18TH EURALEX INTERNATIONAL CONGRESS

Lexicography in global contexts
17-21 July 2018
Ljubljana, Slovenia

The Dictionary of the Serbian Academy: from the Text to the Lexical Database

https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/2975-1

Ranka Stanković[1] Rada Stijović[2] Duško Vitas[1] Cvetana Krstev[1] Olga Sabo[2]

[1] Human Language Technology group at the University of Belgrade
[2] Institute for Serbian Language, Serbian Academy of Sciences and Arts Serbia

EURALEX

Motivation

- Dictionary of Serbian Academy or DSA
  - ✓19 volumes printed, 20th in print, ~15 more to come
  - ✓covers written resources of the standard Serbo-Croatian (from 19th century to the present day) of all Shtokavian dialects
  - ✓complex microstructure
- Task:
  - ✓ automatic transformation of the digitized text of DSA into various standard structured formats and into a lexical database
  - ✓ use the lexical base of the Dictionary for research purposes and for the production of different derived lexicographic products

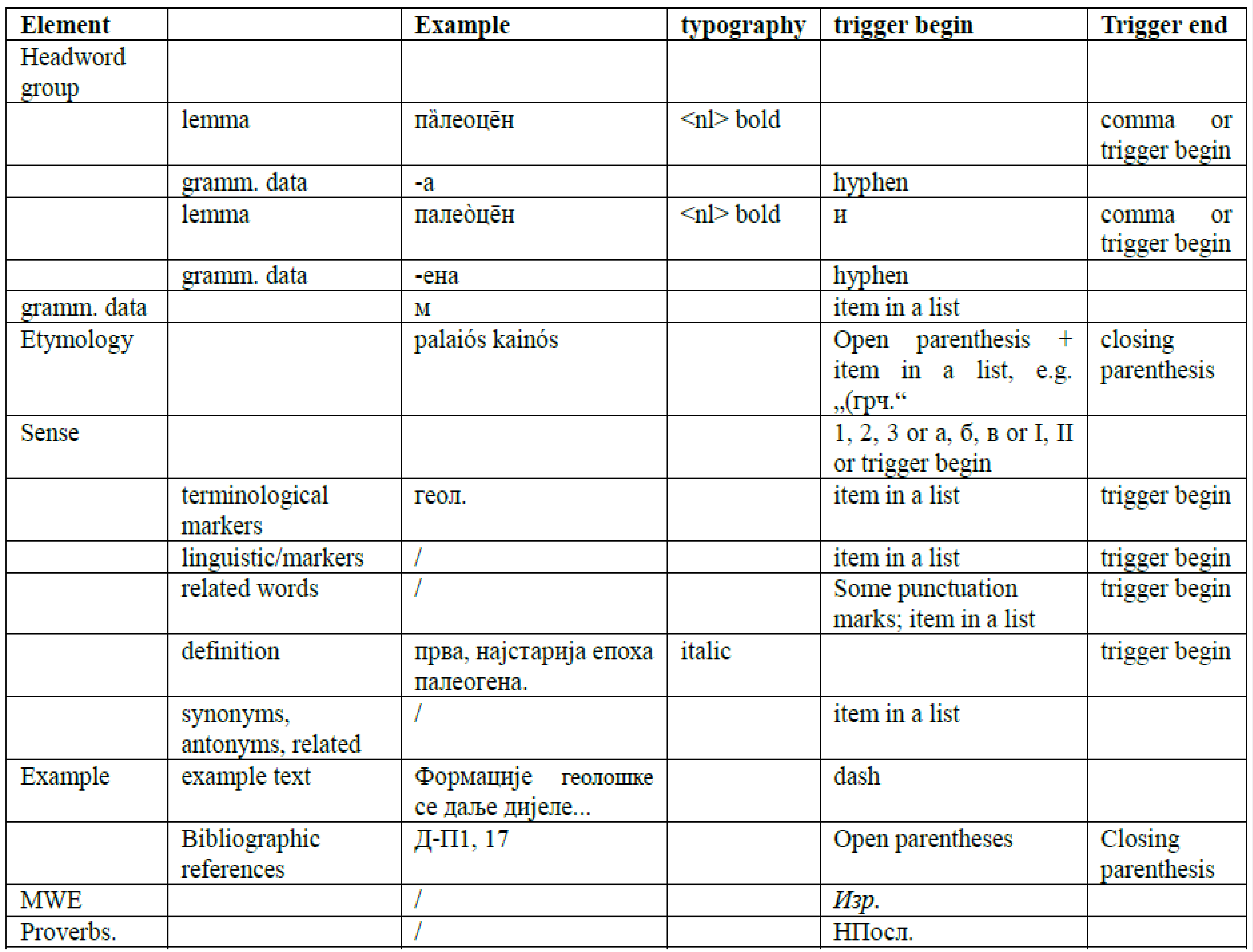
Model of Lexical Database

- **Analysis of formatting conventions used for typesetting dictionary entries and identification of triggers:**
  1. **headword group,**
  2. **grammatical data,**
  3. **etymology,**
  4. **sense,**
  5. **multiword expressions,**
  6. **proverbs.**

Model of Lexical Database

- **Dictionary markers**
  - ✓ **semantic, phonetic, morphological, syntactic, normative, functional, stylistic, domain of use,..**
  - ✓ **371 abbreviations (markers) mapped as data category values**
  - ✓ **30 data categories, further grouped in data-category set**

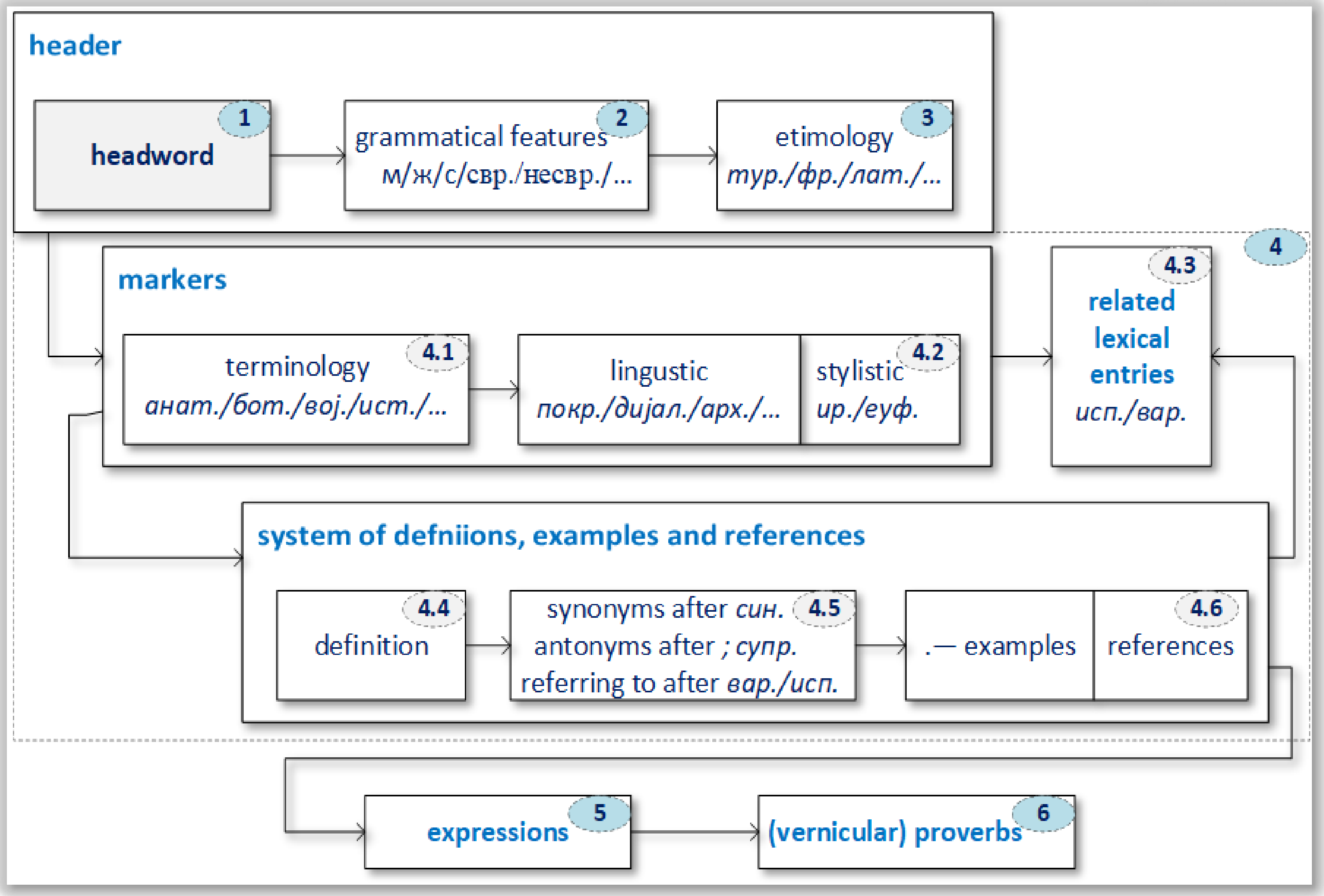
Typographic conventions and triggers as applied to nouns



