Uniwersytet Wrocławski

Reading Bibliographies: Methods of Semi-Automatic Categorization of Short Texts

prof. dr hab. Adam Pawłowski, dr Piotr Malak, dr Elżbieta Herden,
dr Tomasz Walkowiak, dr hab. Krzysztof Topolski

Acknowledgements:
this presentation was partly financed by the National Science Center Poland, project UMO-2016/23/B/HS2/01323 (Methods and tools of corpus linguistics in the research of bibliography of Polish publications from the period 1997-2017).

QUALICO 2018 , University of Wrocław, 5-8 July 2018

INTRODUCTION

Corpora – small reminder

1. Large definition:
   text corpus = any set of linguistic data;
2. Great reference corpora:
   text corpus = great, balanced collection of texts
   (*the bigger, the better* principle works)
3. Authorial corpora:
   text corpus = collection of texts of a single author
4. Monostyle corpora:
   text corpus = one style / genre collection (spoken, written, press, blogs, literary etc.)
5. Odd (unclassified) corpora:
   Sets of texts which have some common features but were not considered as potential corpora before

Corpus of data

1. Dataset: metadata records from Polish National Library (BN);
2. Corpus size: 553 000 records;
3. Contents: bibliographical records of books printed in Poland within the period of 20 years (1997-2017);
4. Format: MARC21 transcribed into JSON format;
   Coverage: all bibliographical data concerning books (not periodicals);

Access channel: BN API, http://data.bn.org.pl/docs/bibs

A complete record in a human readable form

BINMORE, Ken (1940-)
Teoria gier / Ken Binmore ; translation Iwona Konarzewska. – Łódź : Wydawnictwo Uniwersytetu Łódzkiego, 2017. – 206, [1] page : graphics, photos, charts ; 21 cm. – (Krótkie Wprowadzenie; 8)
Title of the original: *Game theory : a very short introduction*.

References on pages 195-200. Index.
Available also as e-book. – Publication financed by
Wydawnictwo Uniwersytetu Łódzkiego
ISBN 978-83-8088-594-3
ISBN 978-83-8088-595-0 (e-ISBN)
Type: Publikacje popularnonaukowe
Genre: Opracowanie
Creation time: 2007
Subject: Teoria gier
Domain: Filozofia i etyka

(620 characters)

519.83

A record in a machine readable form

Basic form of a record (98 characters):

Binmore Ken (2017), Teoria gier. Tłum. Iwona Konarzewska. Łódź: Wydawnictwo Uniwersytetu Łódzkiego

Full bibliographical record in MARC format (9324 characters, due to high redundancy):

{"id":5675461,"createdDate":"2017-08-07T13:50:59.000+02:00","updatedDate":"2017-11-06T14:54:28.000+01:00","language":"polski","subject":"Teoria gier","subjectPlace":"","subjectTime":"","subjectWork":"","isbnIssn":"9788380885943 9788380885950","author":"Binmore, Ken (1940-). Konarzewska, Iwona. Wydawnictwo Uniwersytetu Łódzkiego.","placeOfPublication":"Łódź : Polska","location":"","title":"Teoria gier / Game theory : a very short introduction, Krótkie Wprowadzenie ; 8","udc":"519.83 02","publisher":"Wydawnictwo Uniwersytetu Łódzkiego. Wydawnictwo Uniwersytetu Łódzkiego.","kind":"książka","domain":"Filozofia i etyka","formOfWork":"Książki Publikacje popularnonaukowe","genre":"Opracowanie","timePeriodOfCreation":"2007","audienceGroup":"","demographicGroup":"","nationalBibliographyNumber":"PB 2017/27081","publicationYear":"2017","languageOfOriginal":"angielski","fixedFields":[{"label":"LANG","value":"pol","display":"Polish","id":"24"},{"label":"COUNTRY","value":"pl ","display":"Polska","id":"89"},{"label":"CAT DATE","value":"2017-08-07","id":"28"},{"label":"CREATED","value":"2017-08-07T11:50:59Z","id":"83"},{"label":"MARCTYPE","value":" ","id":"107"},{"label":"REVISIONS","value":"9","id":"85"},{"label":"SUPPRESS","value":"b","id":"31"},{"label":"SKIP","value":"0","id":"25"},{"label":"REC TYPE","value":"b","id":"80"},{"label":"MAT TYPE","value":"a","display":"Book","id":"30"},{"label":"COPIES","value":"0","id":"27"},{"label":"PDATE","value":"2017-08-30T10:55:06Z","id":"98"},{"label":"BIB LVL","value":"m","display":"Monograph","id":"29"},{"label":"AGENCY","value":"1","id":"86"},{"label":"UPDATED","value":"2017-11-06T13:54:28Z","id":"84"},{"label":"RECORD #","value":"5675461","id":"81"},{"label":"LOCATION","value":"multi","id":"26"}],"varFields":[{"fieldTag":"a","marcTag":"100","ind1":"1","ind2":" ","subfields":[{"tag":"a","content":"Binmore, Ken"},{"tag":"d","content":"(1940-).","},{"tag":"e","content":"Autor"}]},{"fieldTag":"b","marcTag":"700","ind1":"1","ind2":" ","subfields":[{"tag":"a","content":"Konarzewska, Iwona.","},{"tag":"e","content":"Tłumaczenie"}]},{"fieldTag":"b","marcTag":"710","ind1":"2","ind2":" ","subfields":[{"tag":"a","content":"Wydawnictwo Uniwersytetu Łódzkiego.","},{"tag":"e","content":"Wydawca"},{"tag":"4","content":"pbl"}]},{"fieldTag":"d","marcTag":"380","ind1":" ","ind2":" ","subfields":[{"tag":"a","content":"Książki"}]},{"fieldTag":"d","marcTag":"380","ind1":" ","ind2":" ","subfields":[{"tag":"a","content":"Publikacje popularnonaukowe"}]},{"fieldTag":"d","marcTag":"388","ind1":"1","ind2":" ","subfields":[{"tag":"a","content":"2001-"}]},{"fieldTag":"d","marcTag":"650","ind1":" ","ind2":"4","subfields":[{"tag":"a","content":"Teoria gier"}]},{"fieldTag":"d","marcTag":"655","ind1":" ","ind2":"4","subfields":[{"tag":"a","content":"Opracowanie"}]},{"fieldTag":"d","marcTag":"658","ind1":" ","ind2":" ","subfields":[{"tag":"a","content":"Filozofia i etyka"}]},{"fieldTag":"g","marcTag":"015","ind1":" ","ind2":" ","subfields":[{"tag":"a","content":"PB 2017/27081"}]},{"fieldTag":"i","marcTag":"020","ind1":" ","ind2":" ","subfields":[{"tag":"a","content":"9788380885943"}]},{"fieldTag":"i","marcTag":"020","ind1":" ","ind2":" ","subfields":[{"tag":"a","content":"9788380885950"},{"tag":"q","content":"e-ISBN"}]},{"fieldTag":"j","marcTag":"080","ind1":" ","ind2":" ","subfields":[{"tag":"a","content":"519.83"}]},{"fieldTag":"l","marcTag":"998","ind1":" ","ind2":" ","subfields":[{"tag":"a","content":"ik"}]},{"fieldTag":"n","marcTag":"504","ind1":" ","ind2":" ","subfields":[{"tag":"a","content":"Bibliografia na stronach 195-200. Indeks."}]},{"fieldTag":"n","marcTag":"530","ind1":" ","ind2":" ","subfields":[{"tag":"a","content":"Dostępne także jako e-book."}]},{"fieldTag":"n","marcTag":"536","ind1":" ","ind2":" ","subfields":[{"tag":"a","content":"Publikacja sfinansowana ze środków Wydawnictwa Uniwersytetu Łódzkiego"}]},{"fieldTag":"p","marcTag":"260","ind1":" ","ind2":" ","subfields":[{"tag":"a","content":"Łódź :"},{"tag":"b","content":"Wydawnictwo Uniwersytetu Łódzkiego,"},{"tag":"c","content":"2017."}]},{"fieldTag":"r","marcTag":"300","ind1":" ","ind2":" ","subfields":[{"tag":"a","content":"206, [1] strona :"},{"tag":"b","content":"ilustracje, fotografie, wykresy ;"},{"tag":"c","content":"21 cm."}]},{"fieldTag":"s","marcTag":"490","ind1":"1","ind2":" ","subfields":[{"tag":"a","content":"Krótkie Wprowadzenie ;"},{"tag":"v","content":"8"}]},{"fieldTag":"s","marcTag":"830","ind1":" ","ind2":"0","subfields":[{"tag":"a","content":"Krótkie Wprowadzenie ;"},{"tag":"v","content":"8"}]},{"fieldTag":"t","marcTag":"245","ind1":"1","ind2":"0","subfields":[{"tag":"a","content":"Teoria gier /"},{"tag":"c","content":"Ken Binmore ; tłumaczenie Iwona Konarzewska."}]},{"fieldTag":"u","marcTag":"246","ind1":"1","ind2":" ","subfields":[{"tag":"i","content":"Tytuł oryginału:"},{"tag":"a","content":"Game theory :"},{"tag":"b","content":"a very short introduction,"},{"tag":"f","content":"2007"}]},{"fieldTag":"y","marcTag":"008","ind1":" ","ind2":" ","content":"170807s2017 pl aod |001 0 pol nam i "},{"fieldTag":"y","marcTag":"040","ind1":" ","ind2":" ","subfields":[{"tag":"a","content":"WA N"},{"tag":"c","content":"WA N"}]},{"fieldTag":"y","marcTag":"041","ind1":"1","ind2":" ","subfields":[{"tag":"a","content":"pol"},{"tag":"h","content":"eng"}]},{"fieldTag":"y","marcTag":"046","ind1":" ","ind2":" ","subfields":[{"tag":"k","content":"2007"}]},{"fieldTag":"y","marcTag":"336","ind1":" ","ind2":" ","subfields":[{"tag":"a","content":"Tekst"},{"tag":"b","content":"txt"},{"tag":"2","content":"rdacontent"}]},{"fieldTag":"y","marcTag":"337","ind1":" ","ind2":" ","subfields":[{"tag":"a","content":"Bez urządzenia pośredniczącego"},{"tag":"b","content":"n"},{"tag":"2","content":"rdamedia"}]},{"fieldTag":"y","marcTag":"338","ind1":" ","ind2":" ","subfields":[{"tag":"a","content":"Wolumin"},{"tag":"b","content":"nc"},{"tag":"2","content":"rdacarrier"}]},{"fieldTag":"y","marcTag":"920","ind1":" ","ind2":" ","subfields":[{"tag":"a","content":"978-83-8088-594-3"}]},{"fieldTag":"y","marcTag":"920","ind1":" ","ind2":" ","subfields":[{"tag":"a","content":"978-83-8088-595-0 (e-ISBN)"}]},{"fieldTag":"y","marcTag":"999","ind1":" ","ind2":" ","subfields":[{"tag":"a","content":"zkd"},{"tag":"b","content":"eoaw"},{"tag":"x","content":"33"},{"tag":"y","content":"17"}]},{"fieldTag":"y","marcTag":"084","ind1":" ","ind2":" ","subfields":[{"tag":"a","content":"02"}]},{"fieldTag":"_","content":"00000nam a2200517 i 4500"}],"marc":{"leader":"00000nam a2200517 i 4500","fields":[{"001":"b56754619"},{"008":"170807s2017 pl aod |001 0 pol nam i "},{"015":{"ind1":" ","ind2":" ","subfields":[{"a":"PB 2017/27081"}]}},{"020":{"ind1":" ","ind2":" ","subfields":[{"a":"9788380885943"}]}},{"020":{"ind1":" ","ind2":" ","subfields":[{"a":"9788380885950"},{"q":"e-ISBN"}]}},{"040":{"ind1":" ","ind2":" ","subfields":[{"a":"WA N"},{"c":"WA N"}]}},{"041":{"ind1":"1","ind2":" ","subfields":[{"a":"pol"},{"h":"eng"}]}},{"046":{"ind1":" ","ind2":" ","subfields":[{"k":"2007"}]}},{"080":{"ind1":" ","ind2":" ","subfields":[{"a":"519.83"}]}},{"084":{"ind1":" ","ind2":" ","subfields":[{"a":"02"}]}},{"100":{"ind1":"1","ind2":" ","subfields":[{"a":"Binmore, Ken"},{"d":"(1940-).","},{"e":"Autor"}]}},{"245":{"ind1":"1","ind2":"0","subfields":[{"a":"Teoria gier /"},{"c":"Ken Binmore ; tłumaczenie Iwona Konarzewska."}]}},{"246":{"ind1":"1","ind2":" ","subfields":[{"i":"Tytuł oryginału:"},{"a":"Game theory :"},{"b":"a very short introduction,"},{"f":"2007"}]}},{"260":{"ind1":" ","ind2":" ","subfields":[{"a":"Łódź :"},{"b":"Wydawnictwo Uniwersytetu Łódzkiego,"},{"c":"2017."}]}},{"300":{"ind1":" ","ind2":" ","subfields":[{"a":"206, [1] strona :"},{"b":"ilustracje, fotografie, wykresy ;"},{"c":"21 cm."}]}},{"336":{"ind1":"

