodes, and links between nodes.
  - Nodes: the words of the sentence.
  - Links: the dependency relations between words.
- The structural center (root node) of the sentence is the verb, the other words are connected to the verb directly or indirectly.
- Types of nodes:
  - Heads: superordinated nodes
  - Dependents: subordinated node
- A word can be a head and a dependent at the same time.

Dependency tree

- *The youngest girl went to the supermarket.*

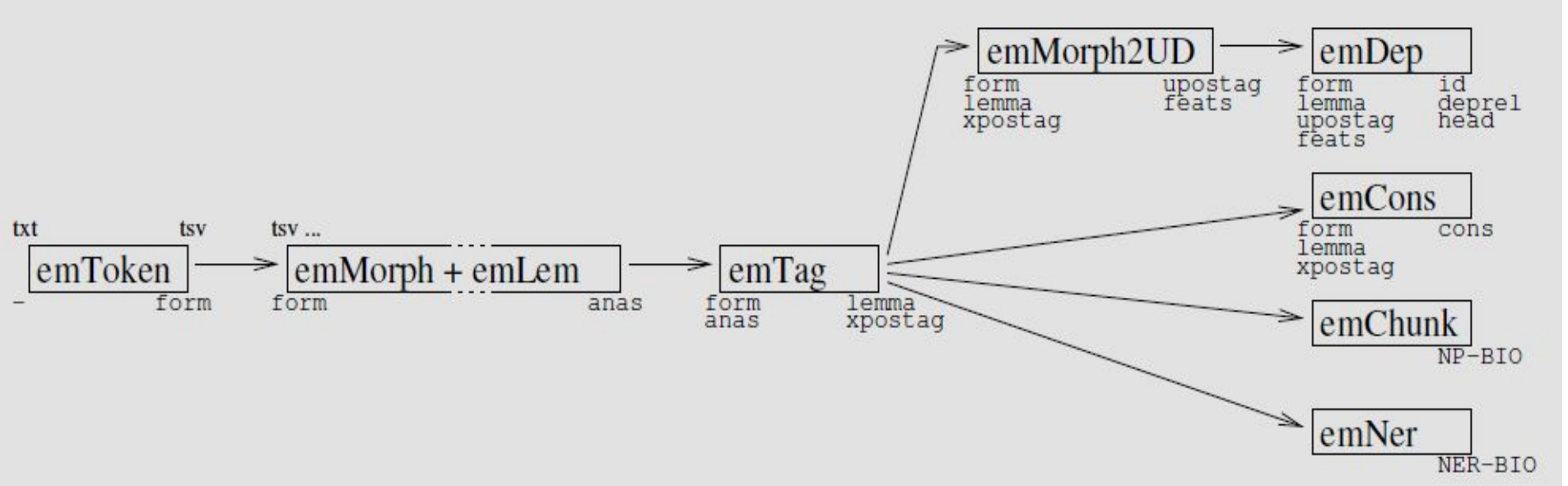


Output of syntactic tagging

- *The youngest girl went to the supermarket.*

1.	The	3	
2.	youngest		3
3.	girl	4	
4.	went	0	
5.	to	7	
6.	the	7	
7.	supermarket		4
8.	.	4	

E-magyar – NLP tool for Hungarian



- Váradi–Simon–Sass et al 2017, Mittelholcz 2017, Novák–Rebrus–Ludányi 2017, Indig–Sass–Simon et al 2019)
- https://e-magyar.hu/hu/
- https://github.com/dlt-rilmta/emtsv

The change of document format in the process of corpus creation

XML, HTML, RTF stb. → TXT → NLP tool → TSV → XML

TXT: text

TSV: tabulator separated values

XML: extensible markup language

TSV output of NLP tools

Id	Form	Lemma	Postag	Feats	Head
1	The	the	DET	Definite=Def|PronType=Art	3
2	youngest	young	ADJ	Degree=Sup	3
3	girl	girl	NOUN	Number=Sing	4
4	went	go	VERB	Mood=Ind|Tense=Past|VerbForm=Fin	0
5	to	to	ADP	_	7
6	the	the	DET	Definite=Def|PronType=Art	7
7	supermarket	supermarket	NOUN	Number=Sing	4
8	.	.	PUNCT	_	4