Element		Example	typography	trigger begin	Trigger end
Headword group					
	lemma	па̏леоце̄н	<nl> bold		comma or trigger begin
	gramm. data	-а		hyphen	
	lemma	палѐоце̄н	<nl> bold	и	comma or trigger begin
	gramm. data	-ена		hyphen	
gramm. data		м		item in a list	
Etymology		palaiós kainós		Open parenthesis + item in a list, e.g. „(грч.“	closing parenthesis
Sense				1, 2, 3 or а, б, в or I, II or trigger begin	
	terminological markers	геол.		item in a list	trigger begin
	linguistic/markers	/		item in a list	trigger begin
	related words	/		Some punctuation marks; item in a list	trigger begin
	definition	*прва, најстарија епоха палеогена.*	italic		trigger begin
	synonyms, antonyms, related	/		item in a list	
Example	example text	Формације геолошке се даље дијеле...		dash	
	Bibliographic references	Д-П1, 17		Open parentheses	Closing parenthesis
MWE		/		*Изр.*	
Proverbs.		/		НПосл.	

Microstructure of dictionary articles

EURALEX



header

headword 1

grammatical features 2
м/ж/с/свр./несвр./...

etimology 3
тур./фр./лат./...

4

markers

terminology 4.1
анат./бот./воj./ист./...

lingustic
покр./диjал./арх./...

stylistic 4.2
ир./еуф.

4.3
related
lexical
entries
исп./вар.

system of defniions, examples and references

4.4
definition

synonyms after *син.* 4.5
antonyms after *; супр.*
referring to after *вар./исп.*

