www.slovenščina.eu

Using GDEX in (semi)-automatic creation of database entries

Iztok Kosem

Trojina, Institute for Applied Slovene Studies

iztok.kosem@trojina.si

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA IZOBRAŽEVANJE, ZNANOST, KULTURO IN ŠPORT

Investing in your future
OPERATION PART FINANCED BY THE EUROPEAN UNION
European Social Fund

New Lexical Database for Slovene

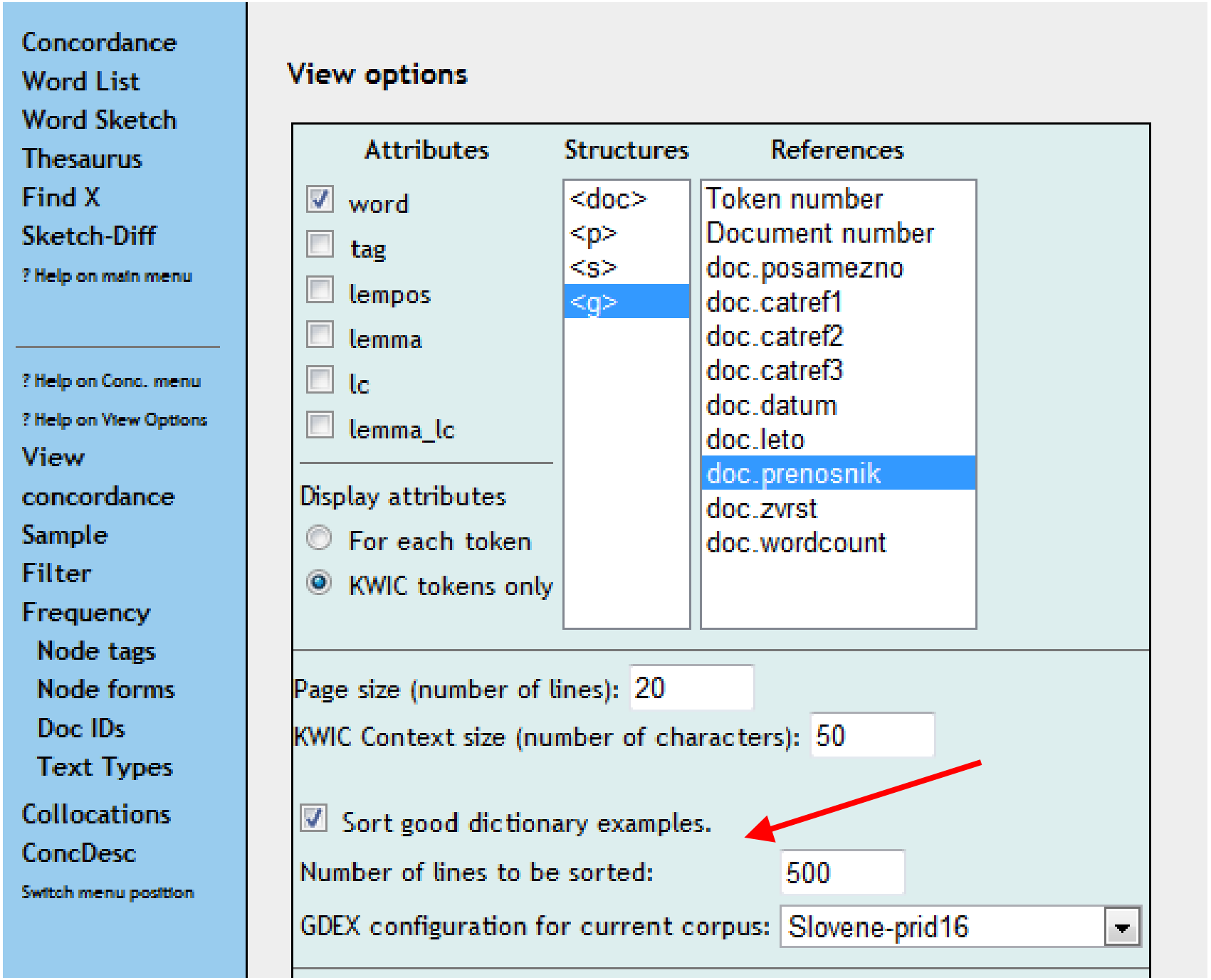
- Activity of the “Communication in Slovene project
- Time frame: 2008-2012
- Funding: European Social Fund and Ministry of Education and Sport of the Republic of Slovenia
- Aims:
  - making a framework for comprehensive corpus-based record of Slovene
  - Devising standard procedures for corpus analysis

Phase 1 (pre-GDEX)

- FidaPLUS corpus of Slovene (620 million words)
- Procedure:
  - Random selection of 300 examples
  - Identifying meanings (WSD)
  - Word sketch: identifying collocates, constructions/grammatical relations, patterns, compounds, phrases (for each meaning)
  - finding examples for collocates/grammatical relations (using TickBox Lexicography or Concordance)

Phase 2 (current)

- Same as Phase 1, BUT
- Examples: TBL + **GDEX for Slovene**


Concordance
Word List
Word Sketch
Thesaurus
Find X
Sketch-Diff
? Help on main menu
? Help on Conc. menu
? Help on View Options
View concordance
Sample
Filter
Frequency
Node tags
Node forms
Doc IDs
Text Types
Collocations
ConcDesc
Switch menu position
View options
Attributes
Structures
References
word
tag
lempos
lemma
lc
lemma_lc
<doc>
<p>
<s>
<g>
Token number
Document number
doc.posamezno
doc.catref1
doc.catref2
doc.catref3
doc.datum
doc.leto
doc.prenosnik
doc.zvrst
doc.wordcount
Display attributes
For each token
KWIC tokens only
Page size (number of lines): 20
KWIC Context size (number of characters): 50
Sort good dictionary examples.
Number of lines to be sorted: 500
GDEX configuration for current corpus: Slovene-prid16

Automatic entry production

- What can be automatized:

Meaning?	NOT YET
Grammatical relations, constructions, collocates?	YES – Word Sketch
Examples?	YES – GDEX

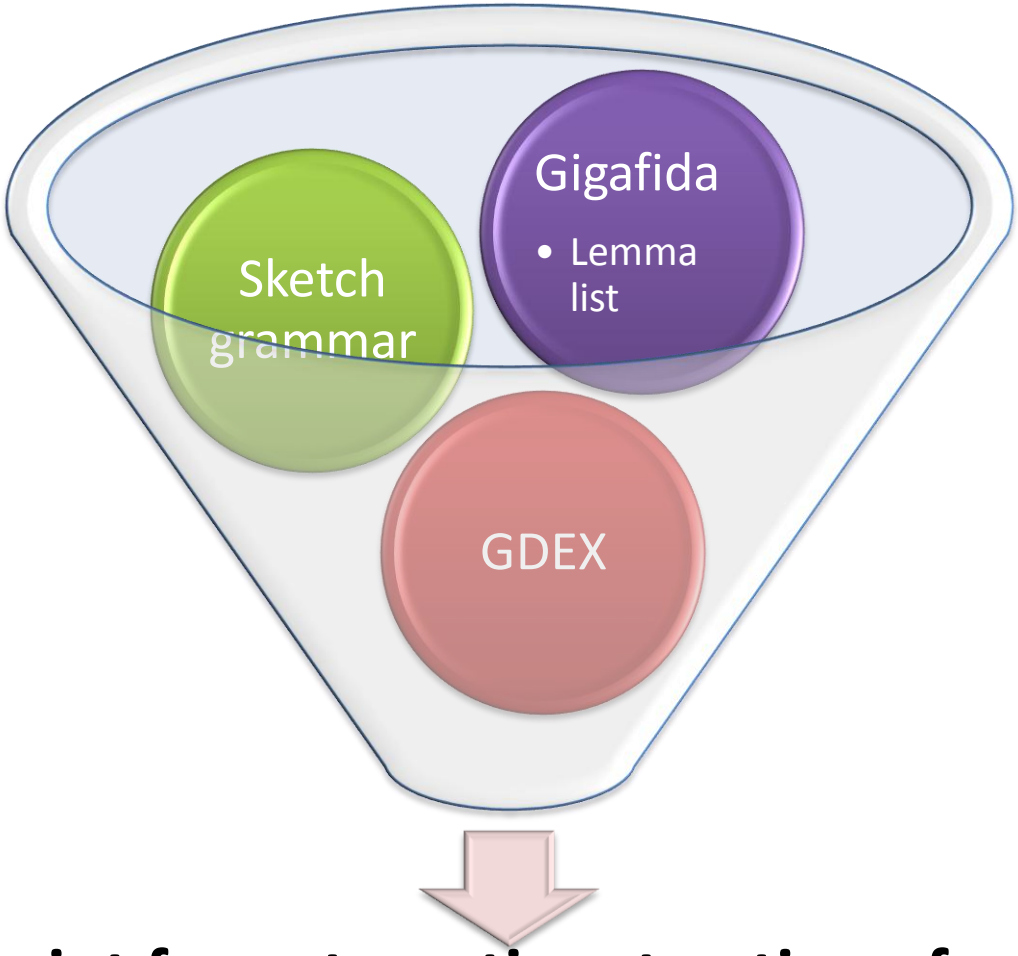
Phase 3 (in testing)

