UNED

LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

DH@Madrid Summer School 2016

Quantitative research on versification: The Corpus of Czech Verse

Tecnologías digitales aplicadas estudio de la poesía

Petr Plecháč
plechac@ucl.cas.cz

erc European Research Council

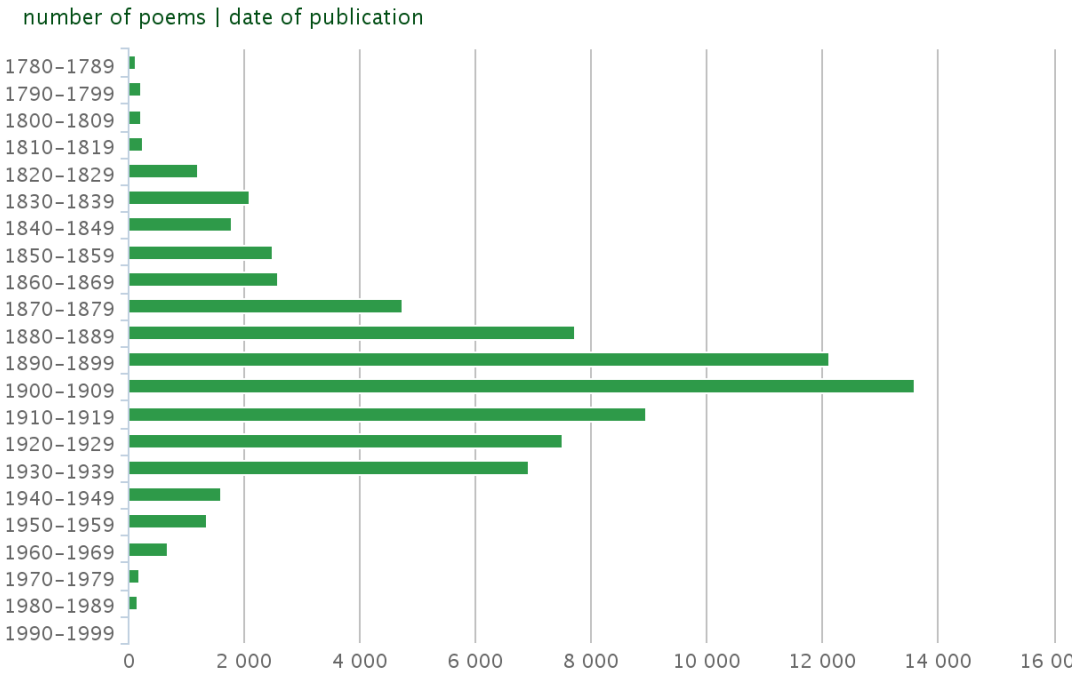
Madrid, 27 junio al 1 de julio de 2016

DiXiT Digital Scholarly Editions Initial Training Network

Corpus of Czech Verse

- Czech poetry of 19th and the beginning of 20th century

- Books: 1689
- Poems: 76 699
- Verse lines: 2 664 989
- Words: 14 592 037

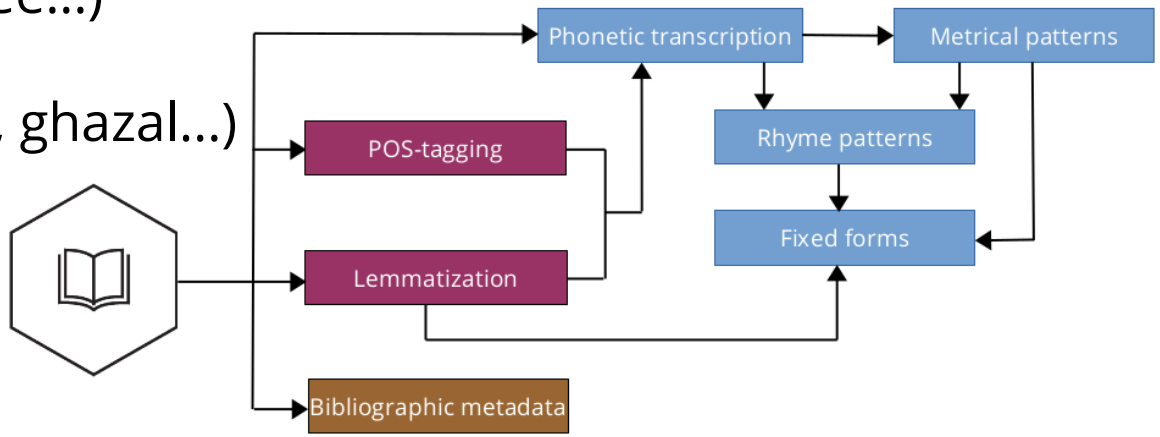


European Research Council

Corpus of Czech Verse

Various levels of annotation

- Lemmatization
- Morphological tagging (POS & others)
- Phonetic transcription
- Metres (iamb, trochee…)
- Rhymes
- Fixed forms (sonnet, ghazal…)



European Research Council

LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