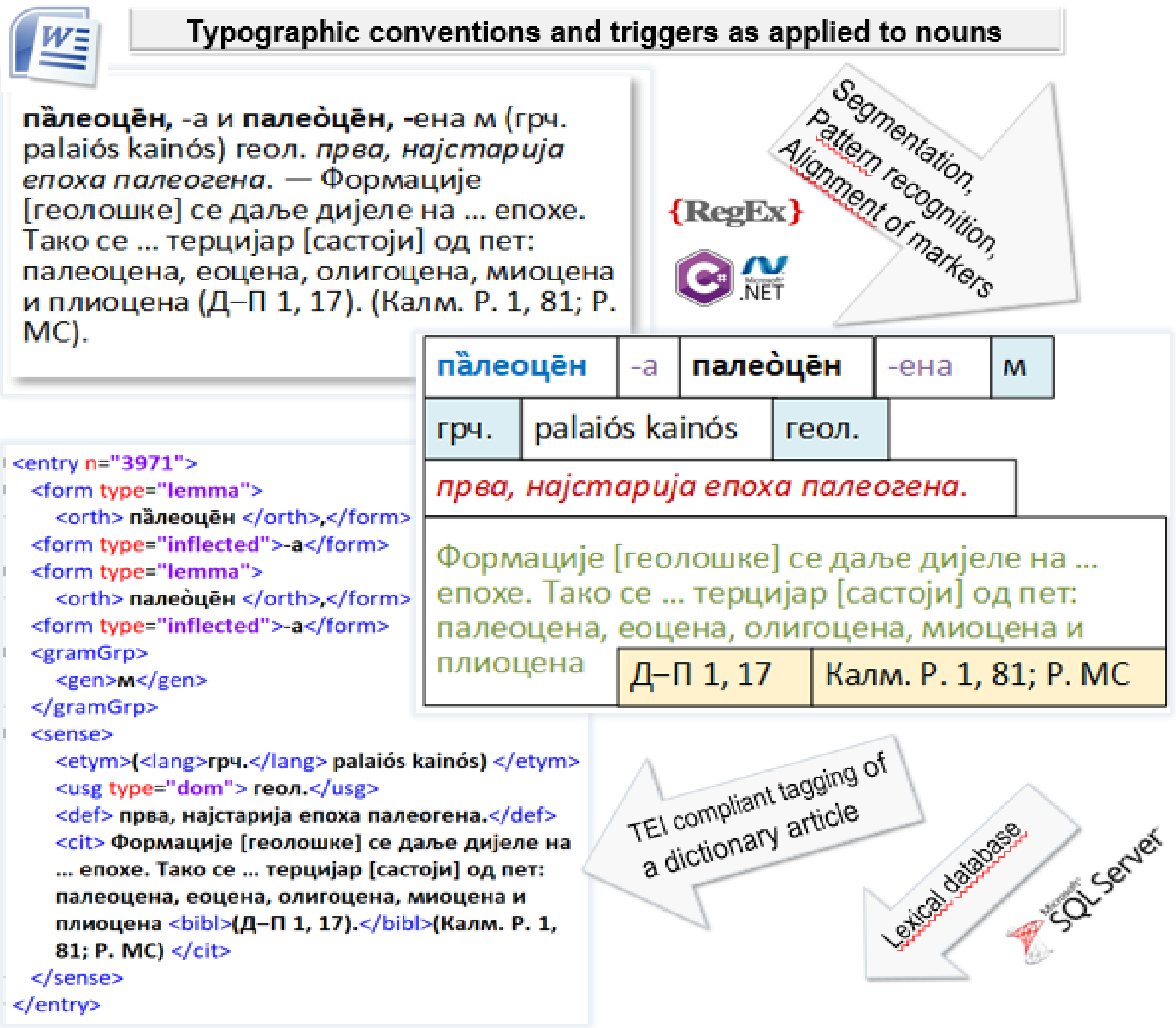
.— examples

4.6
references

expressions 5

(vernicular) proverbs 6

Transformation: dictionary article text form -> lexical database



Typographic conventions and triggers as applied to nouns

па̏леоцēн, -а и **палеòцēн,** -ена м (грч. palaiós kainós) геол. *прва, најстарија епоха палеогена.* — Формације [геолошке] се даље дијеле на ... епохе. Тако се ... терцијар [састоји] од пет: палеоцена, еоцена, олигоцена, миоцена и плиоцена (Д–П 1, 17). (Калм. Р. 1, 81; Р. МС).