- Gigafida - New 1,1-billion-word corpus of Slovene
- Making a selection of 500 lemmas:
  - 200 nouns, 150 verbs, 125 adjectives, 25 adverbs
- WSD - selecting monosemous or less polysemous words:
  - Lemmas from Gigafida (frequency: 1000 - 50.000)
  - sloWNet: words belonging to 1 or 2 synsets
  - Dictionary of Standard Slovene: entries with 1 or 2 senses
  - Not already in the lexical database

Phase 3 (in testing)

- Gigafida - New 1,1-billion-word corpus of Slovene
- Making a selection of 500 lemmas:
  - 200 nouns, 150 verbs, 125 adjectives, 25 adverbs
- Word Sketch: new sketch grammar
- GDEX: improved GDEX

Phase 3 (in testing)



API script for automatic extraction of entries

(Miloš Husák)

GDEX

- Ranking tool
- Heuristics classifiers: punctuation, sentence length, word length, etc.
- Procedure:
  - Score on each classifier
  - Weight scores
  - Weighted average of classifier scores
  - Rank sentences
- GDEX Tools for devising configurations

Select configuration: Slovene-prid17

Configuration name: Slovene-prid17

Save

- total (>)
  - bollean all (>)
    - whole_sentence (>)
    - Never (>)
    - min occurances_score (>)
    - max sentence length score (>)
    - min word length score (>)
      - Classifier50 (>)
      - Classifier51 (>)
    - blacklist_score (>)
    - keyword_repetition_score (>)
  - sentence_length score (>)
  - long words score (>)
  - almost never score (>)
  - capital_letters_score (>)
  - mixed symbols score (>)
  - proper nouns score (>)
  - pronoun score (>)
  - first_word_sentence_score (>)
  - sentence_punctuation (>)
  - low_freq_lemma_score (>)
  - first_phrase_score (>)

Returns 1 if all subclassifiers return non-zero results.

Classifier name: bollean all

Classifier class: _all

Classifier type: 1

Classifier weight: 100

Weight of the classifier for weighted average (currently not used by any other classifier).

Add child classifier:

Compositional Operators

- _all
- _any
- _if_then_else
- _max
- _min
- _sum
- _weighted_average
- _weka
- ---
- total
- bollean all

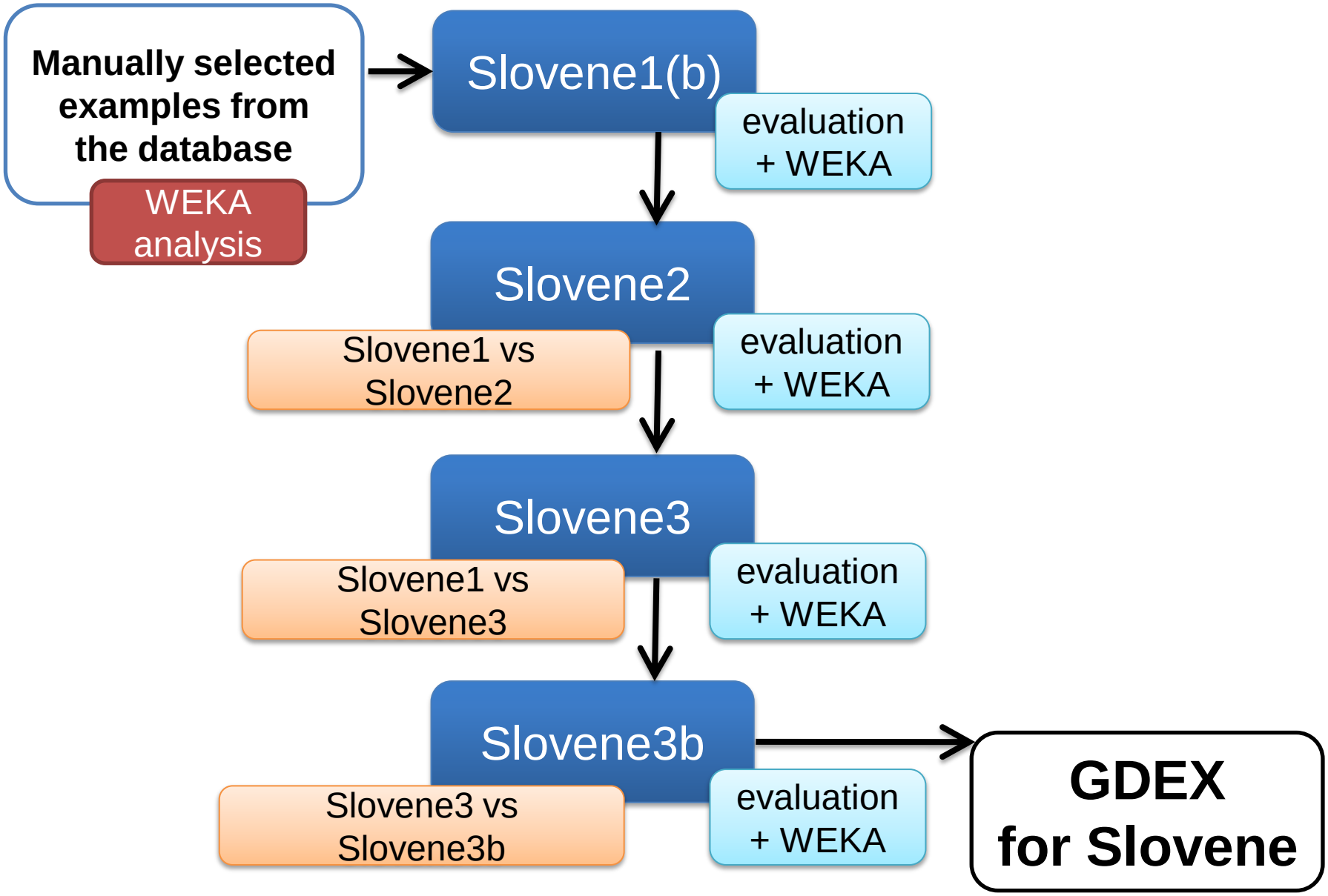
Sentence Operators

- op_equal
- op_fraction
- op_larger_than
- op_normalize
- op_optimal_interval
- op_slot_map
- op_smaller_than
- op_transformation
- ---
- Never
- min occurances_score

Sentence A

- s_c
- s_k
- s_k
- s_se
- s_to
- s_w
- ---
- who
- sent

GDEX for Slovene – manual selection



GDEX: Slovene3

prebivalstvo

- V večini držav v razvoju naj bi siva ekonomija vključevala med 30 - 70 odstotkov **aktivnega** prebivalstva.
- Leta 2001 naj bi bilo v EU brezposelnih 8 odstotkov **aktivnega** prebivalstva oziroma 15 milijonov oseb.
- Zakon določa, da je lahko le pet odstotkov **aktivnega** prebivalstva tujcev, torej približno 41.000 ljudi.
- Brez službe je okoli 388.000 ljudi ali več kakor 22 odstotkov **aktivnega** prebivalstva.
- Do leta 2010 naj bi imelo 85 odstotkov **aktivnega** prebivalstva doseženo srednješolsko izobrazbo.
- Brezposelnost na širšem območju Maribora je pred dobrima dvema letoma zajela že skoraj četrtino **aktivnega** prebivalstva.
- Rast zaposlenosti v ZDA je letos že presegla naravno povečanje **aktivnega** prebivalstva za skoraj pol odstotne točke.
- Februarja 2001 je tako v Sloveniji internet uporabljalo okoli 19 % **aktivnega** prebivalstva.
- V črnomaljski in semiški občini je zaposlenih 5.634 ljudi ali 80,5 odst **aktivnega** prebivalstva.
- Tedaj bo razmerje med **aktivnim** prebivalstvom in upokojenci 10 proti 7.