How to work with the data?

erc European Research Council

On-line tools

Babel
Console application – direct SQL queries to the corpus.

Database of Czech Metres
Searching and statistics on metrical and stanzaic level of annotation

Euphonometer
Measuring (statistically defined) euphony of text

Frequency lists
of lemmata / words retrieved from particular books

Gunstick
Searching and statistics on rhymes

Hex
Searching and statistics on keywords in particular poems

European Research Council

LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

Gunstick

European Research Council

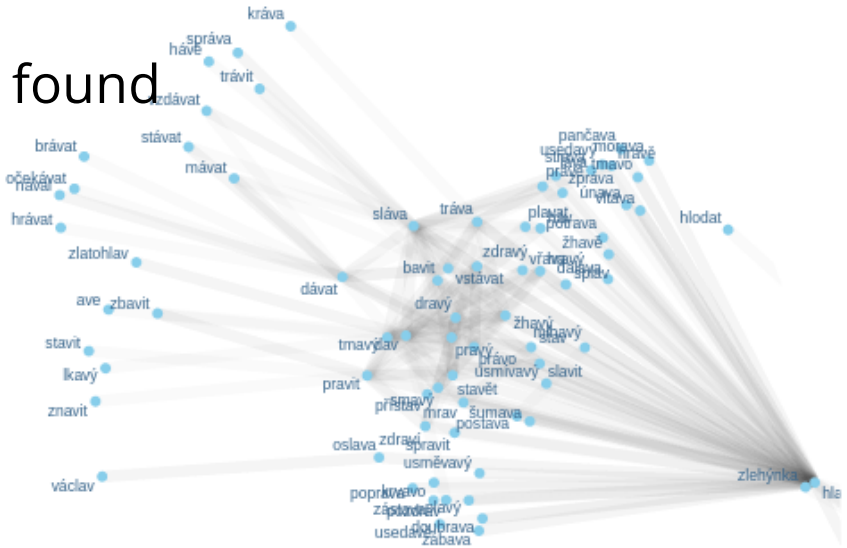
Gunstick

http://versologie.cz/gunstick/index_en.php

Search for RHYMES on word specified by the user
(over one million rhyme pairs)

Results:

- Frequency of rhymes on particular words
- Their distribution in time
- List of lines in which rhymes were found
  + links to full text of poems