Meaningful elements of a record

{"id":5675461,"createdDate":"2017-08-07T13:50:59.000+02:00","updatedDate":"2017-11-06T14:54:28.000+01:00","language":"polski","subject":"Teoria gier","subjectPlace":"","subjectTime":"","subjectWork":"","isbnIssn":"9788380885943 9788380885950","author":"Binmore, Ken (1940-). Konarzewska, Iwona. Wydawnictwo Uniwersytetu Łódzkiego.","placeOfPublication":"Łódź : Polska","location":"","title":"Teoria gier / Game theory : a very short introduction, Krótkie Wprowadzenie ; 8","udc":"519.83 02","publisher":"Wydawnictwo Uniwersytetu Łódzkiego. Wydawnictwo Uniwersytetu Łódzkiego,","kind":"książka","domain":"Filozofia i etyka","formOfWork":"Książki Publikacje popularnonaukowe","genre":"Opracowanie","timePeriodOfCreation":"2007","audienceGroup":"","demographicGroup":"","nationalBibliographyNumber":"PB 2017/27081","publicationYear":"2017","languageOfOriginal":"angielski",

Linguistically valuable parts

What is appropriate for linguistic analysis?

Polish title: *Teoria gier: krótkie wprowadzenie*

Original title: *Game theory: a very short introduction*

Some metadata:

- Author
- Publisher
- Place of publication
- Year of publication
- Genre
- Subject
- Domain
- Universal Decimal Classification number

METHODS

Methods

1. Preprocessing

– MARC-to-XML translation
– extraction and structuring of relevant fields
– linguistic preprocessing (POS tagging, lemmatization)

Problems

Records provide automatically generated *author* field, but it contains all contributors to the book (author, translator, etc.)

Methods

2. Data processing and quantitative analysis

- basic statistics
- POS statistics, frequency list, concordances
- categorisation (based on metadata)
- classification of short texts (fastText)
- additionally: distribution fitting to discriminate between general language and titles

TITLES & GENERAL LANGUAGE: COMPARING TWO CORPORA

Comparing two corpora

Criteria:

1) Vocabulary
2) Basic statistics
3) Statistical distributions of word spectra

Uniwersytet Wrocławski

Corpus of bibliography: the most frequent words

word	frequency	word	frequency	word	frequency
i (and)	157249	**2** (=vol.)	17856	**podręcznik** (handbook)	11513
w (in)	141716	**część** (part)	17222	**a** (and, or, vs.)	11157
z (with, from)	67378	**Polska** (Poland)	17222	**jak** (how, as)	10360
na (on)	39821	**T** (vol.)	15678	**ćwiczenia** (excercises)	10299
dla (for)	35629	**lato** (summer / year)	14068	**od** (from, since)	10052
do (to)	33065	**szkoła** (school)	13855	**wybrana** (chosen)	9638
o (about)	27324	**historia** (history)	12768	**rok** (year)	9543
polski (Polish)	21794	**klasa** (class)	12554	**być** (to be)	9394
1 (=vol.)	20037	**materiał** (contents)	12313	**zbiorowy** (collective)	9375
praca (work)	19933	**życie** (life)	11787	**dziecko** (child)	9021

Uniwersytet Wrocławski

Corpus of bibliography: POS frequencies

POS	Frequency	Fraction
subst	2037937	53,9%
adj	479559	12,7%
prep	391053	10,3%
num	232774	6,2%
conj	212496	5,6%
adv	47934	1,3%
ppas	28402	0,8%
ger	27881	0,7%
brev	26716	0,7%
fin	24977	0,7%
inf	20141	0,5%
qub	17600	0,5%
ppron3	8978	0,2%
comp	8928	0,2%
depr	8811	0,2%
impt	8328	0,2%
other	200680	5,3%