GDEX: Slovene2

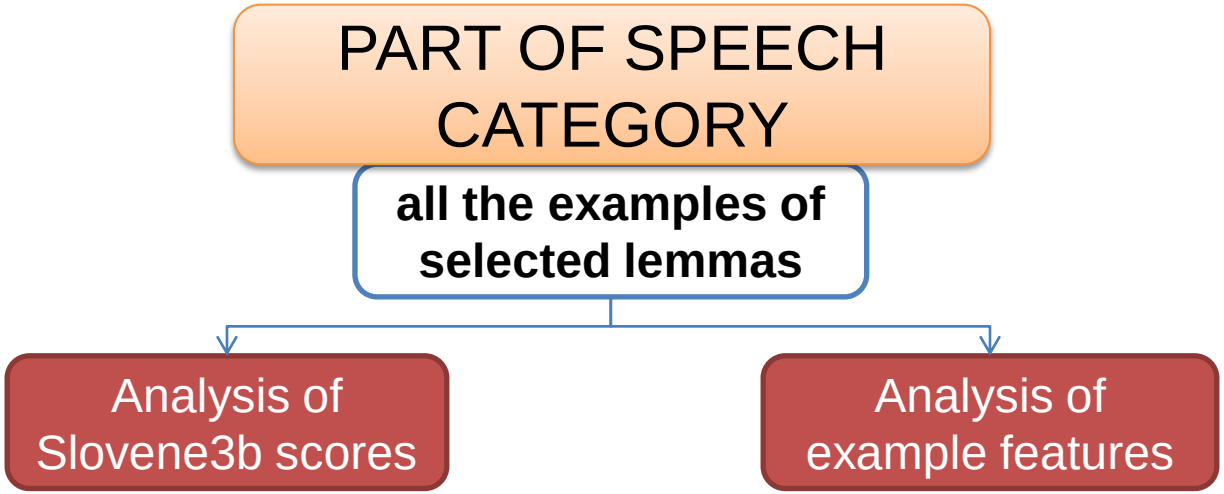
prebivalstvo

- Zaradi bolečin v križu so največje težave pri **aktivnem** prebivalstvu.
- Tedaj bo razmerje med **aktivnim** prebivalstvom in upokojenci 10 proti 7.
- Zakaj je **aktivno** prebivalstvo na udaru?
- To pa je okrog 60 odstotkov **aktivnega** prebivalstva.
- Brez službe je okoli 388.000 ljudi ali več kakor 22 odstotkov **aktivnega** prebivalstva.
- Delovno aktivni in brezposelni sestavljajo skupaj **aktivno** prebivalstvo.
- V večini držav v razvoju naj bi siva ekonomija vključevala med 30 - 70 odstotkov **aktivnega** prebivalstva.
- Delež aktivnih žensk v skupnem številu **aktivnega** prebivalstva je 47 odstotkov.
- Do leta 2010 naj bi imelo 85 odstotkov **aktivnega** prebivalstva doseženo srednješolsko izobrazbo.
- Najbolj ogroža **aktivno** prebivalstvo, predvsem ljudi, stare od 35 do 45 let.

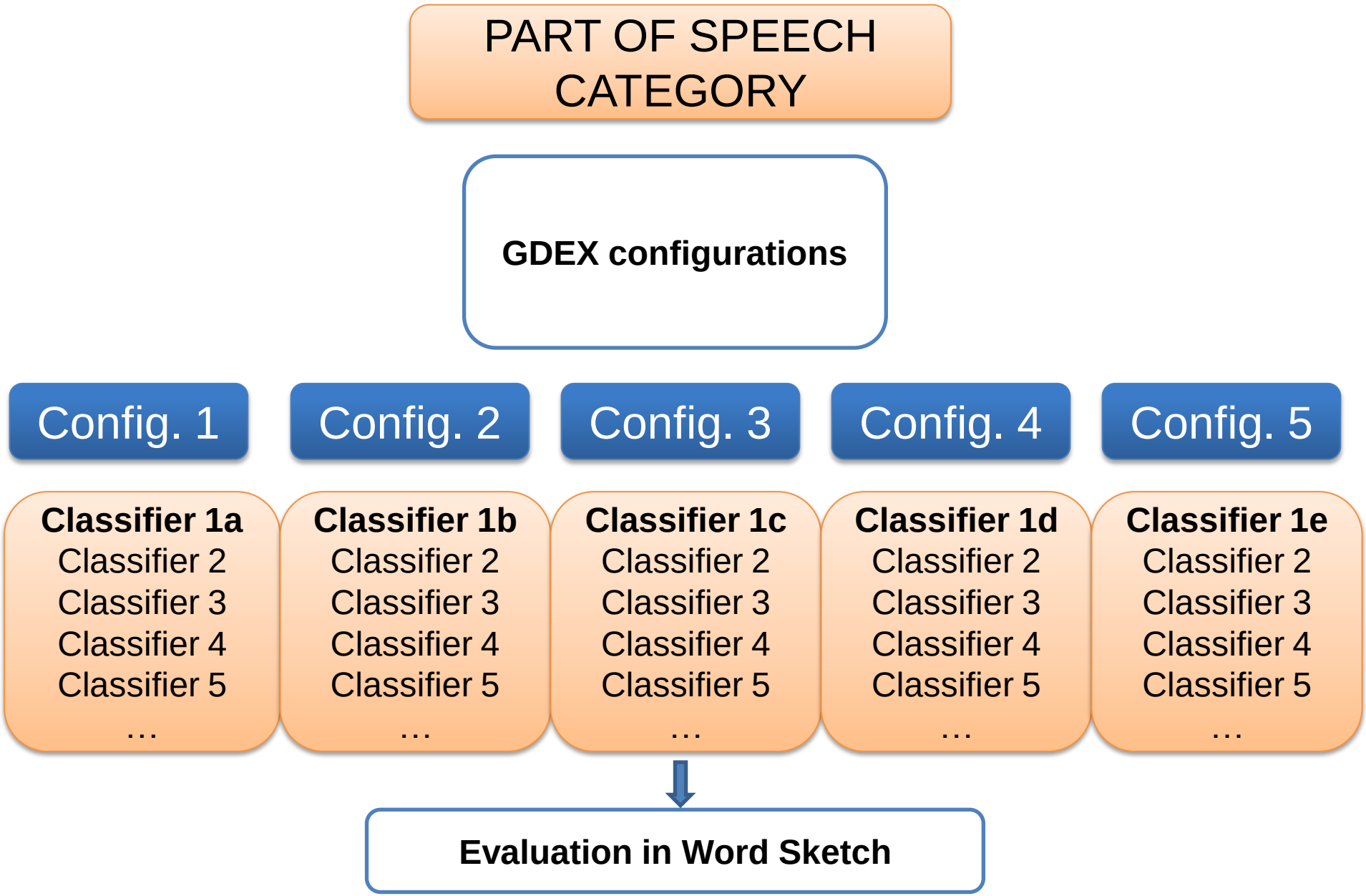
GDEXes for Phase 3

- Point of departure → GDEX for Slovene (Kosem, Husák and McCarthy, 2011)
- Aim: separate GDEX configurations for nouns, verbs, adjectives, adverbs
- Different task: **first 5 examples of each collocate need to be good** (not any 3 out of 10 examples)