Segmentation, Pattern recognition, Alignment of markers

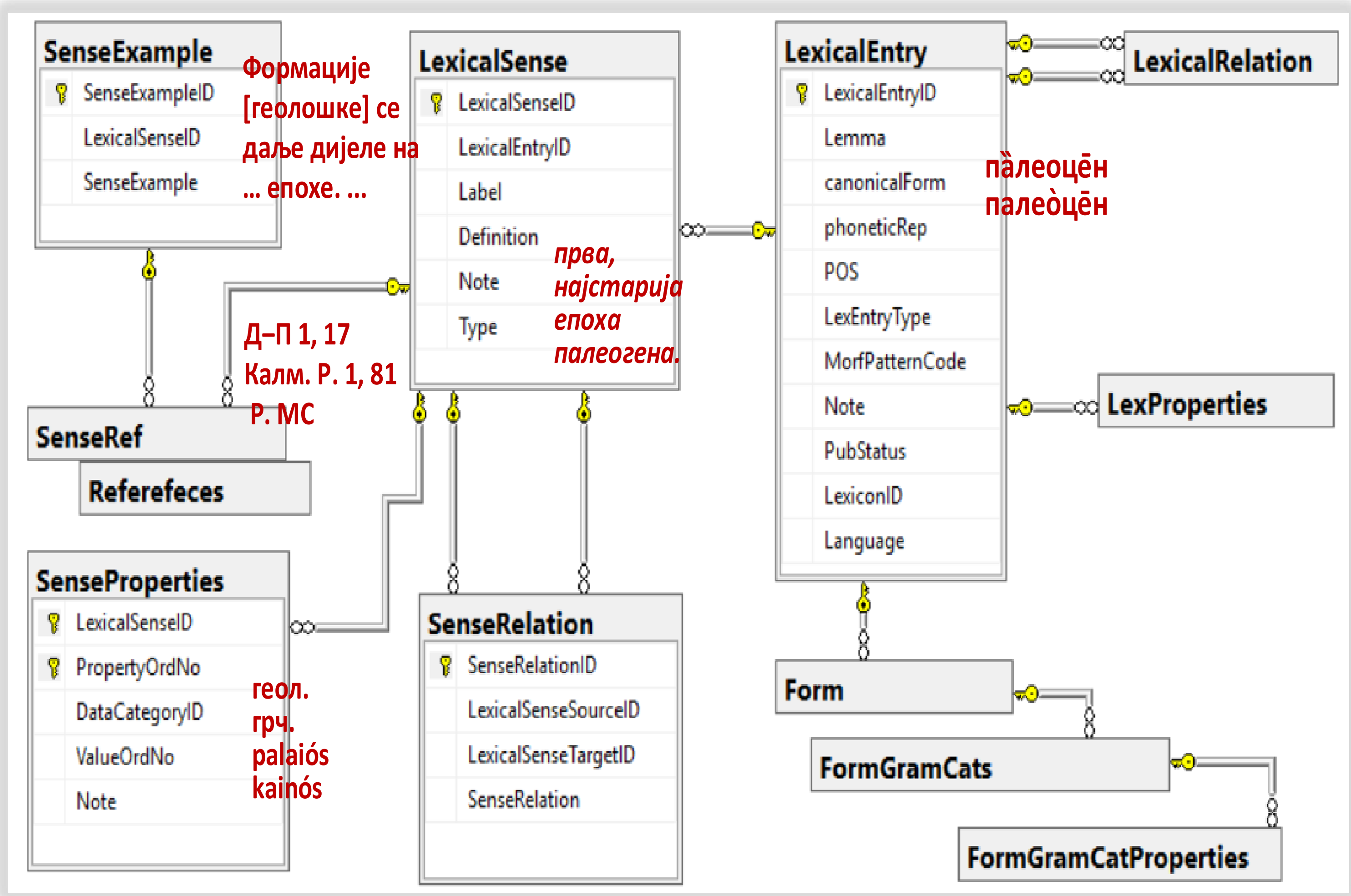
па̏леоцēн	-а	**палеòцēн**	-ена	м
грч.	palaiós kainós	геол.		
прва, најстарија епоха палеогена.				
Формације [геолошке] се даље дијеле на ... епохе. Тако се ... терцијар [састоји] од пет: палеоцена, еоцена, олигоцена, миоцена и плиоцена	Д–П 1, 17	Калм. Р. 1, 81; Р. МС		