Uniwersytet Wrocławski

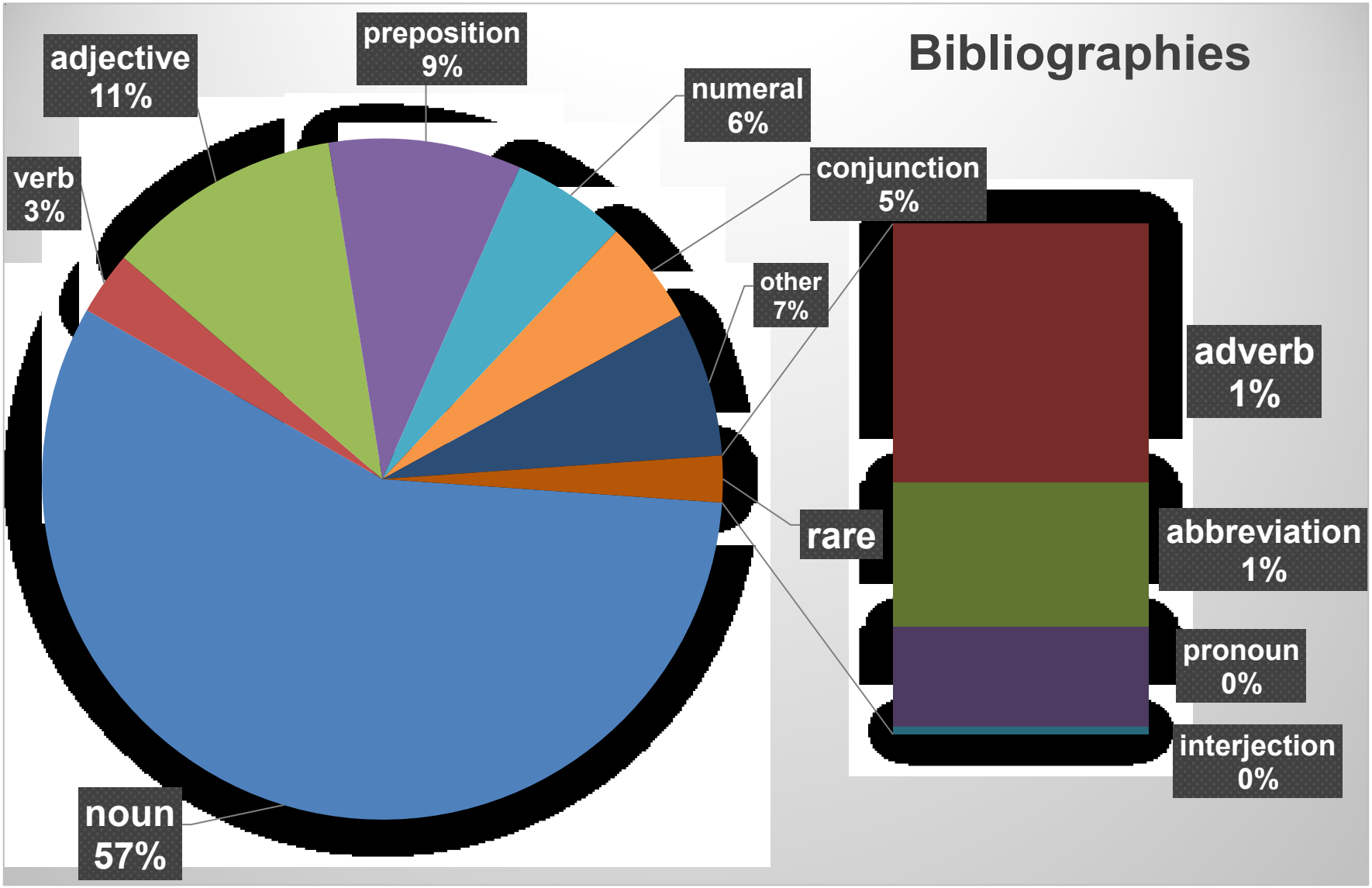
General language vs titles: POS frequencies

NKJP (263,754,400 tokens)		
POS	Frequency	Fraction
noun	114,607,420	43,45%
verb	40,203,046	15,24%
preposition	28,762,239	10,90%
adjective	28,341,959	10,75%
other	21,853,298	8,29%
conjunction	10,442,593	3,96%
adverb	10,308,830	3,91%
pronoun	5,201,486	1,97%
abbreviation	2,201,422	0,83%
numeral	1,627,941	0,62%
interjection	204,166	0,08%

Bibliographies (4,278,774 tokens)		
POS	Frequency	Fraction
noun	2,445,525	57.15%
verb	128,439	3.00%
adjective	479,559	11.21%
preposition	391,053	9.14%
numeral	232,774	5.44%
conjunction	212,496	4.97%
other	294,341	6.88%
adverb	47,934	1.12%
abbreviation	26,716	0.62%
pronoun	18,555	0.43%
interjection	1,382	0.03%