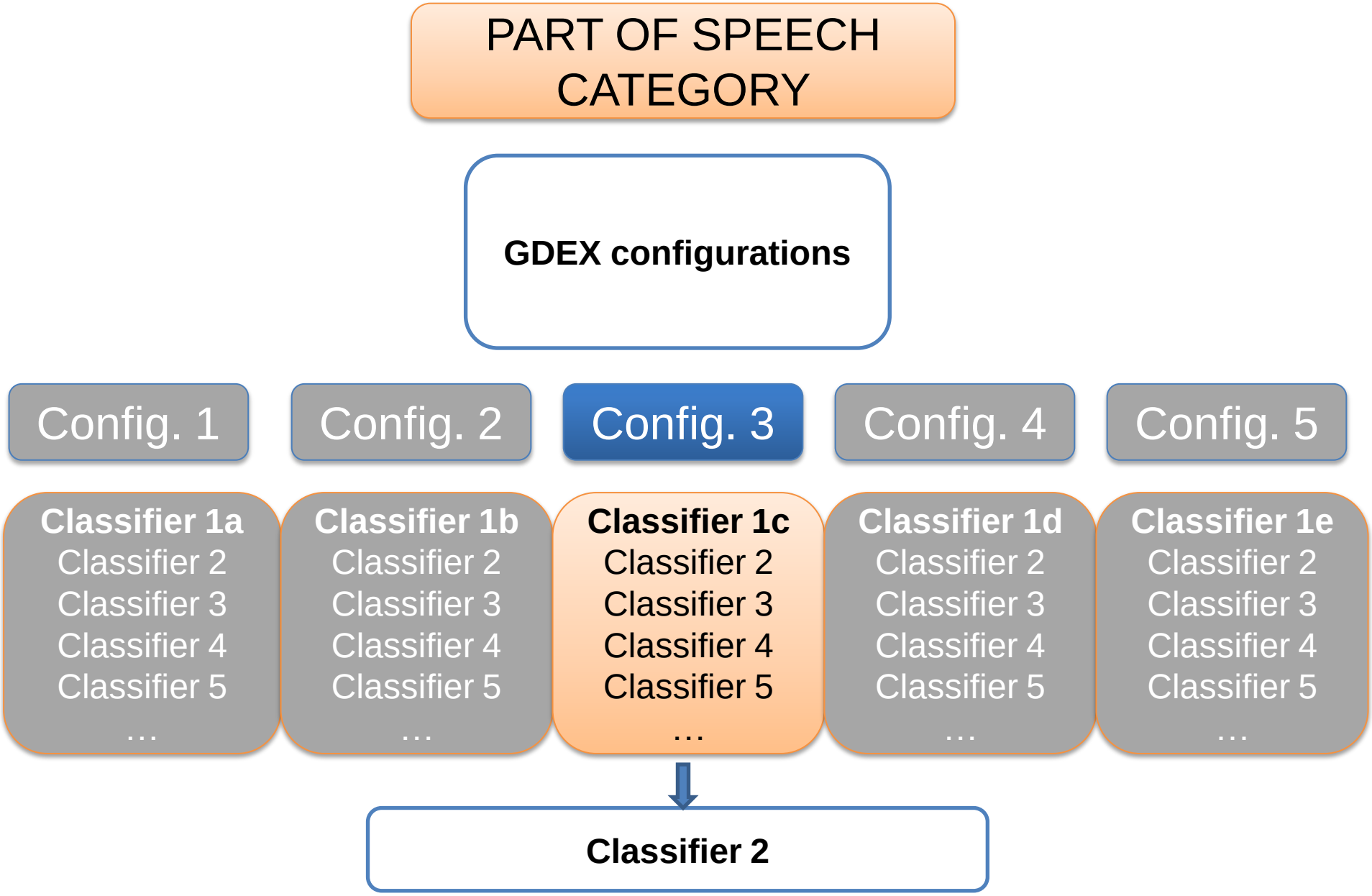
GDEX for Slovene – automatic selection



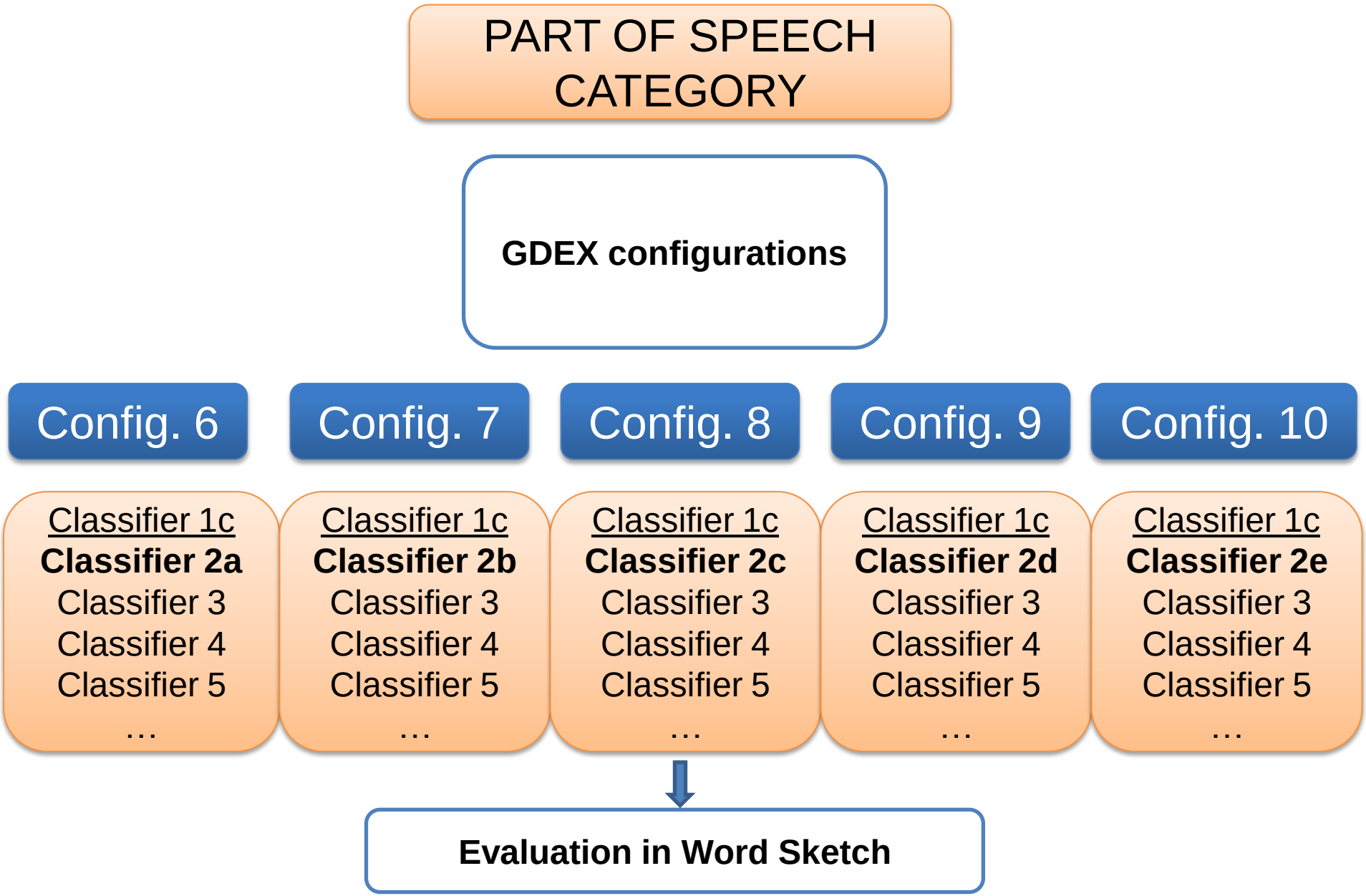
GDEX for Slovene – automatic selection



GDEX for Slovene – automatic selection



GDEX for Slovene – automatic selection



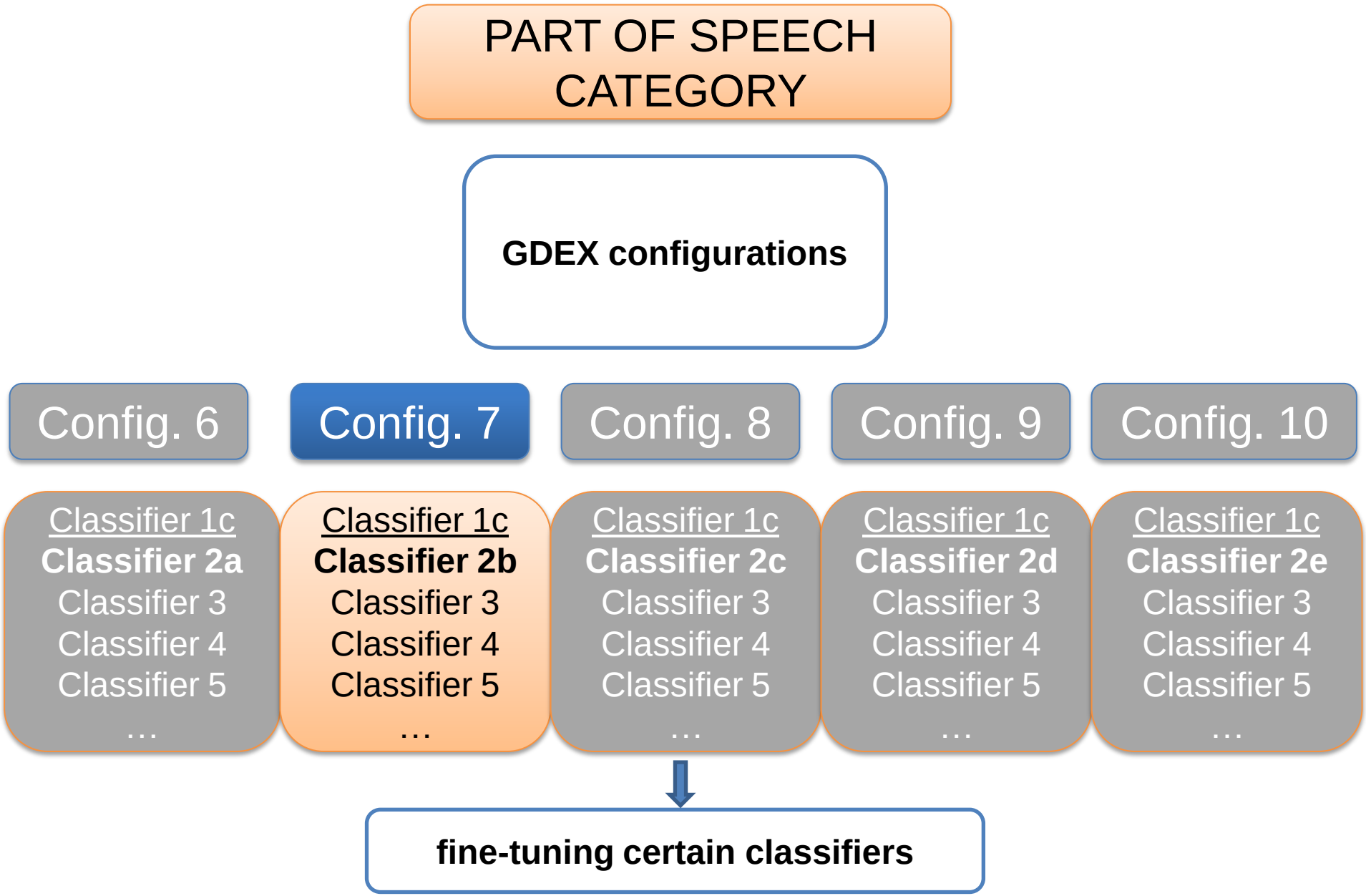
GDEX for Slovene – automatic selection

PART OF SPEECH CATEGORY

GDEX configurations

Config. 6	Config. 7	Config. 8	Config. 9	Config. 10
Classifier 1c	Classifier 1c	Classifier 1c	Classifier 1c	Classifier 1c
Classifier 2a	**Classifier 2b**	**Classifier 2c**	**Classifier 2d**	**Classifier 2e**
Classifier 3	Classifier 3	Classifier 3	Classifier 3	Classifier 3
Classifier 4	Classifier 4	Classifier 4	Classifier 4	Classifier 4
Classifier 5	Classifier 5	Classifier 5	Classifier 5	Classifier 5
…	…	…	…	…

GDEX for Slovene – automatic selection