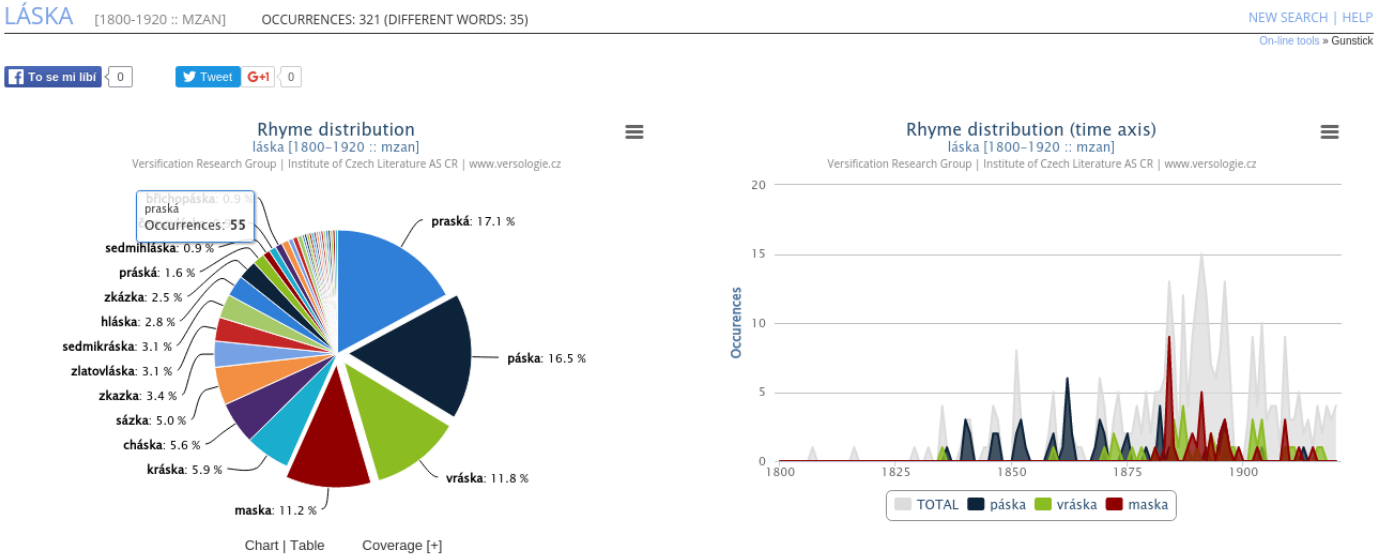


European Research Council

Gunstick

http://versologie.cz/gunstick/index_en.php

QUERY: láska (love)



Rhyme	Line 1	Line 2	Author	Poem	Collection	Year	End of a line
páska	buď velebena věčná Boha Láska,	nás pevná k nebi víry pojí páska	Bouška, Sigismund	Ó kniho Bohem psaná...	Pietas	1897	z
páska	čistá, pravá láska.	ta nebeská páska,	Chládek, František	Svátost a láska.	Básně samouka Františka Chládka	1884	z