TEI compliant tagging of a dictionary article

```
<entry n="3971">
  <form type="lemma">
     <orth> па̏леоцēн </orth>,</form>
  <form type="inflected">-а</form>
  <form type="lemma">
     <orth> палеòцēн </orth>,</form>
  <form type="inflected">-а</form>
  <gramGrp>
     <gen>м</gen>
  </gramGrp>
  <sense>
     <etym>(<lang>грч.</lang> palaiós kainós) </etym>
     <usg type="dom"> геол.</usg>
     <def> прва, најстарија епоха палеогена.</def>
     <cit> Формације [геолошке] се даље дијеле на
     ... епохе. Тако се ... терцијар [састоји] од пет:
     палеоцена, еоцена, олигоцена, миоцена и
     плиоцена <bibl>(Д–П 1, 17).</bibl>(Калм. Р. 1,
     81; Р. МС) </cit>
  </sense>
</entry>
```

Lexical database

Transformation:
dictionary article text form -> lexical database



SenseExample
- SenseExampleID
- LexicalSenseID
- SenseExample

Формације [геолошке] се даље дијеле на ... епохе. ...

LexicalSense
- LexicalSenseID
- LexicalEntryID
- Label
- Definition
- Note
- Type

прва, најстарија епоха палеогена.

LexicalEntry
- LexicalEntryID
- Lemma
- canonicalForm
- phoneticRep
- POS
- LexEntryType
- MorfPatternCode
- Note
- PubStatus
- LexiconID
- Language

па̏леоцѐн
палео̀цѐн

LexicalRelation

LexProperties

SenseRef

Referefeces

Д–П 1, 17
Калм. Р. 1, 81
Р. МС

SenseProperties
- LexicalSenseID
- PropertyOrdNo
- DataCategoryID
- ValueOrdNo
- Note

геол.
грч.
palaiós
kainós

SenseRelation
- SenseRelationID
- LexicalSenseSourceID
- LexicalSenseTargetID
- SenseRelation