Converting TSV into TEI XML

```
<s>
    <w lemma="the" pos="DET" msd="Definite=Def|PronType=Art">The</w>
    <w lemma="young" pos="ADJ" msd="">youngest</w>
    <w lemma="girl" pos="NOUN" msd="Degree=Sup">girl</w>
    <w lemma="go" pos="VERB" msd="Number=Sing">went</w>
    <w lemma="to" pos="ADP" msd="Mood=Ind|Tense=Past|VerbForm=Fin">to</w>
    <w lemma="the" pos="DET" msd="Definite=Def|PronType=Art">the</w>
    <w lemma="supermarket" pos="NOUN" msd="Number=Sing">supermarket</w>
    <pc pos="PUNCT" join="left">.</pc>
</s>
```

Two further methods in NLP

- **Named entity recognition:** The automatic recognition and classification of proper names in texts.
- **Sentiment analysis:** The automatic analysis of emotions in texts.

An online NLP tool

- http://ufal.mff.cuni.cz/udpipe
- http://lindat.mff.cuni.cz/services/udpipe/
- Model: english-gum-ud-2.4-190531

ELTE Poetry Corpus

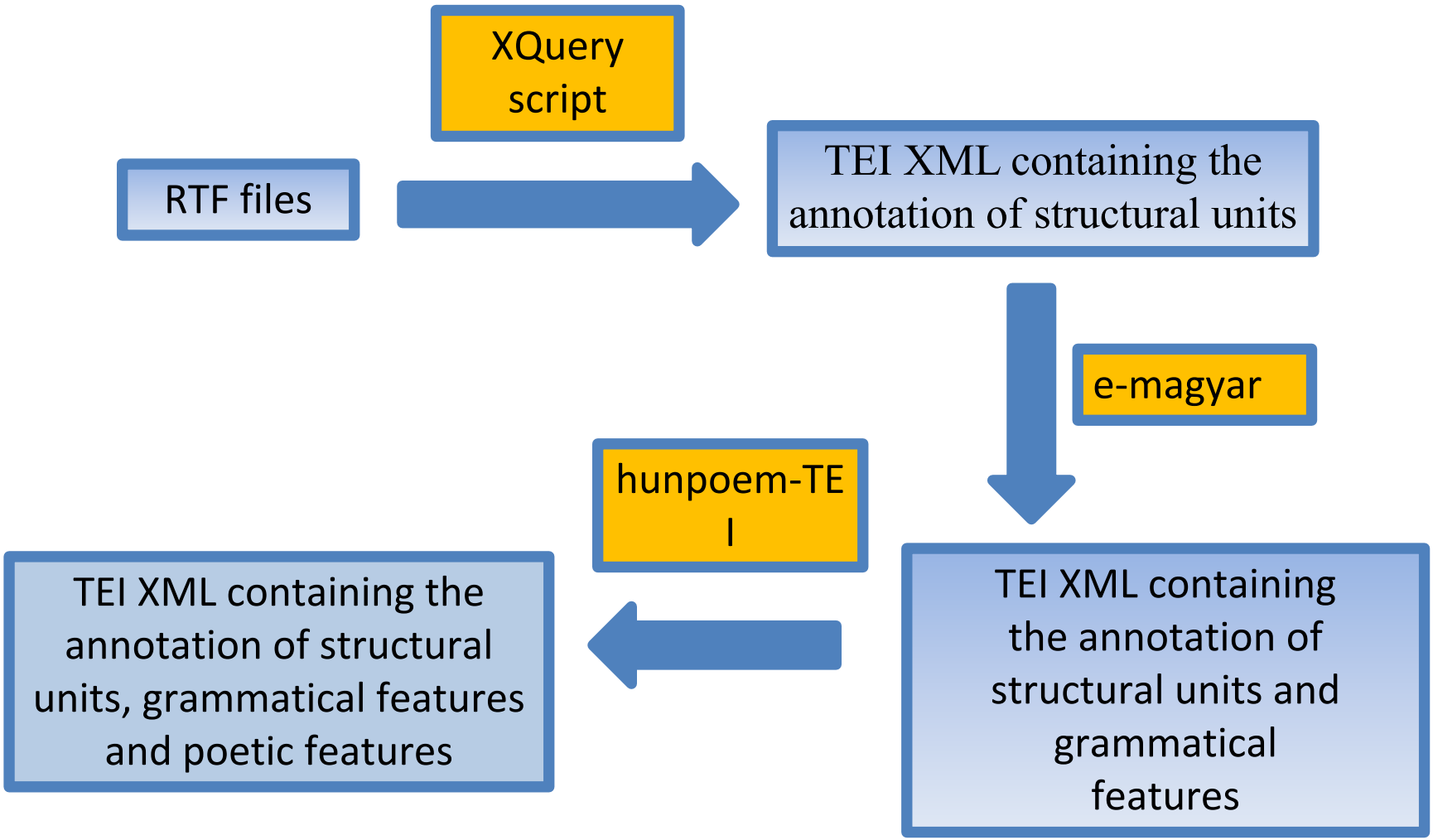
ELTE Poetry Corpus

- **Goal:** The aim of the project is to create an annotated corpus containing the complete poems of numerous Hungarian poets from before the mid-20th century.
- **Source of the corpus:** digital documents (mostly RTF files) from the Hungarian Electronic Library (https://mek.oszk.hu/)
- **Content:** poems, metadata, annotated properties
- **Annotated properties:** structural elements, grammatical features, poetic features
- **Format of the corpus:** TEI XML containing the tags and attributes of the annotated properties

Tools of the annotation

- **XQuery script**: annotation of structural units
  - Title, stanzas, lines
- **E-magyar**: annotation of grammatical features (Váradi–Simon–Sass et al. 2017, Novák–Rebrus–Ludányi 2017, Indig-Sass-Simon 2019)
  - Lemma, part of speech, morphosyntactic features
- **Python program**: annotation of poetic features