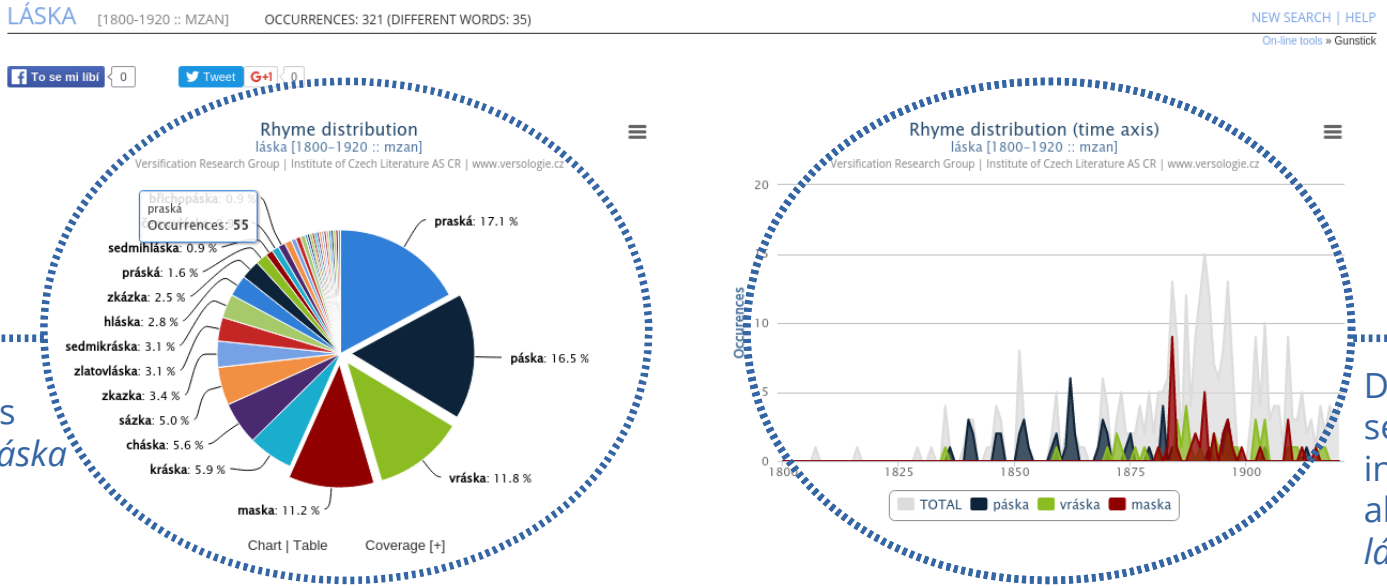
European Research Council

Gunstick

http://versologie.cz/gunstick/index_en.php

LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

QUERY: láska (love)



Frequency of different words rhyming with *láska*

Distribution of selected rhymes in time (grey = all rhymes with *láska*)

List of lines + links to full texts

Rhyme	Line 1	Line 2	Author	Poem	Collection	Year	End of line
páska	buď velebena věčná Boha Láska,	nás pevná k nebi víry pojí páska	Bouška, Sigismund	Ó kniho Bohem psaná...	Pietas	1897	z
páska	čistá, pravá láska.	ta nebeská páska,	Chládek, František	Svátost a láska.	Básně samouka Františka Chládka	1884	z

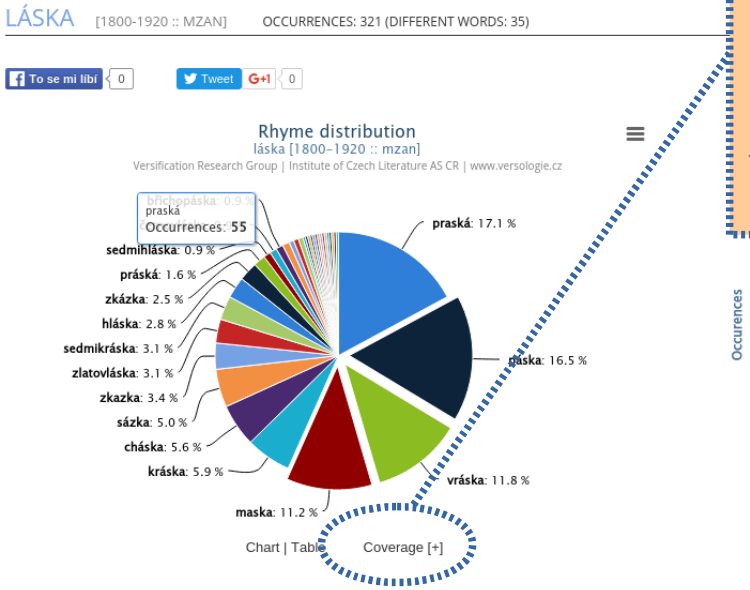
European Research Council

Gunstick

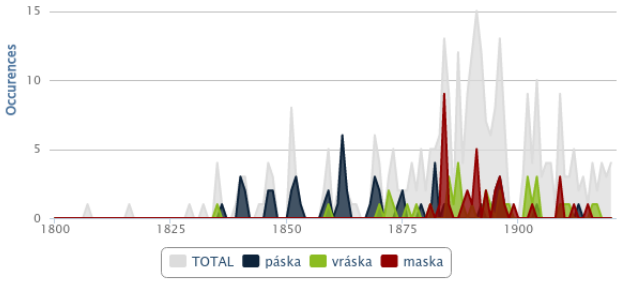
http://versologie.cz/gunstick/index_en.php

LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

QUERY: láska (love)



Corpus is not balanced in each year! Increasing grey area in second chart does not necessarily mean rhymes with *láska* are becoming more and more frequented during the century. One needs to check COVERAGE.