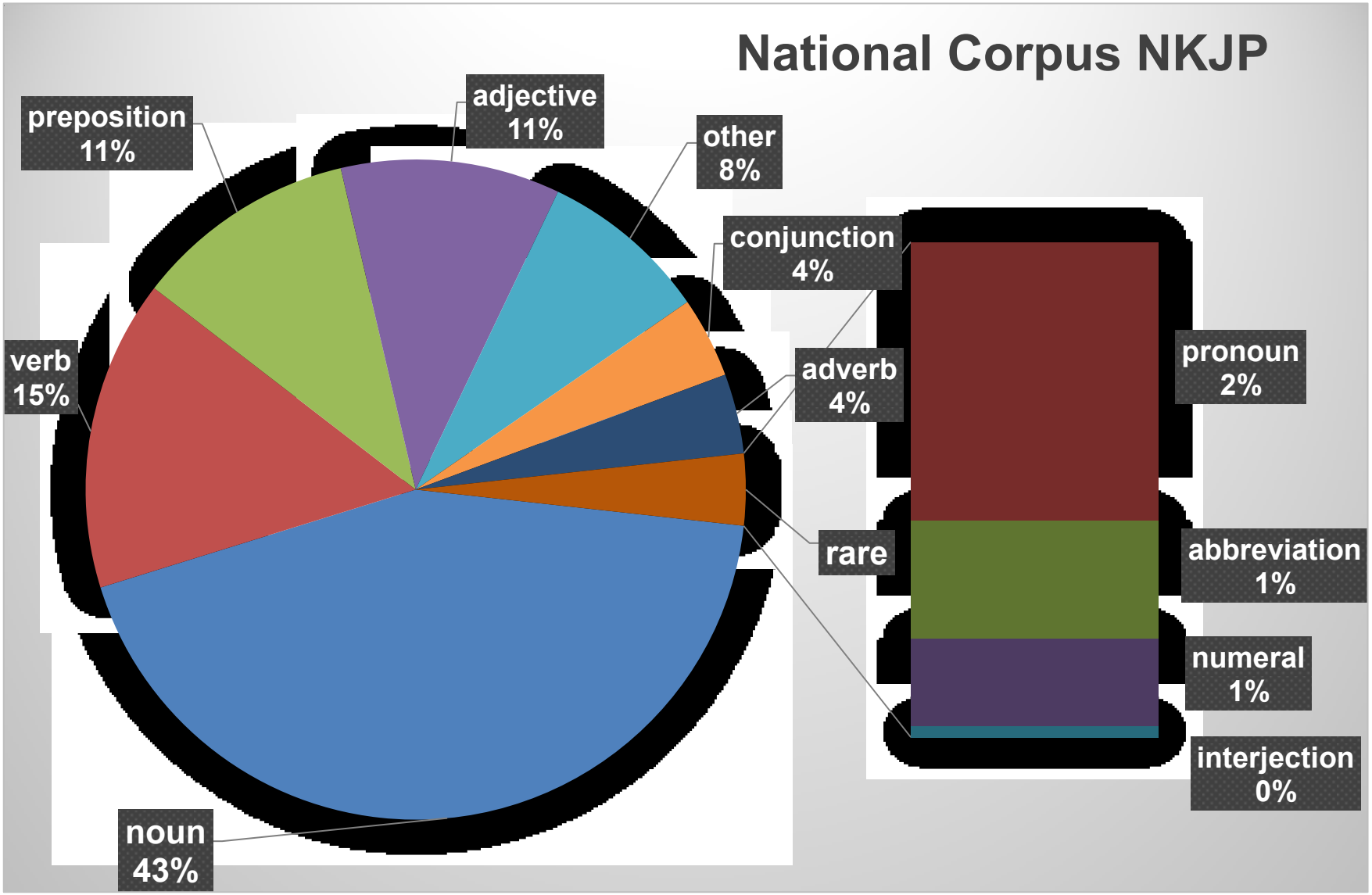
Uniwersytet Wrocławski

Corpus of titles: POS frequencies



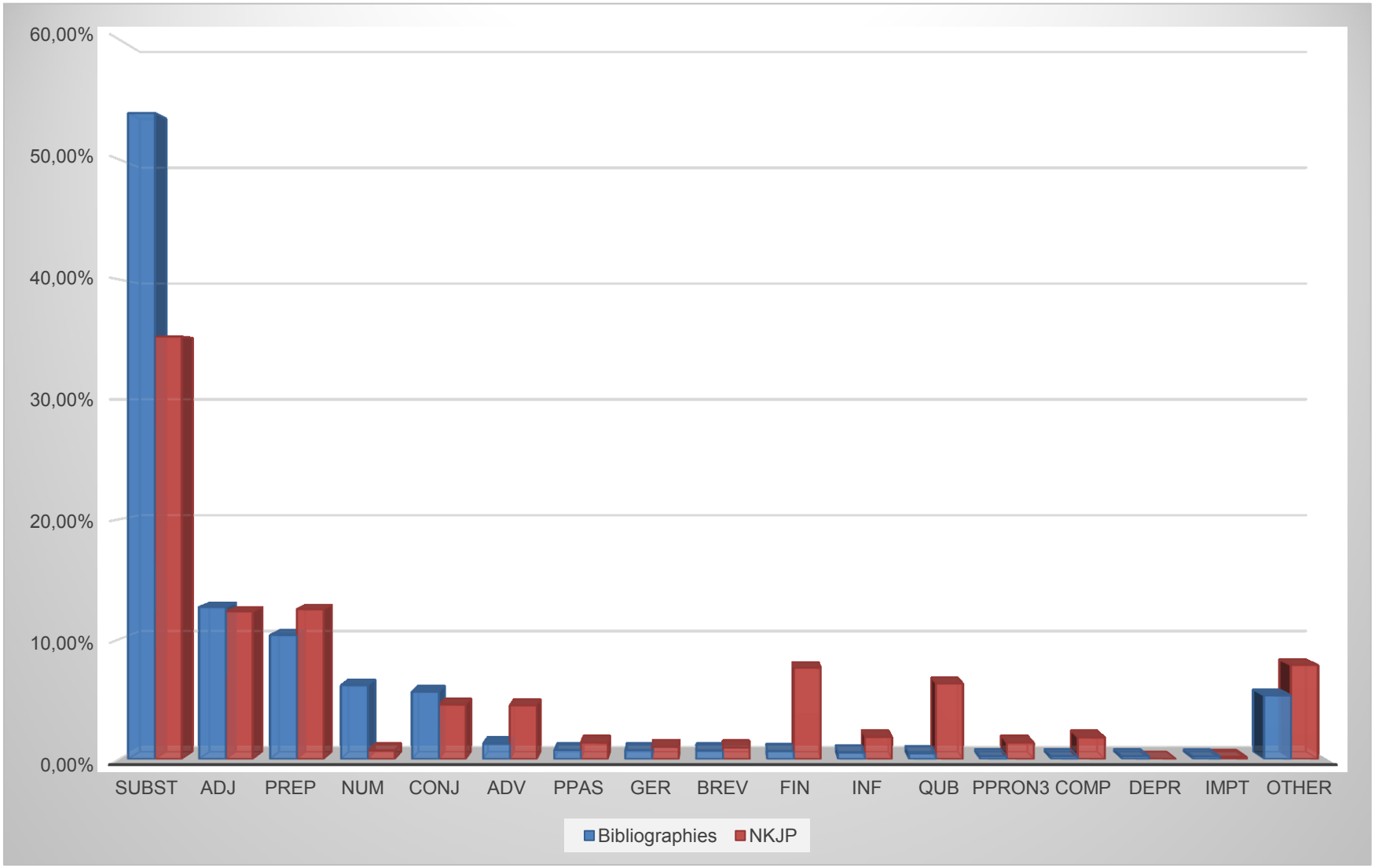
Uniwersytet Wrocławski

General language (NKJP): POS frequencies



Uniwersytet Wrocławski

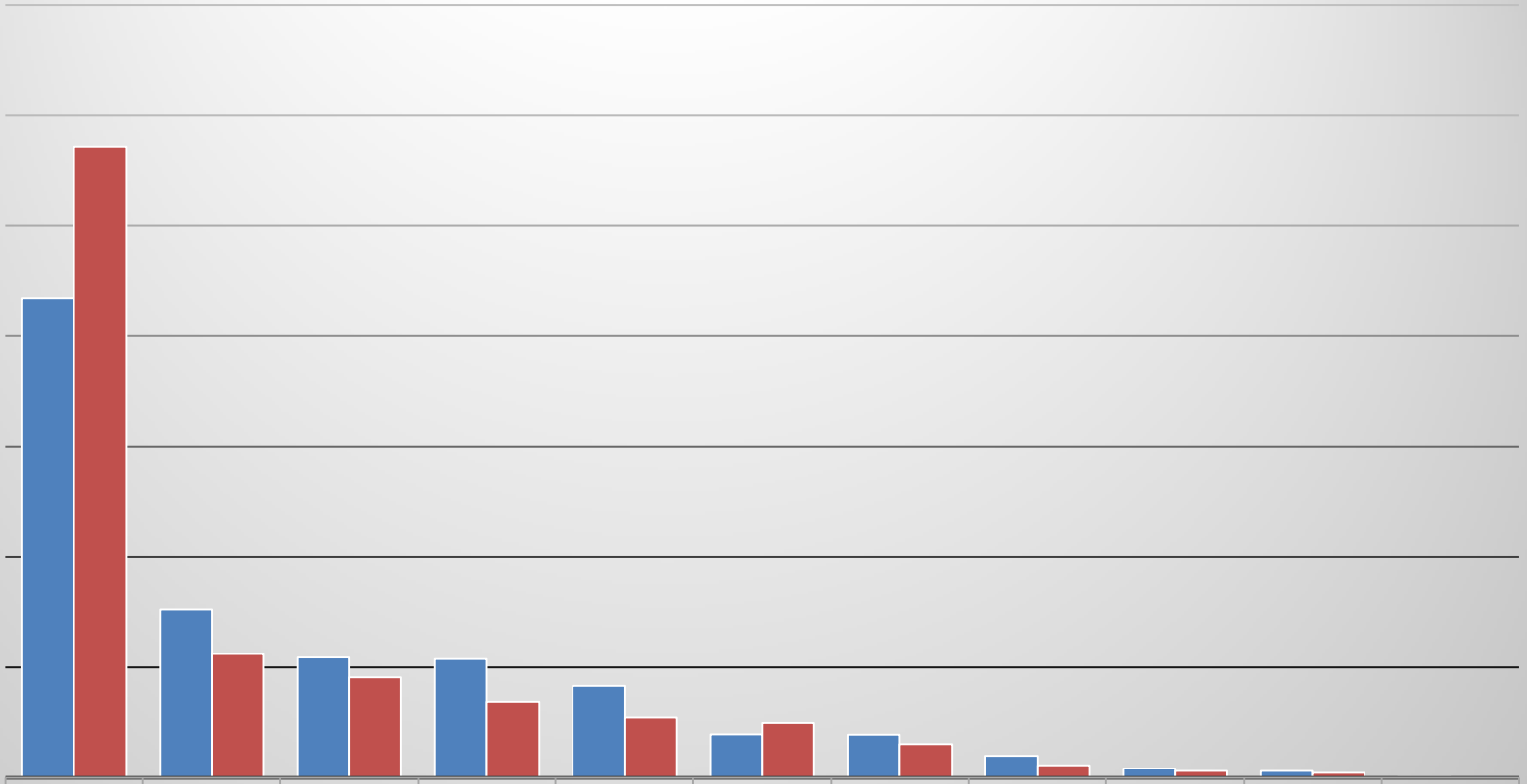
POS in titles and in general language (no verb category)



Uniwersytet Wrocławski

POS in titles and in general language

POS frequencies comparison



	noun	verb	preposition	adjective	other	conjunction	adverb	pronoun	abbreviation	numeral	interjection
NKJP	43,45%	15,24%	10,90%	10,75%	8,29%	3,96%	3,91%	1,97%	0,83%	0,62%	0,08%
Bibliographies	57,15%	11,21%	9,14%	6,88%	5,44%	4,97%	3,00%	1,12%	0,62%	0,43%	0,00032299

NKJP Bibliographies

Conclusions

1. Titles are nominal (high percentage of nouns, fewer pure verbal forms, few adverbs)
2. Relatively high participation of quasi-verbal forms: gerunds and participles
3. Titles include many words related to genre (*handbook*, *material* (PL *materiał*), *exercises*, *selected*, *collective* etc.)