GDEX for Slovene – automatic selection

PART OF SPEECH CATEGORY

GDEX configurations

Config. 11	Config. 12	Config. 13
Classifier 1c	Classifier 1c	Classifier 1c
Classifier 2b	Classifier 2b	Classifier 2c
Classifier 3x	Classifier 3x	Classifier 3x
Classifier 4	**Classifier 4x**	Classifier 4x
Classifier 5	Classifier 5	**Classifier 5x**
Evaluation in Word Sketch	**Evaluation in Word Sketch**	**Evaluation in Word Sketch**

Classifiers – no change

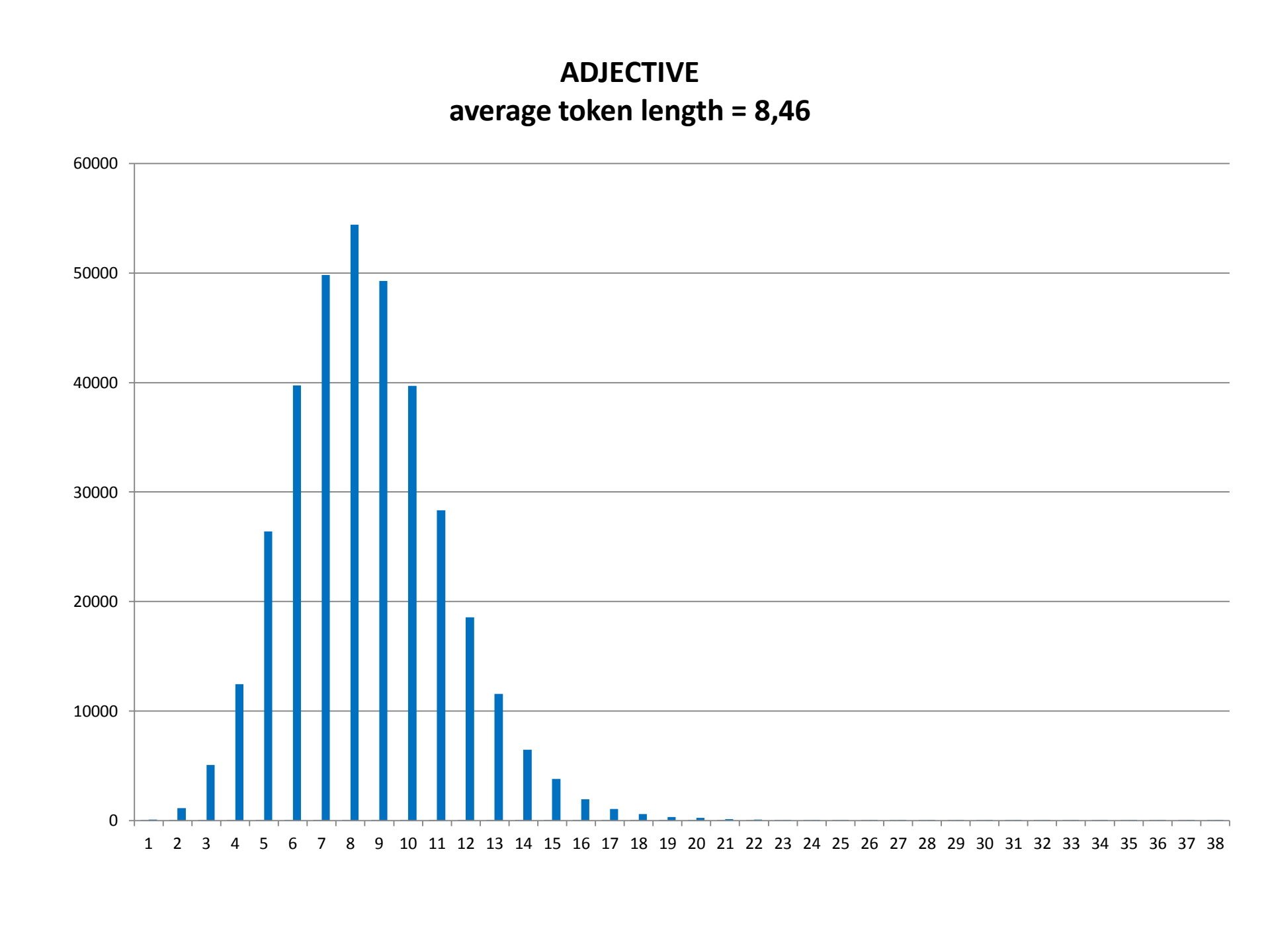
- Boolean classifier group (binary) (weight = 100)
  - Whole sentence
  - Classifier matching regexp ([<|\][>/\\])
  - Any token frequency < 3
  - maximum sentence length = 60 tokens
  - Contains token with less than one character
- “Penalty” classifiers
  - Proper nouns (weight = 2): -0.2 deduction for each proper noun
- Example diversity: Levenshtein distance > 30%