Rhyme	Line 1	Line 2	Author	Poem	Collection	Year	End of a line
páska	buď velebena věčná Boha Láska,	nás pevná k nebi víry pojí páska	Bouška, Sigismund	Ó kniho Bohem psaná...	Pietas	1897	z
páska	čistá, pravá láska.	ta nebeská páska,	Chládek, František	Svátost a láska.	Básně samouka Františka Chládka	1884	z

erc European Research Council

Gunstick

http://versologie.cz/gunstick/index_en.php

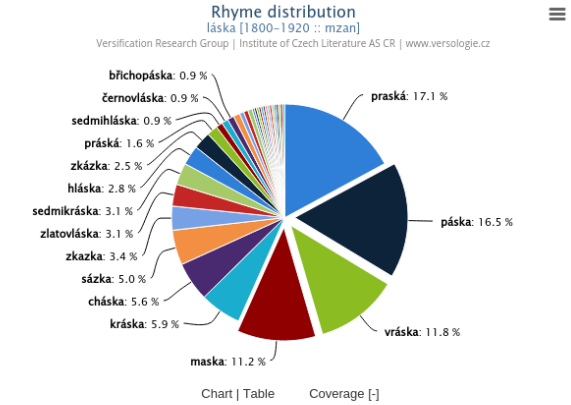
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

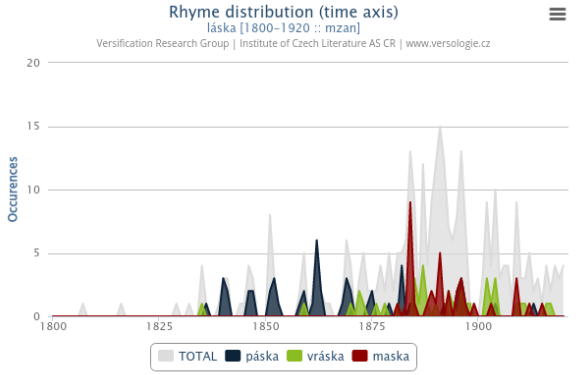
QUERY: láska (love)

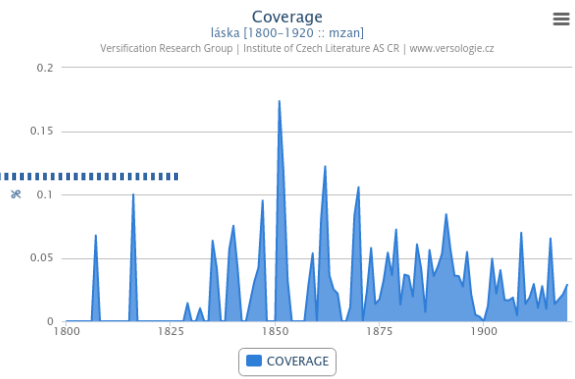
Chart shows
WHAT PERCENTAGE
of all rhymes found in each year
contain the word *láska.*

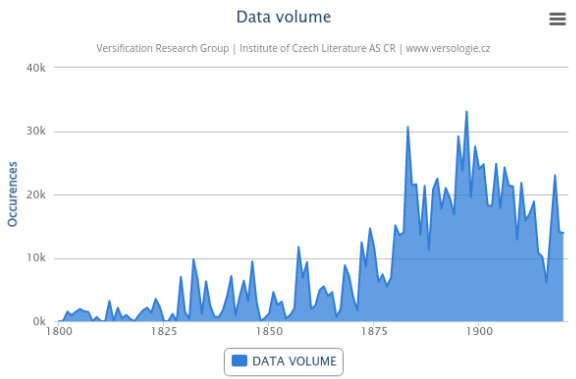
=>

If there's some trend it is rather increase until the middle of the century followed by decrease