COMPARING TWO CORPORA: WORD SPECTRA DISTRIBUTIONS

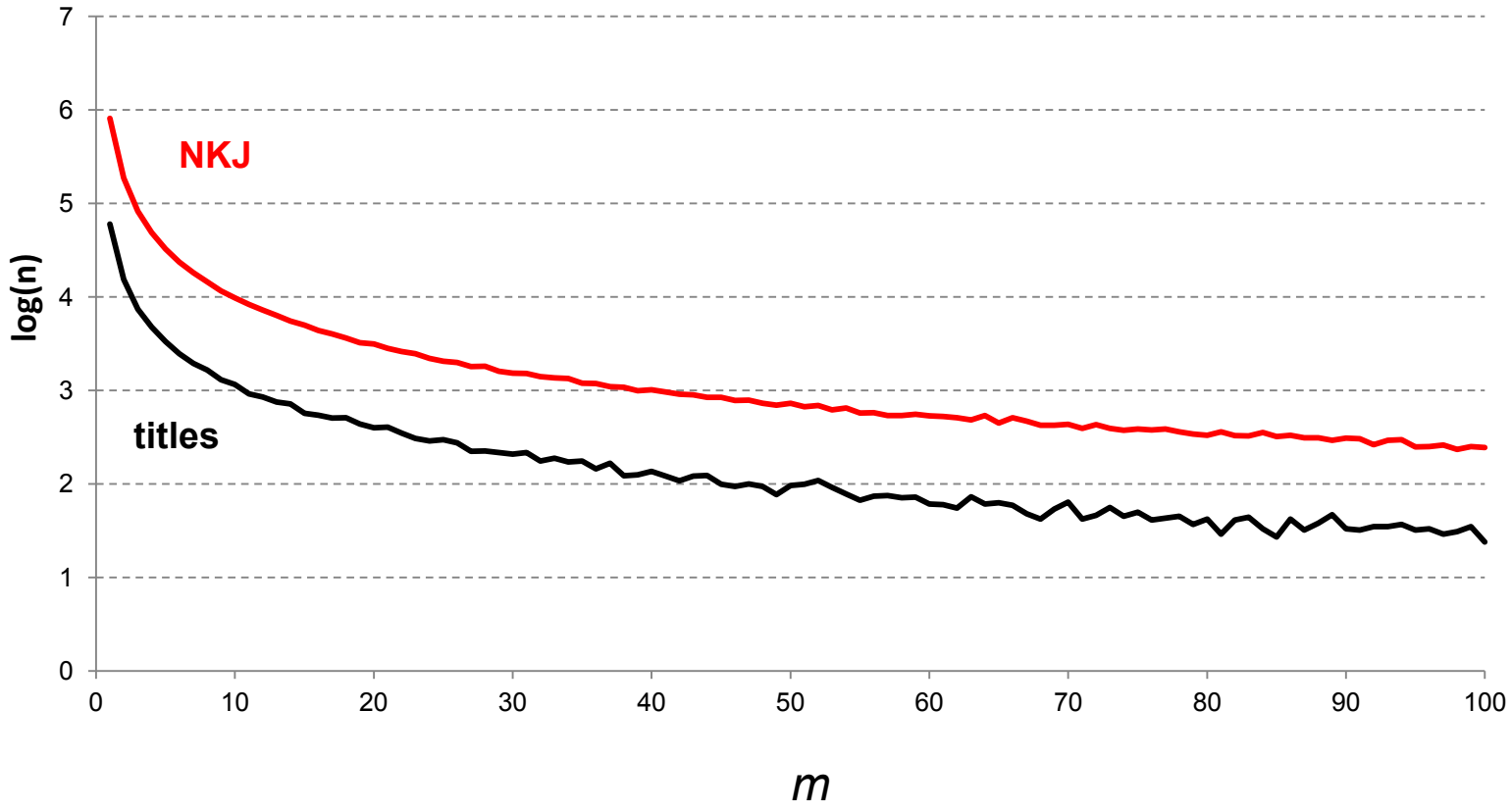
Uniwersytet Wrocławski

Distribution of lemmas frequencies

Book titles (3,539,644 lemmas)		
Occurences	**Frequency of occurences**	**Fraction**
1	74989	2,12%
2	16511	0,47%
3	7652	0,22%
4	4815	0,14%
5	3266	0,09%
6	2430	0,07%
7	1966	0,06%
8	1643	0,05%
9	1241	0,04%
10	1070	0,03%
11	931	0,03%
12	874	0,02%
13	721	0,02%
14	684	0,02%
15	563	0,02%

NKJP (236,956,885 lemmas)		
Occurences	**Frequency of occurences**	**Fraction**
1	808047	0,3410%
2	186939	0,0789%
3	82286	0,0347%
4	48763	0,0206%
5	32497	0,0137%
6	23640	0,0100%
7	18154	0,0077%
8	14450	0,0061%
9	11601	0,0049%
10	9791	0,0041%
11	8293	0,0035%
12	7220	0,0030%
13	6352	0,0027%
14	5499	0,0023%
15	4977	0,0021%

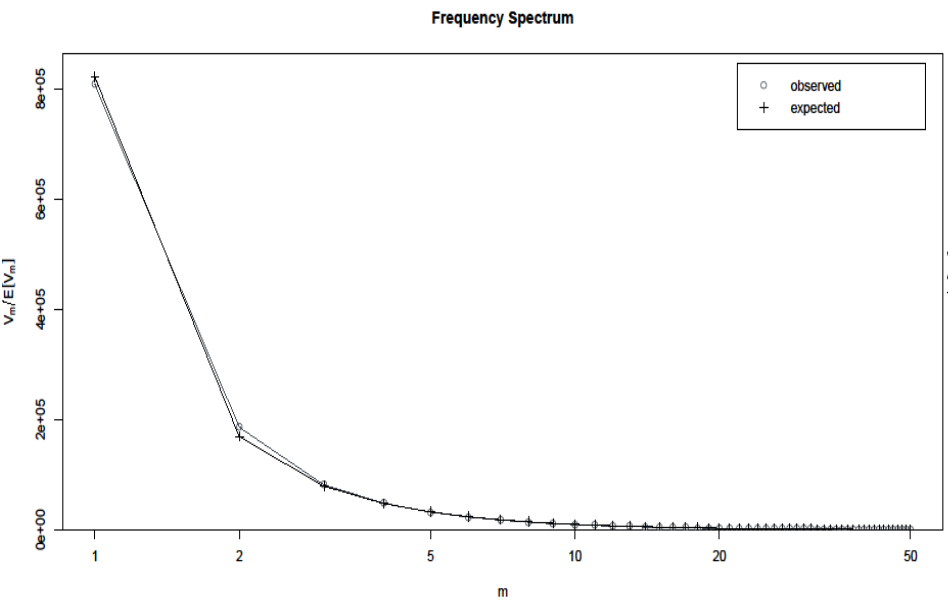
General language vs titles corpus (1)

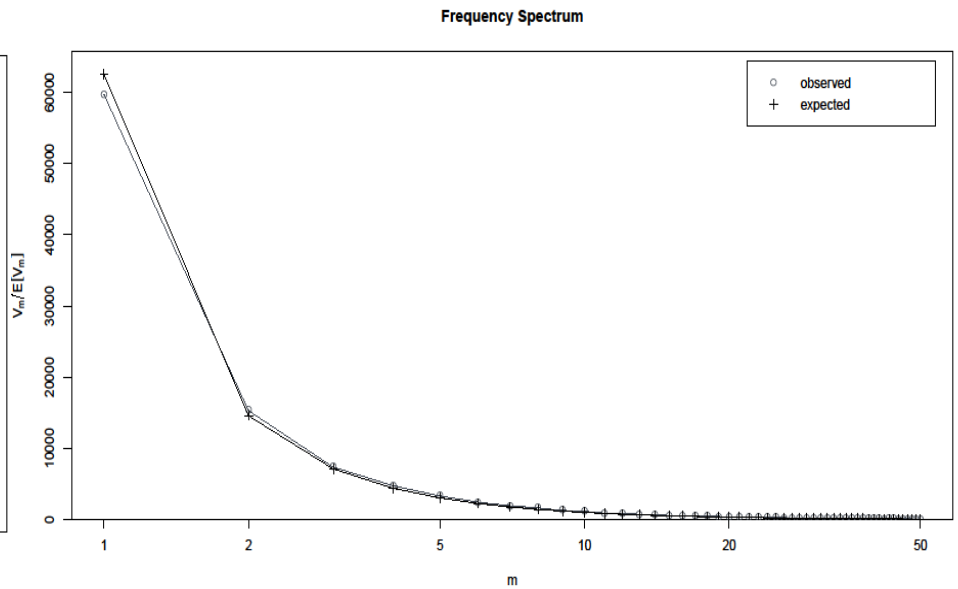


General language vs titles corpus (1)