  - Rhyme patterns, rhyme pairs, rhythm patterns, alliterations, phonological features

The main steps of the annotation process:



Automatic analysis of poetic features

Rhyme patterns, rhyme pairs:

aabb; abcb; aabbccdd [meghalt, megcsalt]

Rhythm patterns:

„Látjátok feleim, egyszerre meghalt”: 1 1 1 0 0 0 1 1 0 1 1

Alliterations:

„ … **m**inket **m**agunkra. **M**egcsalt” : a a a

„**n**em volt **n**agy”: a n a

„**b**ús **d**onna **b**arna **b**alkonon”: a n a a

Fonological structure:

látjátok: [C, VB2, C, C, VB2, C, VB1, C] low word

feleim: [C, VF1, C, VF1, VF1, C] high word

meghalt: [C, VF1, C, C, VB1, C, C] mixed word

Annotation of structural units in TEI XML

```
<div type="poem">
    <head>Húnyt szemmel...</head>
    <lg>
        <l>Húnyt szemmel bérceken futunk</l>
        <l>s mindig csodára vágy szivünk:</l>
        <l>a legjobb, amit nem tudunk,</l>
        <l>a legszebb, amit nem hiszünk.</l>
    </lg>
    <lg>
        <l>Az álmok síkos gyöngyeit</l>
        <l>szorítsd, ki únod a valót:</l>
        <l>hímezz belőlük</l>
        <l>fázó lelkedre gyöngyös takarót.</l>
    </lg>
</div>
```

Annotation of structural units, rhyme patterns and rhythm patterns in TEI XML

```
<div type="poem">
    <head>Húnyt szemmel...</head>
    <lg rhyme="abab">
        <l real="11110101">Húnyt szemmel bérceken futunk</l>
        <l real="11010101">s mindig csodára vágy szivünk:</l>
        <l real="01101101">a legjobb, amit nem tudunk,</l>
        <l real="01101101">a legszebb, amit nem hiszünk.</l>
    </lg>
    <lg rhyme="abcb">
        <l real="01111101">Az álmok síkos gyöngyeit</l>
        <l real="01010001">szorítsd, ki únod a valót:</l>
        <l real="11011">hímezz belőlük</l>
        <l real="1111011001">fázó lelkedre gyöngyös takarót.</l>
    </lg>
</div>
```