European Research Council

LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

Hex

erc European Research Council

Hex

http://versologie.cz/hex/

LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

Search for KEYWORDS

Keyword: Statistically defined „topic“ of the text (poem)

CONCEPTION → Keyword is a word (lemma) which occurs in a text (poem) more often than one would expect to occur by chance alone

DEFINITION → Keyword of a poem is a word (lemma) which occurs in it with statistically higher frequency than in the rest of the corpus (statistical test: $\chi 2$ with Yates correction)

erc European Research Council

Hex

http://versologie.cz/hex/

LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

Search for KEYWORDS

(1) Search for poems containing specified keyword

(2) Search for keywords of specified poems

erc European Research Council

Hex

http://versologie.cz/hex/

LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

Query (search for occurence of keywords)

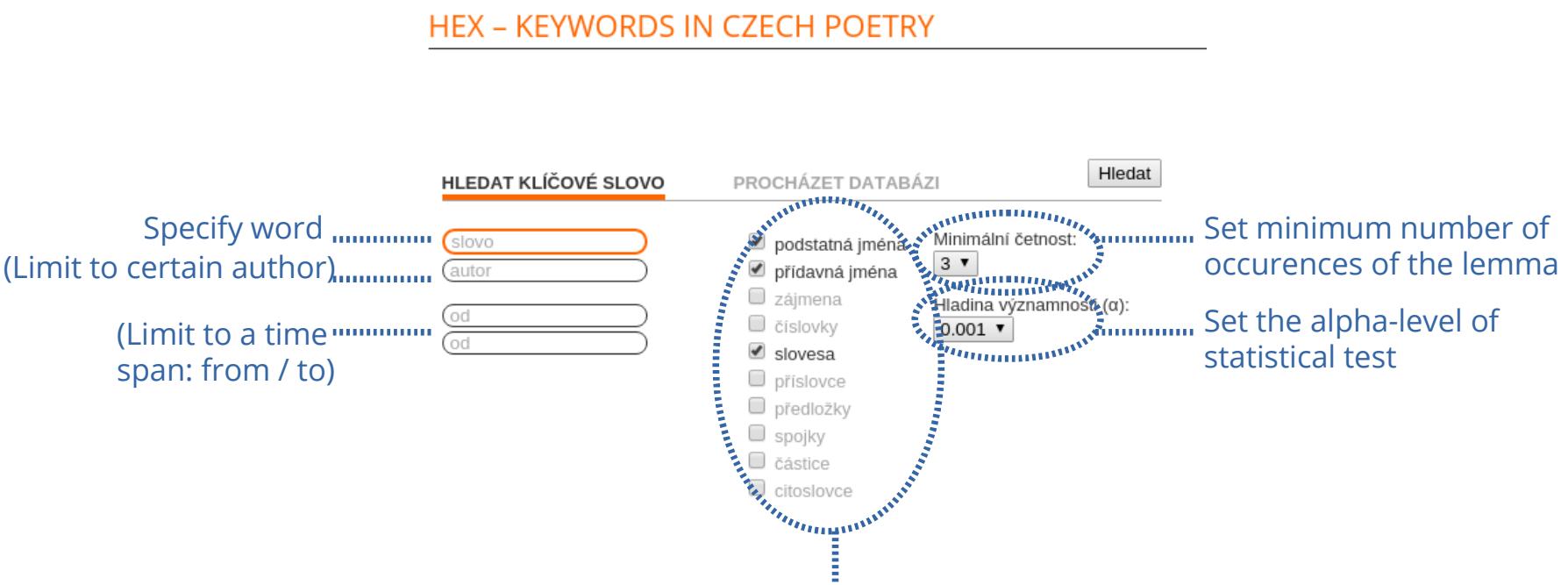
erc European Research Council

Hex

http://versologie.cz/hex/

LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

Query (search for occurence of keywords)



Specify word

(Limit to certain author)

(Limit to a time span: from / to)