Zipf-Mandelbrot distribution

$$f(i, N) = \frac{K}{(i+b)^a},$$



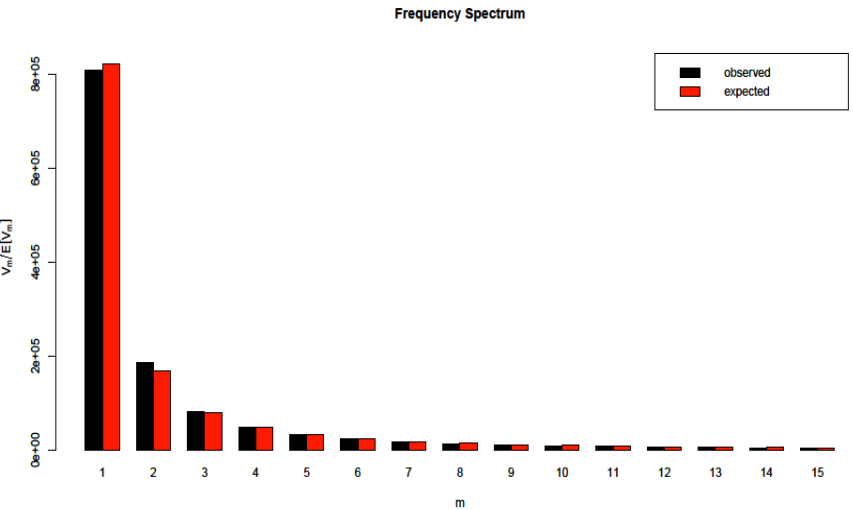


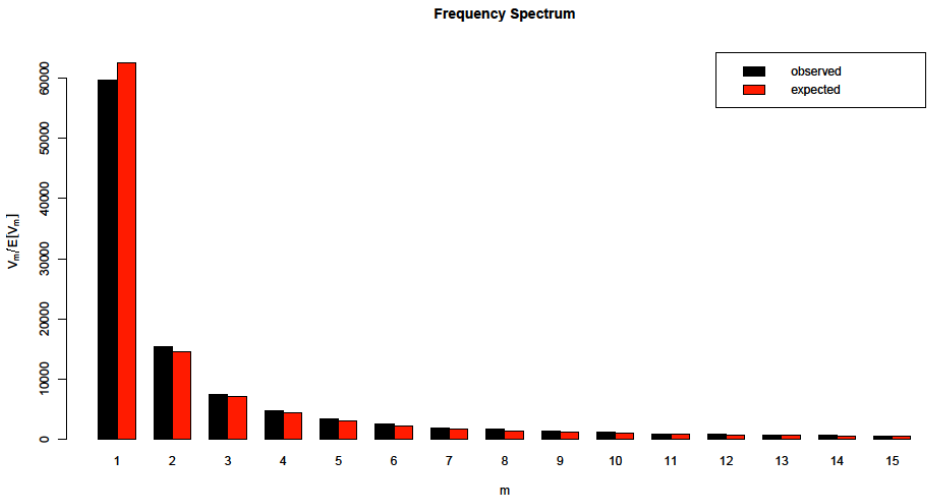
	Distribution	Par. *a*	Par. *b*	X2	df	p
General language	ZM	0,58774	0,00288	3514,09	14	0
Titles	ZM	0,53296	0,00133	1402,91	14	0

General language vs titles corpus (1a)

Zipf-Mandelbrot distribution

$$f(i, N) = \frac{K}{(i+b)^{a}},$$



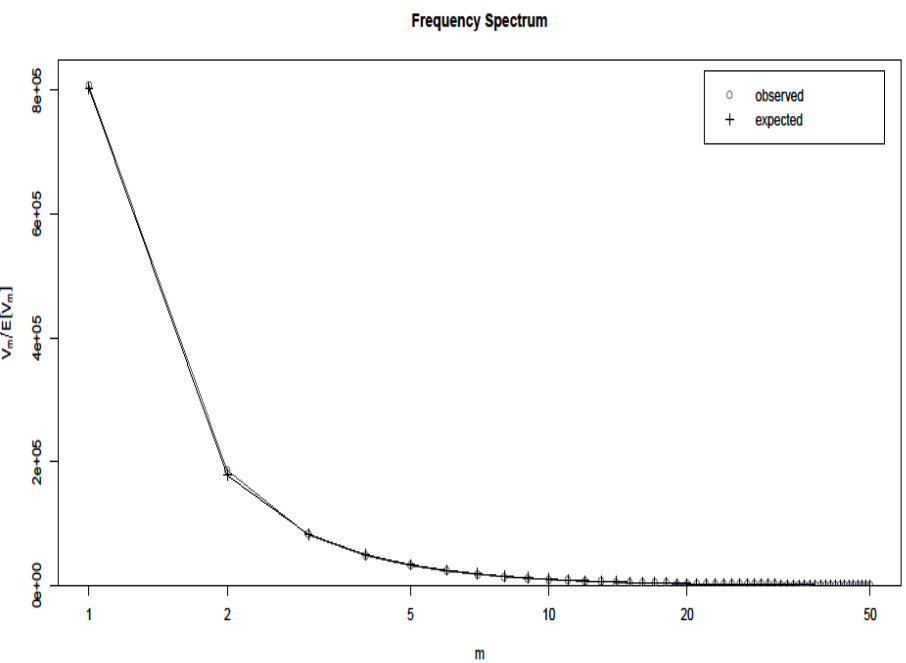


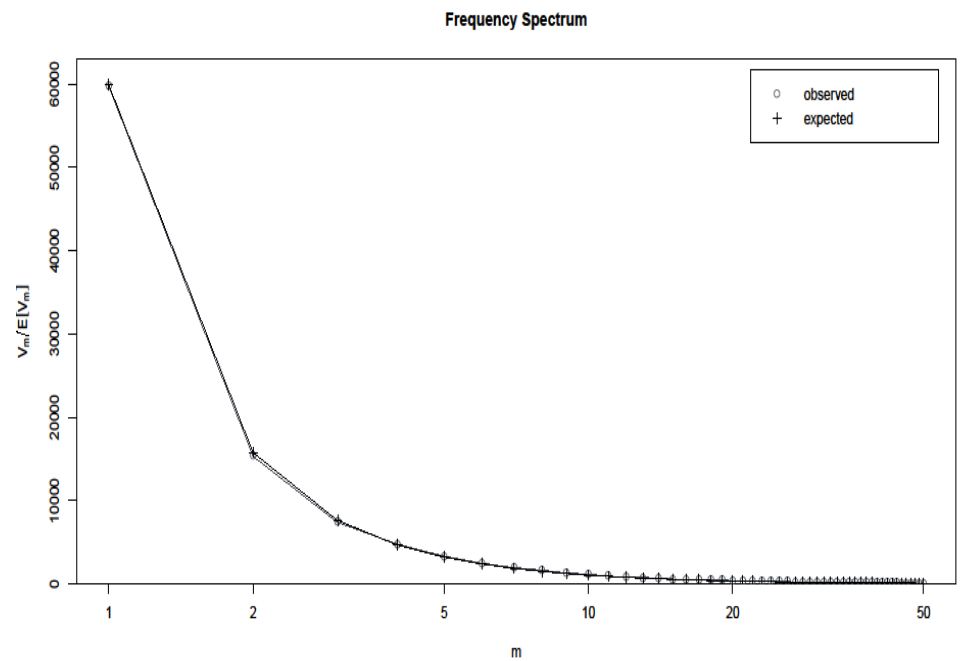
	Distribution	Par. *a*	Par. *b*	X2	df	p
General language	ZM	0,58774	0,00288	3514,09	14	0
Titles	ZM	0,53296	0,00133	1402,91	14	0

General language vs titles corpus (2)