Fine-tuning of classifiers

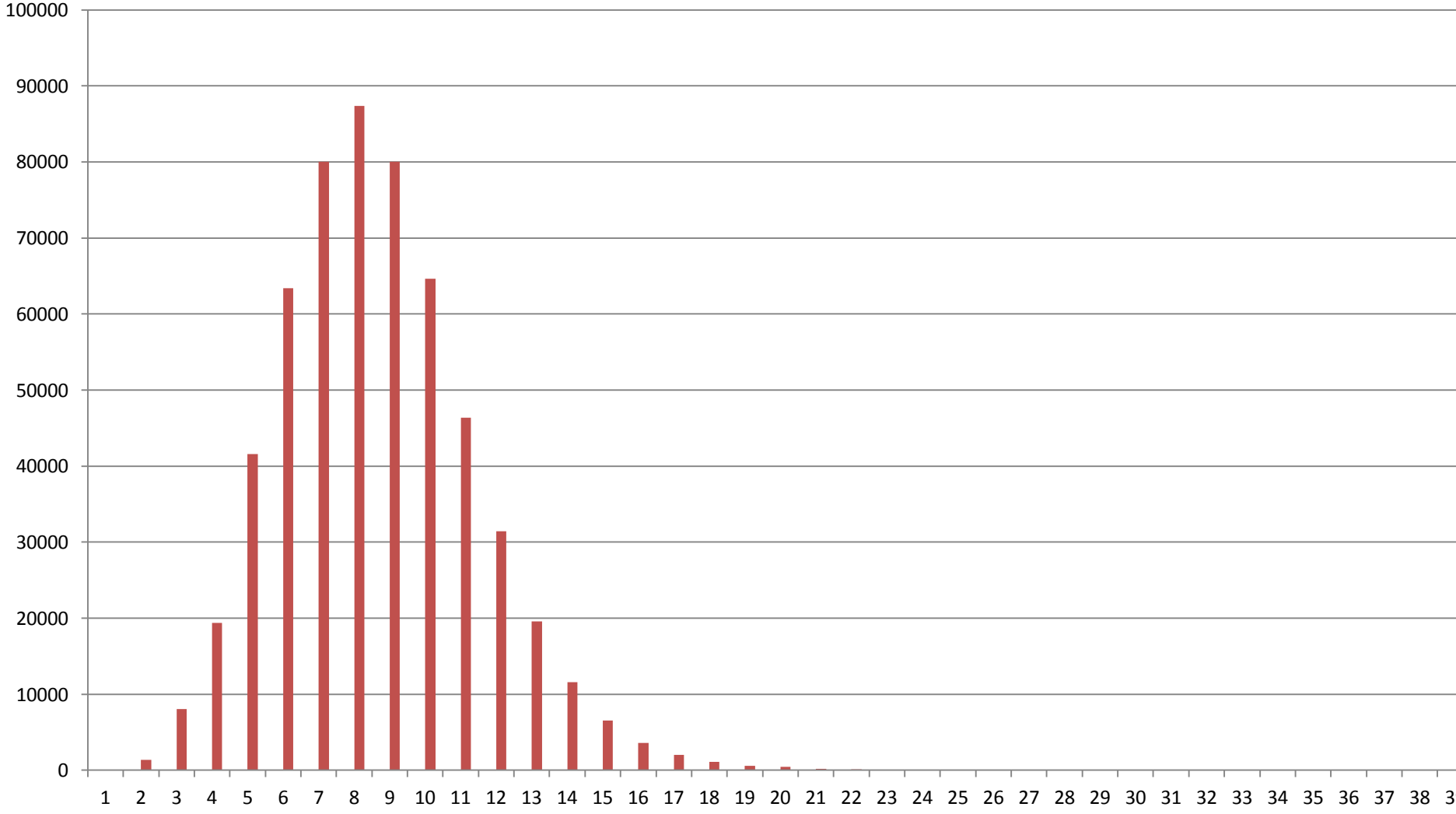
- Removed classifiers:
  - Boolean: maximum token length
  - Percentage of tokens with frequency above 104
- Repaired bugs:
  - Non-ASCII characters not recognized (e.g. ščž ČŠŽ)
    - Whole sentence
    - Capital letters
    - Mixed symbols
- Classifiers moved under boolean:
  - classifier penalizing web addresses, emails
  - keyword repetition (matching lemma, not token)