Set minimum number of occurences of the lemma

Set the alpha-level of statistical test

Limit query to only some parts-of-speech
Due to rich homonymity in Czech, e.g.: *stát* = (1) state, country (noun), (2) stand (verb)

erc European Research Council

Hex

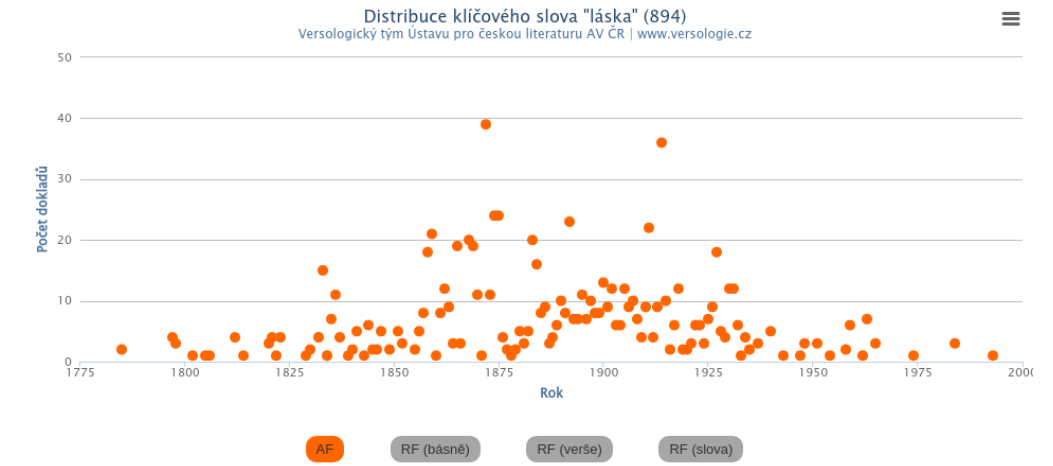
http://versologie.cz/hex/

LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

Results – query: láska (love)

DISTRIBUCE KLÍČOVÉHO SLOVA "LÁSKA" (894) AUT: | SB: | – (*A* = 0.001; *N* ≥ 3; POS: NAV)

NOVÉ HLEDÁNÍ | O APLIKACI



DOKLADY KLÍČOVÉHO SLOVA "LÁSKA" (894) AUT: | SB: | – (*A* = 0.001; *N* ≥ 3; POS: NAV)

Autor	Báseň	Sbírka	Rok	AF	RF	φ
Ambrož, Vilém	Tré hlasů. (III.) »	Pestré kvítí »	1883	5	61728	0.0025
Ambrož, Vilém	Ecce lignum crucis! »	Pestré kvítí »	1883	4	33898	0.0015

erc European Research Council

Hex

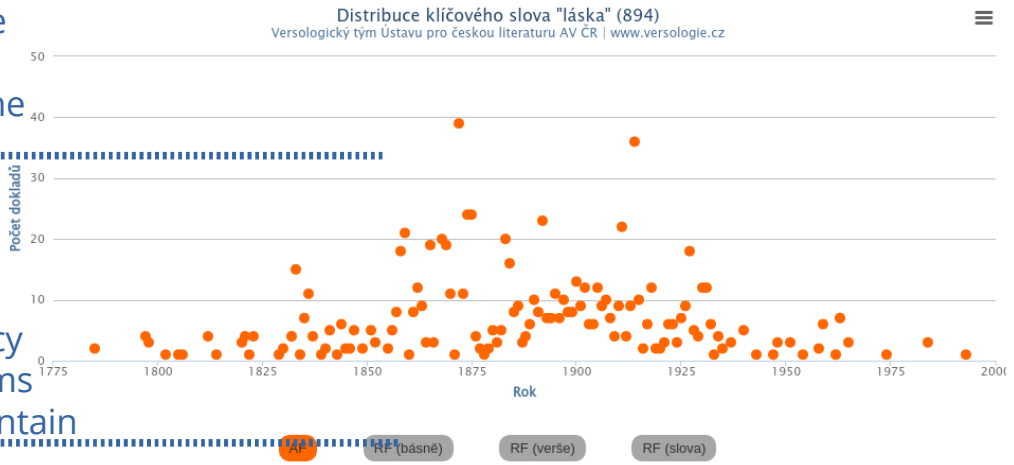
http://versologie.cz/hex/

LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

Results – query: láska (love)