Finite Zipf-Mandelbrot distribution

$$h(i, N) = \frac{K}{i^{a}},$$



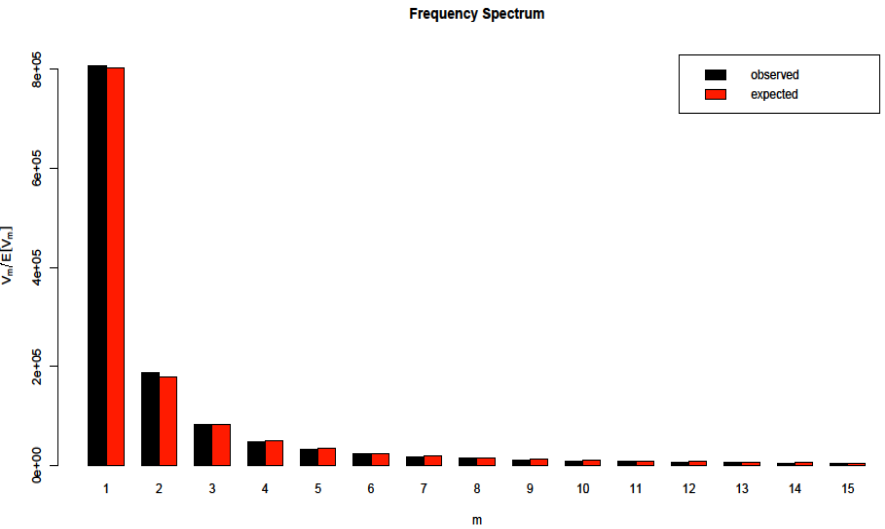


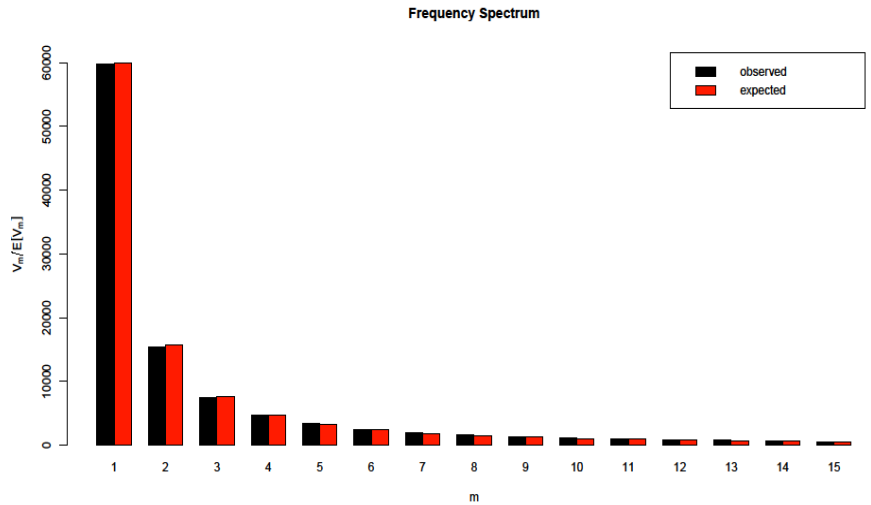
	Distribution	Par. *a*	Par. *b*	Par. *c*	X2	df	p
General language	fZM	0,59888	9,409E-12	3,8852	4440,24	13	0
Titles	fZM	0,55029	2,87E-09	8,5609	303,24	13	0

General language vs titles corpus (2a)

Finite Zipf-Mandelbrot distribution

$$h(i,N) = \frac{K}{i^a}, \quad \mathrm{i} = b, \dots c$$



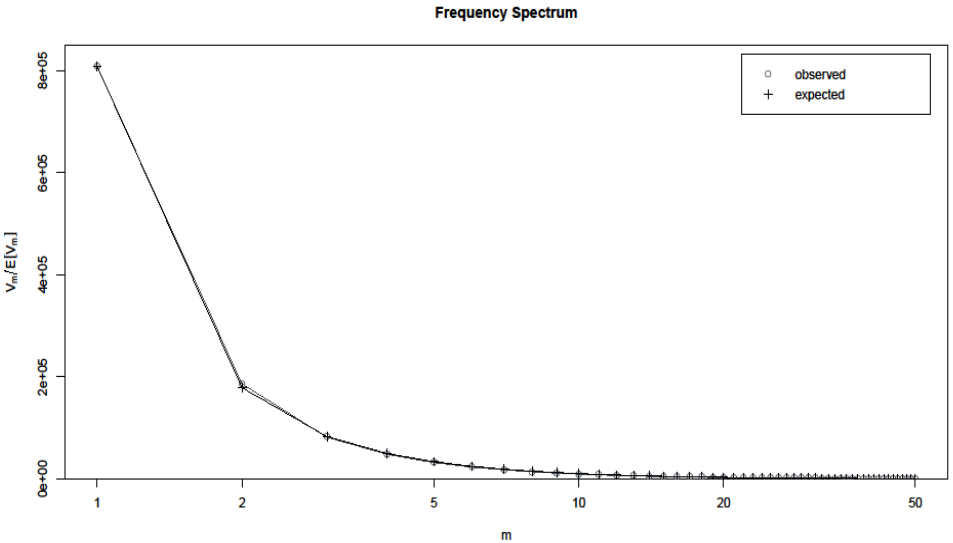


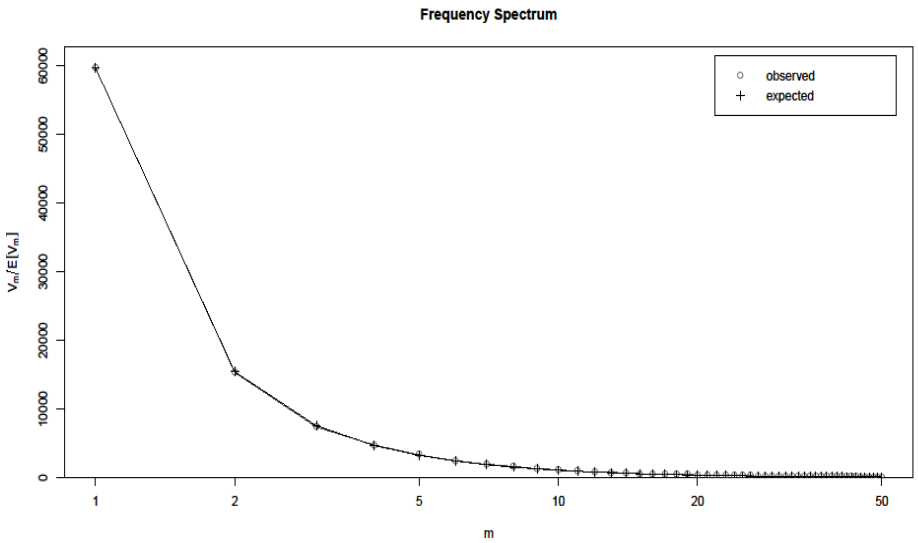
	Distribution	Par. *a*	Par. *b*	Par. *c*	X2	df	p
General language	fZM	0,5989	9,409E-12	3,885	4440,24	13	0
Titles	fZM	0,5503	2,87E-09	8,561	303,24	13	0

General language vs titles corpus (2)

Generalized inversed Gauss-Poisson distribution

$$g(x) = Mx^{a-1}\exp\left(-\frac{x}{c} - \frac{b^2 c}{4x}\right)$$



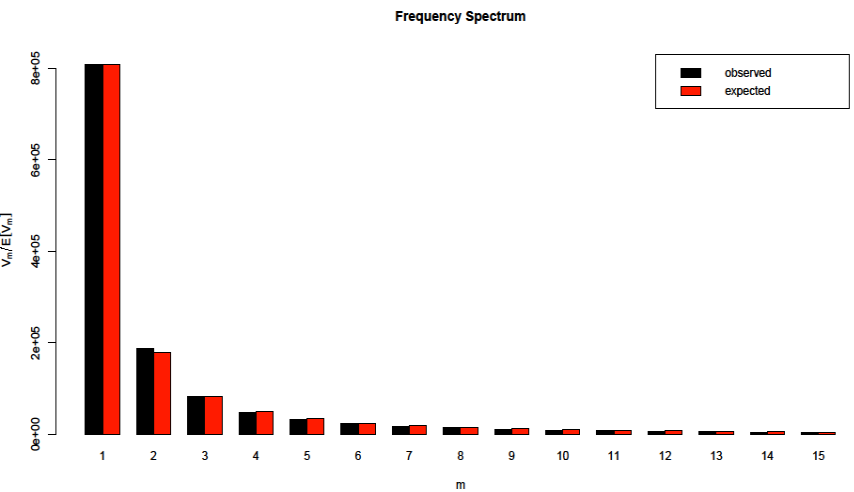


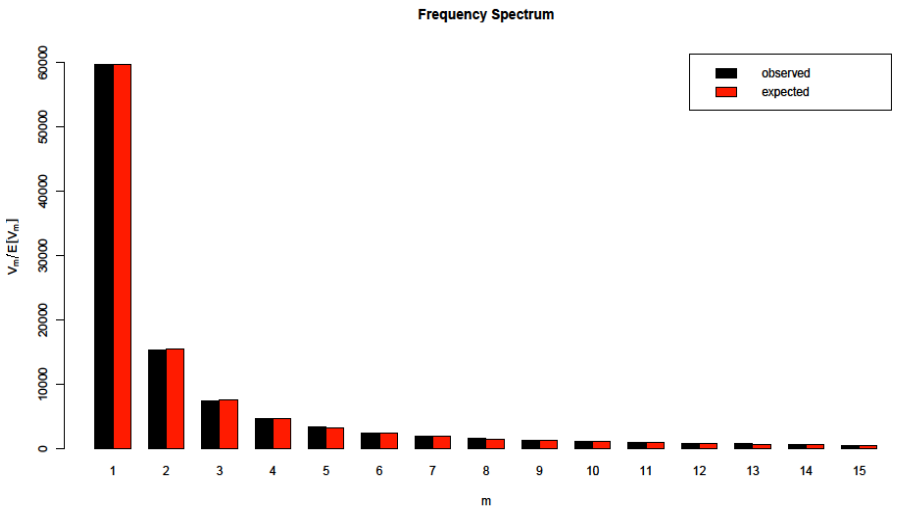
	Distribution	Par. *a*	Par. *b*	Par. *c*	X2	df	p
General language	GIG-P	-0,5995	5,581E-05	0,0047	5007,67	13	0
Titles	GIG-P	-0,5277	0,00113	0,0014	108,73	13	0

General language vs titles corpus (2)

Generalized inversed Gauss-Poisson distribution

$$g(x) = Mx^{a-1}\exp\left(-\frac{x}{c} - \frac{b^2 c}{4x}\right)$$





	Distribution	Par. *a*	Par. *b*	Par. *c*	X2	df	p
General language	GIG-P	-0,5995	5,581E-05	0,0047	5007,67	13	0
Titles	GIG-P	-0,5277	0,00113	0,0014	108,73	13	0

AUTOMATIC CLASSIFICATION OF TITLES

Why are bibliographies so interesting?