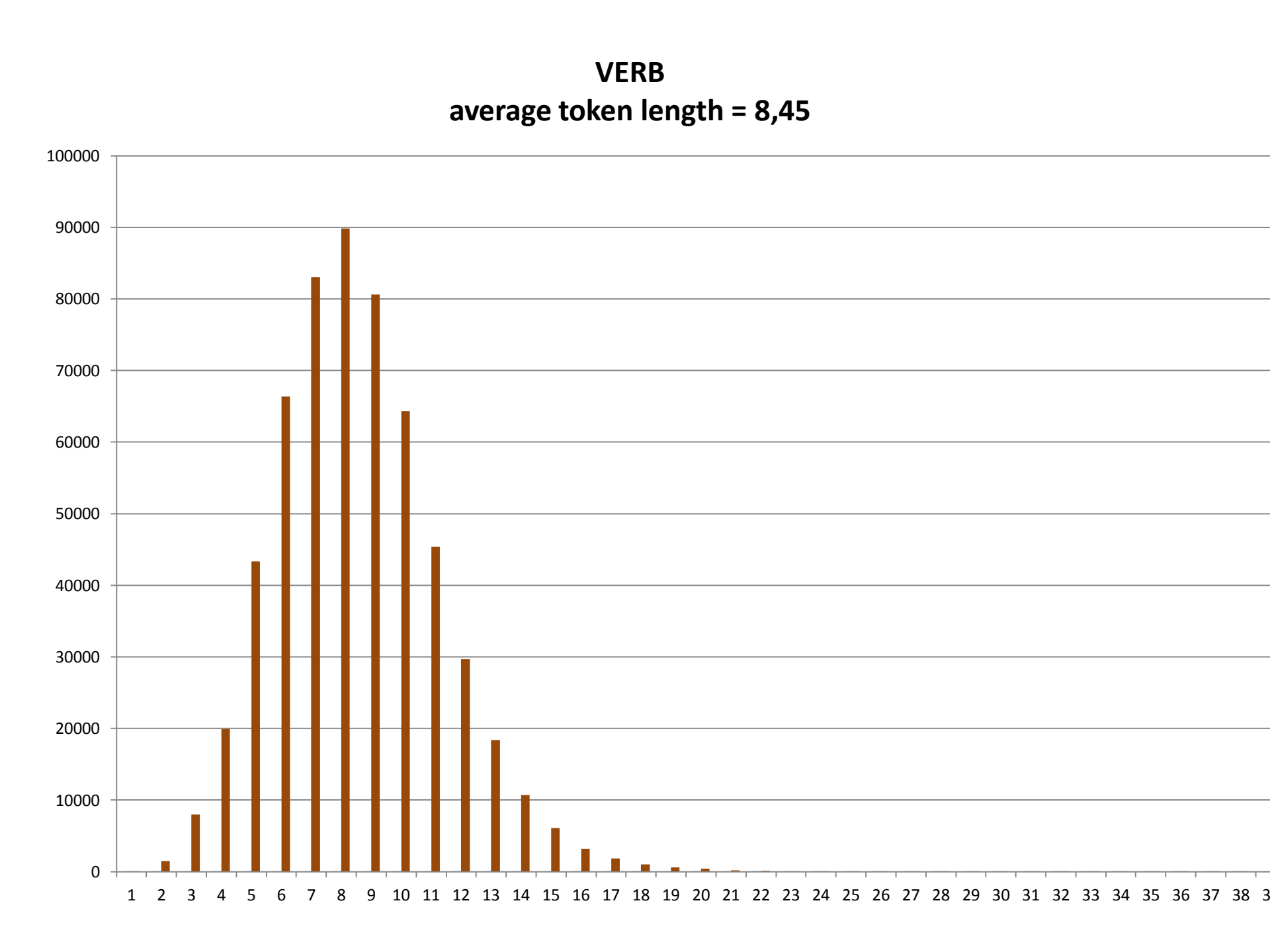
Fine-tuning of classifiers

- Changed classifiers:
  - Token length (originally 6 – from English GDEX)


ADJECTIVE
average token length = 8,46
60000
50000
40000
30000
20000
10000
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38

ADVERB
average token length = 8,54




VERB
average token length = 8,45
100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38

Fine-tuning of classifiers

- Changed classifiers:
  - Token length (originally 6 – from English GDEX)
  - Punctuation and symbols
    - added symbols: “ ’ \ * » « / _ – # “ ” • @ \ |= ’ ~ § × } { …
- Changed weights:
  - Sentence length (2 → 10)
  - Capital letters (2 → 4)
  - Symbols (1 → 5)
  - Punctuation (1 → 5)

New classifiers

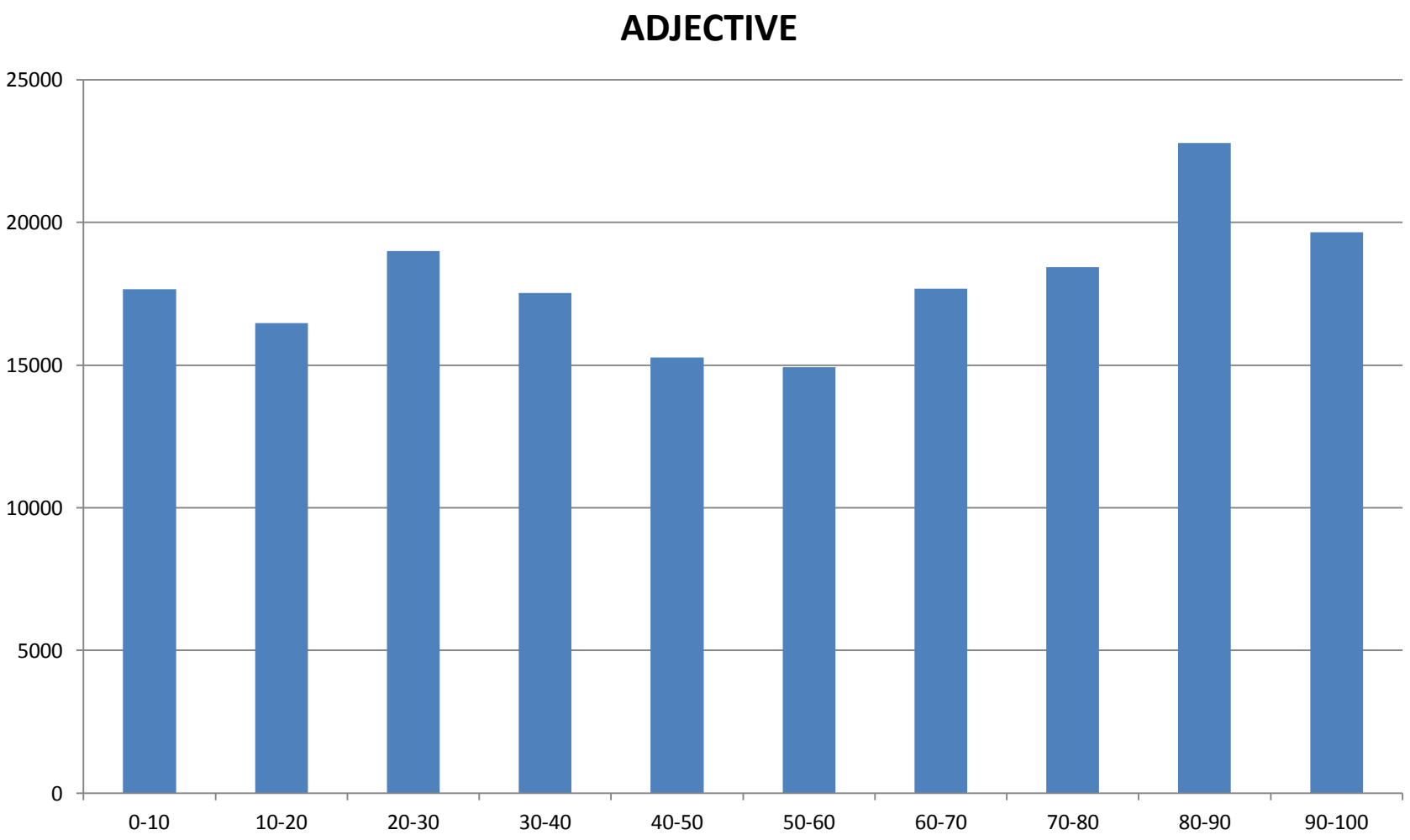
- Separate classifier for punctuation (comma, full stop, question mark, exclamation mark); weight = 5
- Added tag and lemma as an option of token attribute corpus classifier
- Complex phrases classifier
  - Based on tags
  - Sequences such as noun+noun+noun+noun or adj+adj+adj+noun
  - Evaluation showed very little value so the classifier not used at the moment