Annotation of the grammatical properties of words in TEI XML

```
<l real="11110101">
    <w xml:id="w1" lemma= "az" msd="[/Det|Art.Def]"
    pos="DET">az</w>
    <w xml:id="w2" lemma= "álom" msd="[/N][Pl][Nom]"
    pos="NOUN">álmok</w>
    <w xml:id="w3" lemma="síkos" msd="[/Adj][Nom]"
    pos="ADJ">síkos</w>
    <w xml:id="w4" lemma="gyöngy" msd="[/N][Pl.Poss.3Sg][Acc]"
    pos="NOUN">gyöngyeit</w>
</l>
```

Annotation of structural units, poetic features and grammatical features in TEI XML

```
<head>
   <w xml:id="w1" pos="Adj" lemma="halotti" msd="[/Adj][Nom]">Halotti</w>
   <w xml:id="w2" pos="N" lemma="beszéd" msd="[/N][Nom]">beszéd</w>
</head>
<lg xml:id="lg1" rhyme="aabbcddc">
   <l xml:id="l1" real="---uuu--u--">
      <w xml:id="w3" pos="V" lemma="lát" msd="[/V][Prs.Def.2Pl]">Látjátok</w>
      <w xml:id="w4" pos="N" lemma="fél" msd="[/N][Pl.Poss.1Sg][Nom]">feleim</w>
      <pc xml:id="pc1" pos="," join="left">,</pc>
      <w xml:id="w5" pos="Adv" lemma="egyszerre" msd="[/Adv]">egyszerre</w>
      <w xml:id="w6" pos="V" lemma="meghal" msd="[/V][Pst.NDef.3Sg]">meghalt</w>
   </l>
   <l xml:id="l2" real="--u---u-u--">
      <w xml:id="w7" pos="Cnj" lemma="és" msd="[/Cnj]">és</w>
      <w xml:id="w8" pos="Adv" lemma="itt" msd="[/Adv|Pro]">itt</w>
      <w xml:id="w9" pos="V" lemma="hagy" msd="[/V][Pst.NDef.3Sg]">hagyott</w>
      <w xml:id="w10" pos="N" lemma="mi" msd="[/N|Pro|Int][Poss.1Pl][Acc]">minket</w>
      <w xml:id="w11" pos="N" lemma="maga" msd="[/N|Pro][1Pl][Subl]">magunkra</w>
      <pc xml:id="pc2" pos="." join="left">.</pc>
      <w xml:id="w12" pos="V" lemma="megcsal" msd="[/V][Pst.NDef.3Sg]">Megcsalt</w>
      <pc xml:id="pc3" pos="." join="left">.</pc>
   </l>
```

Standoff annotation of rhyme pairs

```
<linkGrp>
    <link target= "#w6 #w12"/>
    <link target= "#w19 #w26"/>
    <link target= "#w29 #w36"/>
    <link target= "#w32 #w34"/>
    <link target= "#w40 #w45"/>
    <link target= "#w53 #w60"/>
    <link target= "#w60 #78"/>
    <link target= "#w67 #w72"/>
    <link target= "#w78 #83"/>
</linkGrp>
```

Corpus queries

Lists of the most frequent nouns

Endre Ady

1. élet (714)
2. lélek (405)
3. világ (362)
4. isten (339)
5. szem (308)
6. álom (298)
7. ember (279)
8. csók (259)
9. sors (254)
10. halál (245)

Mihály Babits

1. lélek (418)
2. ég (299)
3. föld (264)
4. világ (261)
5. nap (257)
6. élet (238)
7. szem (238)
8. isten (212)
9. szó (193)
10. szél (168)

Dezső Kosztolányi

1. szem (379)
2. élet (353)
3. arc (223)
4. ég (207)
5. föld (205)
6. lélek (202)
7. éj (201)
8. kéz (195)
9. fej (172)
10. álom (166)

Lists of the most frequent verbs

Endre Ady

1. van (1872)
2. lesz (651)
3. jön (519)
4. tud (449)
5. szeret (415)
6. él (391)
7. lát (365)
8. vár (315)
9. nincs (290)
10. akar (285)

Mihály Babits

1. van (919)
2. lesz (413)
3. lát (293)
4. tud (290)
5. jön (263)
6. él (219)
7. néz (207)
8. nincs (201)
9. áll (161)
10. mond (135)