DISTRIBUCE KLÍČOVÉHO SLOVA "LÁSKA" (894) AUT: | SB: | – (*A* = 0.001; *N* ≥ 3; POS: NAV) NOVÉ HLEDÁNÍ | O APLIKACI

Each point represents the number of poems found in each year containing the keyword *láska*



Switch to relative frequency (what PORTION of all poems published in given year contain the word *láska*)

DOKLADY KLÍČOVÉHO SLOVA "LÁSKA" (894) AUT: | SB: | – (*A* = 0.001; *N* ≥ 3; POS: NAV)

Autor	Báseň	Sbírka	Rok	AF	RF	φ
Ambrož, Vilém	Tré hlasů. (III.) »	Pestré kvítí »	1883	5	61728	0.0025
Ambrož, Vilém	Ecce lignum crucis! »	Pestré kvítí »	1883	4	33898	0.0015

List of poems + links to full-texts + links to the lists of keywords found in them under same conditions

erc European Research Council

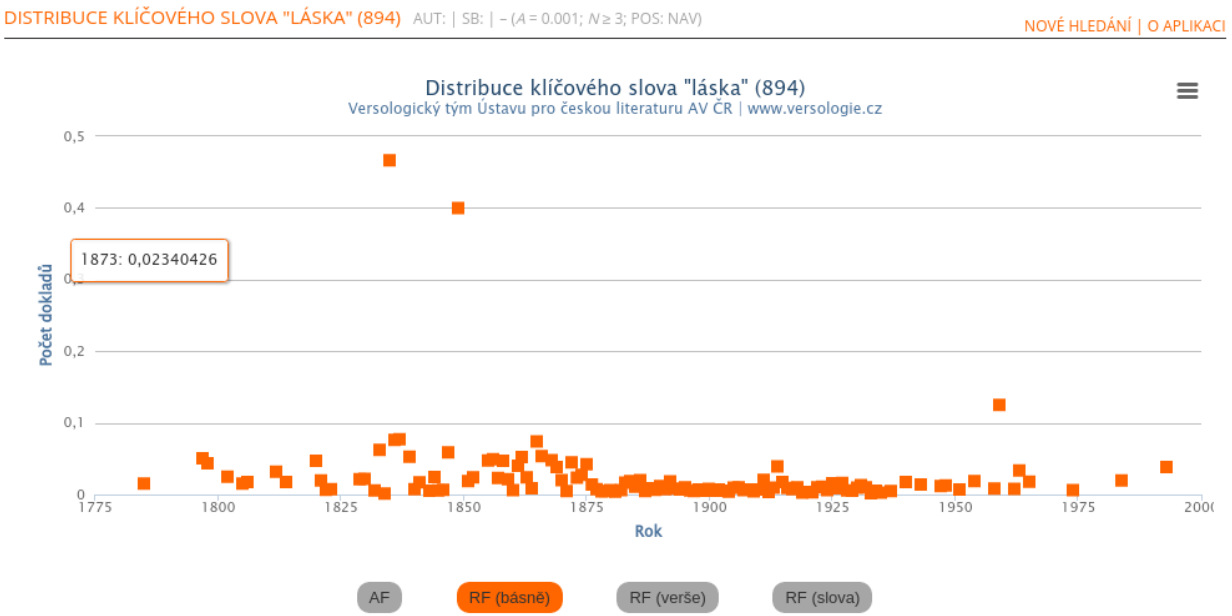
Hex

http://versologie.cz/hex/

LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

Results – query: láska (love) – RELATIVE FREQUENCIES

No clear trend at all, just some outliers…



erc European Research Council

LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

What such findings may be used for?

Among others: the attribution of texts with unkown or disputed authorship...

... let's go on-line:

http://versologie.cz/talks/uned

erc European Research Council