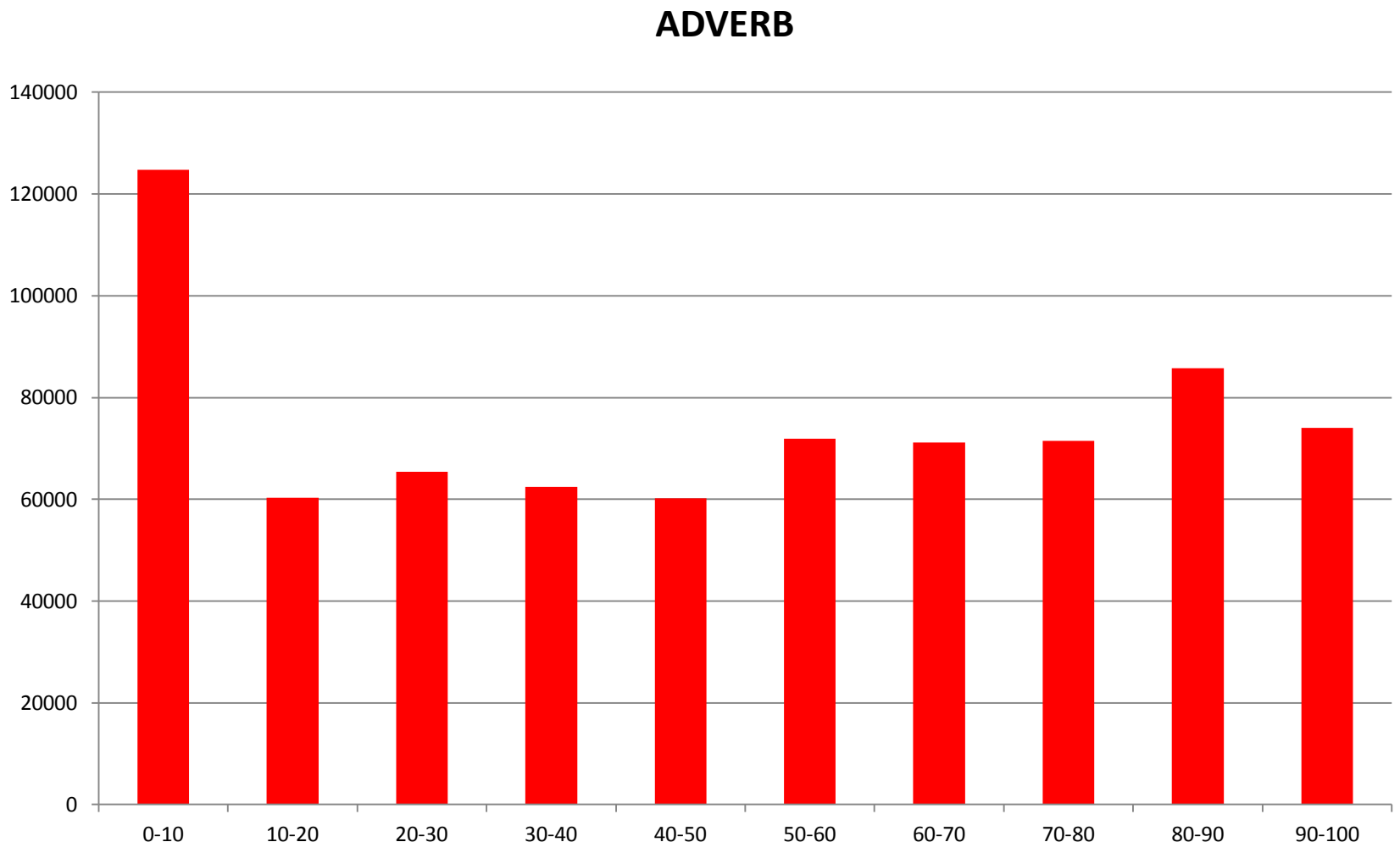
New classifiers

- Blacklist of sentence-initial words:
  - *sledi, zatorej, torej, nato, vendar, gre, oboji, dotlej, zato, tovrsten, to, ta, slednji, tak, takšen, potekati*
  - *both, it follows, thus, therefore, then, but, this is, till then, because, this type of, this, that, latter, it takes place*
- Blacklist of sentence-initial phrases
- Penalty for lemmas with frequency below 600 or 1000 (still testing)

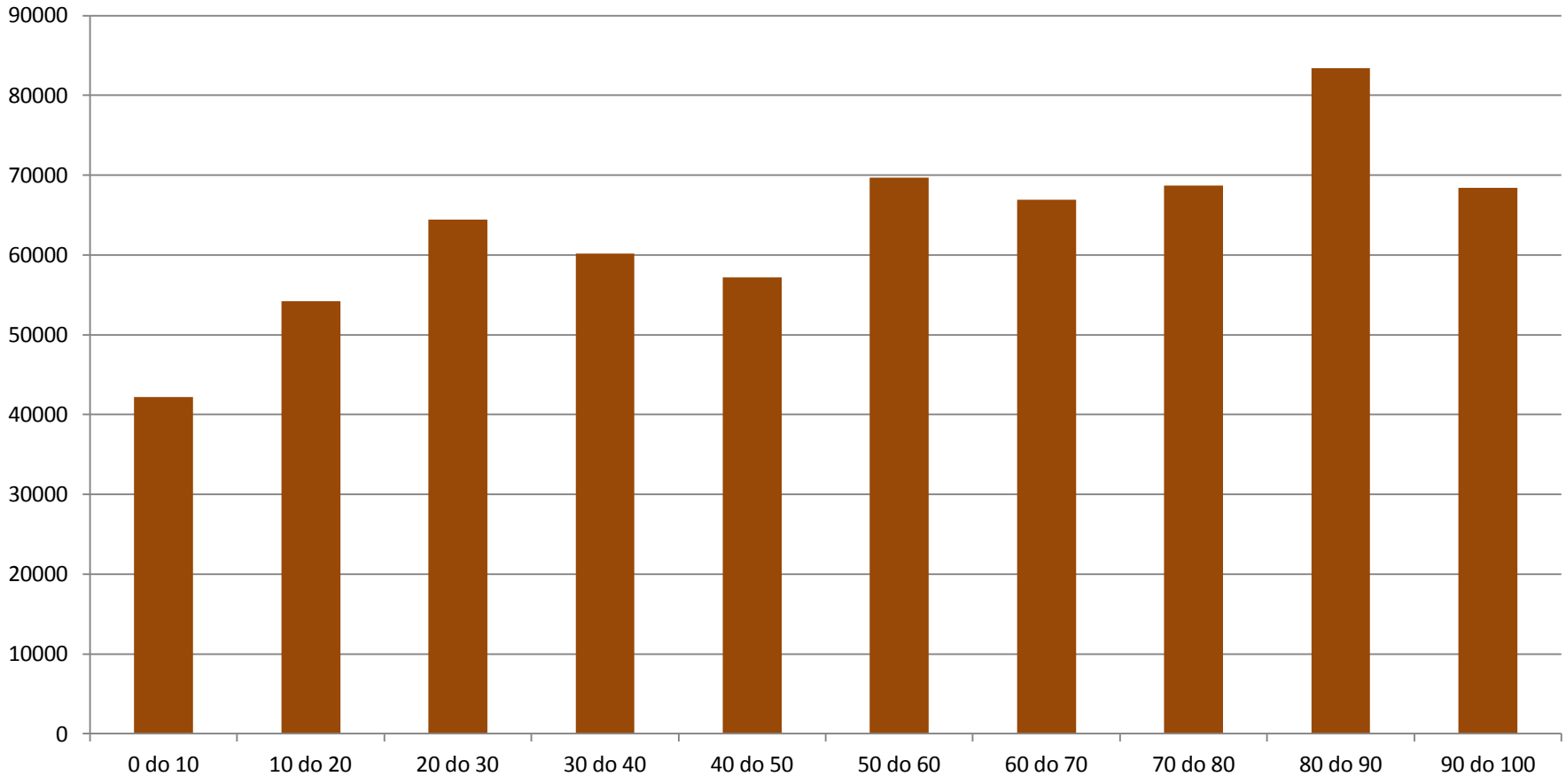
Differences observed between part of speech categories

- Keyword position


ADJECTIVE
25000
20000
15000
10000
5000
0
0-10
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90
90-100


ADVERB
140000
120000
100000
80000
60000
40000
20000
0
0-10
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90
90-100

VERB


90000
80000
70000
60000
50000
40000
30000
20000
10000
0
0 do 10
10 do 20
20 do 30
30 do 40
40 do 50
50 do 60
60 do 70
70 do 80
80 do 90
90 do 100

Differences observed between part of speech categories

- Keyword position
- Optimal sentence length
- Sentence initial words and phrases:
  - Adverbs: phrases containing adverbs need to be removed

To-do list

- More experimenting with weights
- More penalty for pronouns:
  - Penalties for pronouns “to” or “ta” in the first 30-40% of the example
- Preference to examples containing secondary collocates
  - Problem: how to get diversity of examples
- Further study of sentence-initial words and phrases (currently, blacklist based on observations during evaluation)
- Adapt the classifiers for the Gigafida corpus (Testing still done on FidaPLUS)