Dezső Kosztolányi

1. van (770)
2. néz (329)
3. lát (305)
4. lesz (290)
5. tud (212)
6. megy (190)
7. sír (182)
8. szeret (174)
9. él (166)
10. mond (159)

Lists of the most frequent rhyme patterns

Endre Ady

1. abcb (957)
2. aba (376)
3. abcdb (225)
4. abca (208)
5. abb (203)
6. aa (198)
7. abab (191)
8. aabb (181)
9. abcbd (145)
10. abac (133)

Mihály Babits

1. abab (355)
2. aa (256)
3. aabb (224)
4. abcb (145)
5. aba (107)
6. abba (84)
7. abcd (84)
8. abc (80)
9. aaaa (75)
10. # (75)

Dezső Kosztolányi

1. abab (381)
2. abcb (308)
3. aba (280)
4. aab (276)
5. aa (262)
6. ab (116)
7. abba (104)
8. abb (91)
9. # (84)
10. aaa (70)

Lists of the most frequent rhythm patterns

Endre Ady

1. - - - - - - u - - (242)
2. u - u - - - u - - (240)
3. - - u - - - u - - (227)
4. - - u - u - u - - (218)
5. u - - - - - u - - (210)
6. - - u - - - u - (187)
7. - - - - u - u - - (186)
8. u - u - u - u - - (164)
9. u - - - - - u - (154)
10. u - u - - - u - (152)

Mihály Babits

1. u - u - u - u - (142)
2. - - u - - - u - (137)
3. - - u - u - u - (126)
4. u - - - u - u - (107)
5. u - u - - - u - (107)
6. - u - - - u - (103)
7. - - - - - - u - (98)
8. - - u - - - u - - (95)
9. u - u - u - (93)
10. u - - - - - u - (90)

Dezső Kosztolányi

1. - - u - (335)
2. - - u - - - u - u - - (324)
3. - - u - u - u - - (321)
4. u - u - (319)
5. - - u - - - u - - (318)
6. u - u - - - u - u - - (271)
7. u - u - - - u - - (269)
8. - - u - - - u - u - (252)
9. - - u - - - u - (248)
10. - - u - - - - - u - - (246)

Most frequent rhyme pairs in the poems of Attila József

	TOKEN	LEMMA	POS	Syll. Num.
1	ember–tenger (6)	maga–van (8)	noun–noun (1312)	2–3 (1144)
2	hamis–is (5)	ember–tenger (7)	noun–verb (1225)	2–2 (1088)
3	magam–van (5)	lélek–telek (6)	verb–verb (608)	2–4 (601)
4	is – mégis (4)	maga–szó (5)	adjective–noun (372)	3–3 (590)
5	engem–szivemben (4)	agy–van (5)	adverb–noun (331)	1–2 (377)
6	lelkem – telken (4)	hamis–is (5)	noun–pronoun (236)	3–4 (364)
7	végtelenbe–egyre (4)	hisz–visz (4)	adjective–verb (189)	1–3 (292)
8	agyunk–vagyunk (4)	elme–szerelem (4)	adverb–verb (153)	4–4 (192)
9	szerelem–velem (4)	is–mégis (4)	pronoun–verb (126)	1–1 (177)
10	költemény–én (4)	nyom–ottan (4)	adposition–noun (109)	2–5 (116)
11	mindörökre–szemegödre(4)	ellen–szellem (4)		
12		egyre–végtelen (4)		
13		ragyog–van (4)		
14		vágy–ágy (4)		
15		szerelem – én (4)		
16		költemény – én (4)		

Most frequent types of alliterations in the poems of Attila József

	STUCTURE	**POS – without non-alliterating word**	**POS – with non-alliterating word**
1	ana (2237)	adjective, noun (193)	determiner, **noun**, determiner (230)
2	aa (2155)	noun, noun (170)	noun, **determiner**, noun (76)
3	anana (157)	noun, verb (166)	verb, **determiner**, noun (74)
4	aaa (147)	verb, verb (97)	adjective, **noun**, verb (49)
5	aana (133)	pronoun, determiner (83)	noun, **adjective**, noun (43)
6	anaa (120)	determiner, noun (80)	determiner, **noun**, pronoun (38)
7	ananana (16)	verb, pronoun (70)	noun, **noun**, verb (34)
8	anaaa (10)	verb, noun (64)	adverb, **verb**, adverb (34)
9	ananaa (10)	adverb, verb (63)	verb, adverb, verb (32)
10	aaaa (10)	verb, adverb (58)	noun, **noun**, noun (30)