Form

FormGramCats

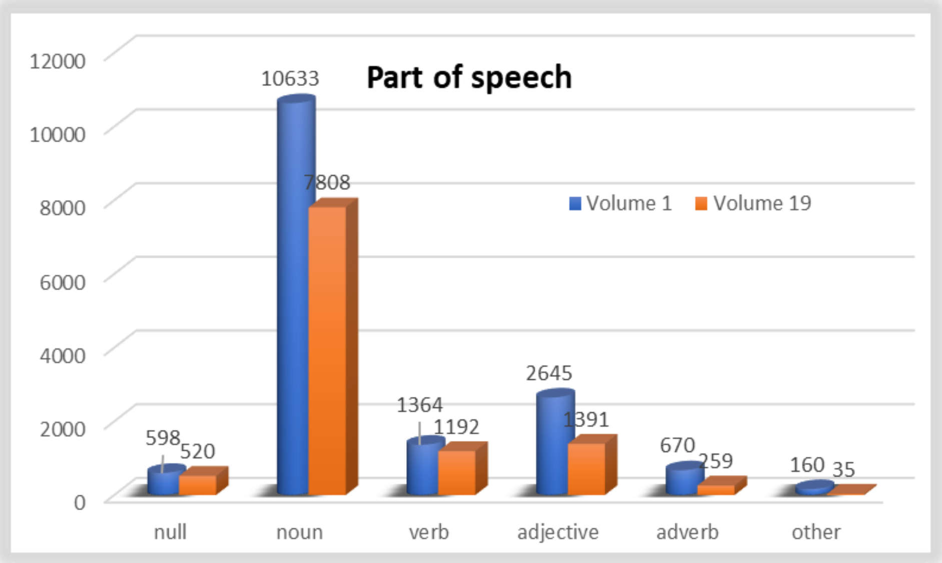
FormGramCatProperties

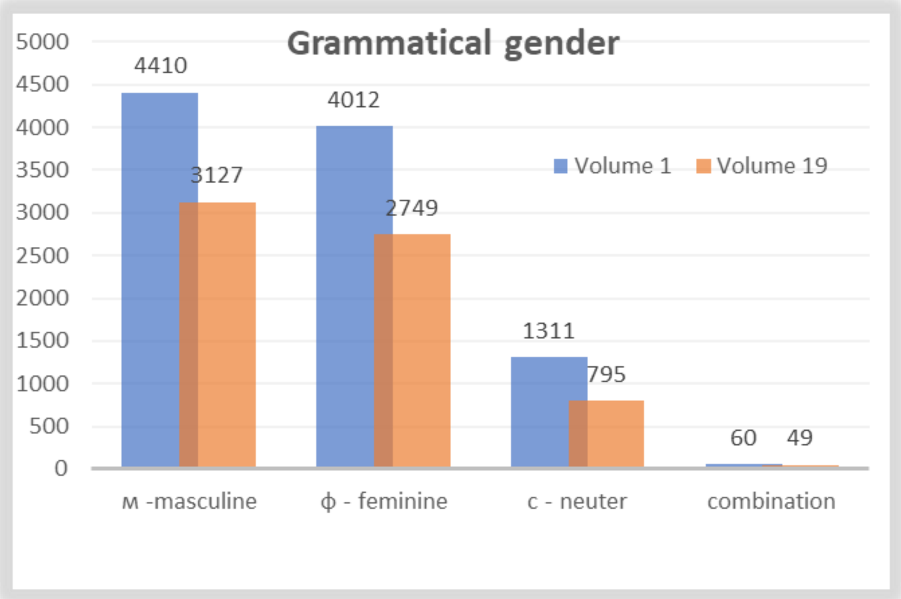
Results and Evaluation

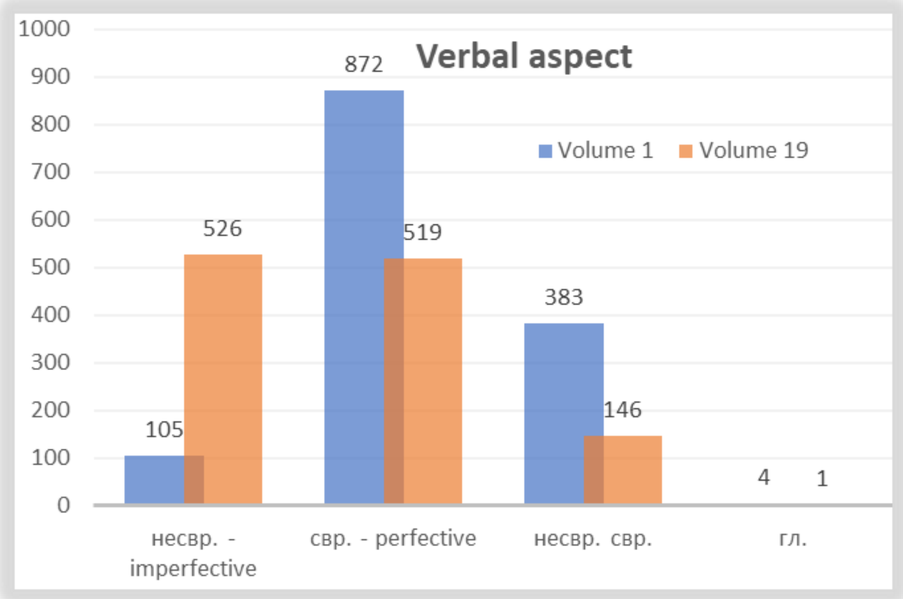
1st (1959) and 19th (2014) volumes

1) Automatic procedure recognized, structured, annotated and stored in the lexical database **15,988** dictionary articles from 1st
and **11,153** from 19th volume
2) Digitization of all previously published volumes is in progress
3) In future, the linear processing of entries may be abandoned which would possibly accelerate the production process.
4) The supplements to the already published volumes could also be produced.
5) Once the DSA is fully populated, users with different levels of accessibility, will be able to search through its lexical data base.

1st (1959) and 19th (2014) volumes


Part of speech
Volume 1
Volume 19
null
noun
verb
adjective
adverb
other


Grammatical gender
Volume 1
Volume 19
м -masculine
ф - feminine
с - neuter
combination


Verbal aspect
Volume 1
Volume 19
несвр. - imperfective
свр. - perfective
несвр. свр.
гл.