The most frequent rhyme patterns in „Hungarian poetry” (on the bases of the complete poems of 44 Hungarian poets)

1. aabb
2. abcb (= xaxa)
3. abab
4. aaaa
5. aa

Verb types and syllable types in Hungarian poetry (on the basis of the complete poems of 44 Hungarian poets)

1. Proportions of verbs in different numbers and persons:

1Sg: 14,6%	1Pl: 3,1%
2Sg: 10,7%	2Pl: 2,0%
3Sg: 59,6%	3Pl: 9,9%

2. Proportion of syllable types:

long syllables:

63,2 %

short syllables:

36,8%

Lists of poems by Endre Ady on the basis of value frequencies

1. List of poems with the highest proportion of verbs:

1. Az ágyam hívogat (52%)
2. Akarom: tisztán lássatok (40%)
3. A legjobb ember (40%)
4. Szeress engem, Istenem (36%)
5. Ki várni tud (35%)

2. List of poems with the highest proportion of long syllables:

1. Vén, bolond úr (81%)
2. Én szép világom… (81%)
3. Léda a kertben (79%)
4. Minden csak volt (78%)
5. Fuimus (78%)

3. List of poems with the highest proportion of high words:

1. Én kifelé megyek (71%)
2. Három őszi könnycsepp (63%)
3. Nem feleltem magamnak (63%)
4. Cseng az élet (62%)
5. A jégcsap-szívű ember (62%)

4. List of poems with the lowest proportion of high words:

1. [Révésznek – Ady] (24%)
2. Dedikáció (24%)
3. Csokonai Vitéz Mihály (25%)
4. Az értől az oceánig (25%)
5. A harcunkat megharcoltuk (26%)

Co-occurrences in the poems of Endre Ady:

1. verb frequency > 30% AND high word frequency > 60% :

Égő tűzben dideregve
Tiltakozni és akarni

2. noun frequency > 50% AND long syllable frequency > 70% :

Ruth és Delila

3. noun frequency < 20% AND long syllable frequency > 70% :

Ha holtan találkozunk
Az én bűnöm
Holnap is így

Co-occurrences in the poems of Endre Ady:

verb frequency > 30%
AND noun frequency < 30%
AND low word frequency > 50%
AND long syllable frequency > 65% :

Fekete hold éjszakáján

References

Indig Balázs – Sass Bálint – Simon Eszter – Mittelholcz Iván – Kundráth Péter – Vadász Noémi 2019. emtsv – egy formátum mind felett. In: Berend Gábor, Gosztolya Gábor, Vincze Veronika (szerk.): *XV. Magyar Számítógépes Nyelvészeti Konferencia*. Szegedi Tudományegyetem TTIK, Informatikai Intézet, Szeged, 235–247.

Mittelholcz Iván 2017. emToken: Unicode-képes tokenizáló magyar nyelvre. In: Vincze Veronika (szerk.): *XIII. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: Szegedi Tudományegyetem, Informatikai Intézet. 61–69.

Novák Attila – Rebrus Péter – Ludányi Zsófia 2017. Az emMorph morfológiai elemző annotációs formalizmusa. In: Vincze Veronika (szerk.): *XIII. Magyar Számítógépes Nyelvészeti Konferencia*. Szegedi Tudományegyetem, Informatikai Intézet, Szeged, 70–78.

Váradi Tamás – Simon Eszter – Sass Bálint – Gerőcs Mátyás – Mittelholtz Iván – Novák Attila – Indig Balázs – Prószéky Gábor – Vincze Veronika 2017. Az e-magyar digitális nyelvfeldolgozó rendszer. In: Vincze Veronika (szerk.): *XIII. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: Szegedi Tudományegyetem, Informatikai Intézet. 49–60.