1. They include titles of different length to classify
2. They include metadata which allow verifying accuracy of classification

Uniwersytet Wrocławski

A complete record in a human readable form

BINMORE, Ken (1940-)
Teoria gier / Ken Binmore ; translation Iwona Konarzewska. – Łódź : Wydawnictwo Uniwersytetu Łódzkiego, 2017. – 206, [1] page : graphics, photos, charts ; 21 cm. – (Krótkie Wprowadzenie; 8)

Title of the original: *Game theory : a very short introduction*.

References on pages 195-200. Index.
Available also as e-book. – Publication financed by Wydawnictwo Uniwersytetu Łódzkiego
ISBN 978-83-8088-594-3
ISBN 978-83-8088-595-0 (e-ISBN)
Type: Publikacje popularnonaukowe
Genre: Opracowanie
Creation time: 2007
Subject: Teoria gier
Domain: Filozofia i etyka

(620 characters)

519.83

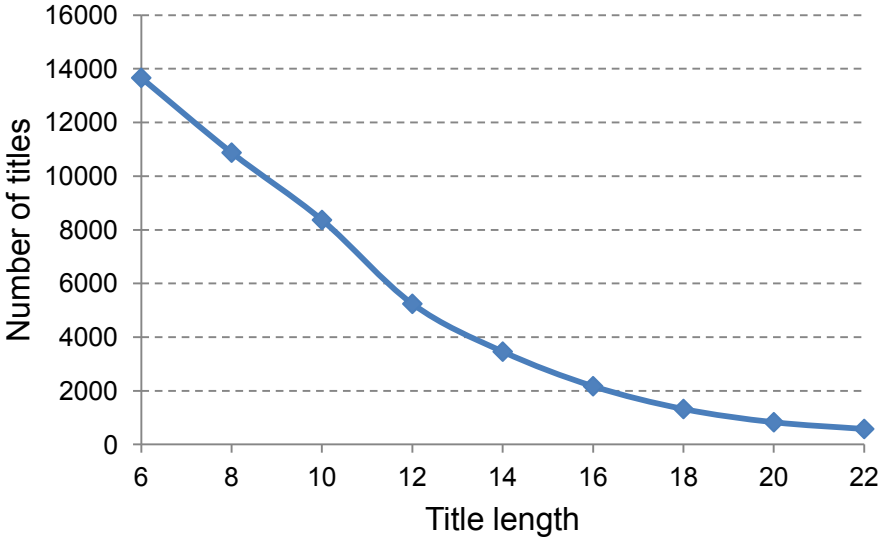
Experiment: classification of titles

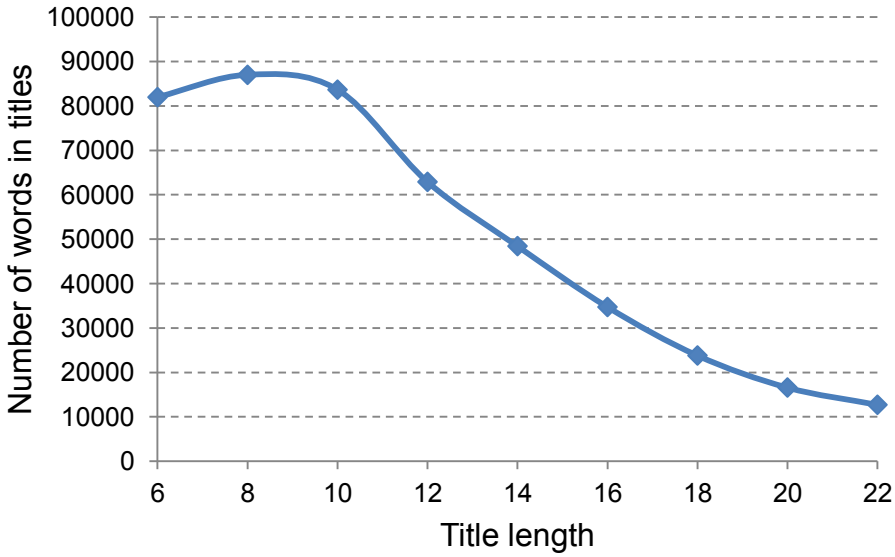
1. Method: fastText algorithm
2. Experiment:
   – variable both title length and the size of a training set
   – variable title length and invariable size of a training set

What is FastText ?

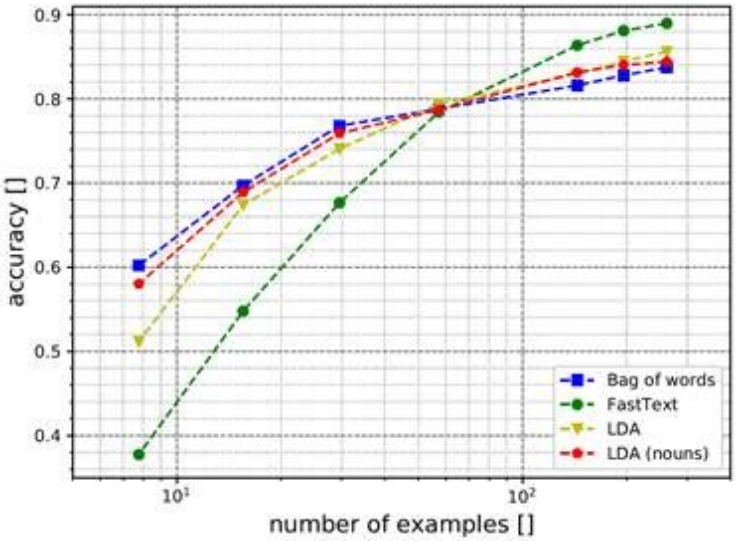
– developed by Facebook's AI Research (FAIR) lab
– recent deep learning method for text classification
– based on word embedding: representation of words (terms) by a multidimensional vector (like Word2Vec)
– representation of documents as an average of word embeddings and uses a linear softmax classifier
– main idea: word representation and classifier learned in parallel
– no NLP knowledge (e.g. jech-ał, jech-ali – different terms)
– available: https://fasttext.cc/docs/en/support.html

Variable both title length and the size of a training set



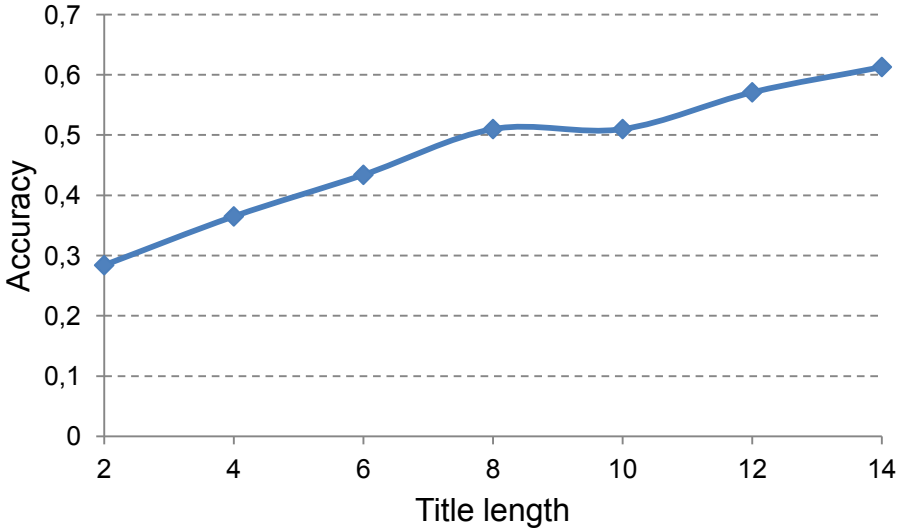


Accuracy of classification tested on Wikipedia

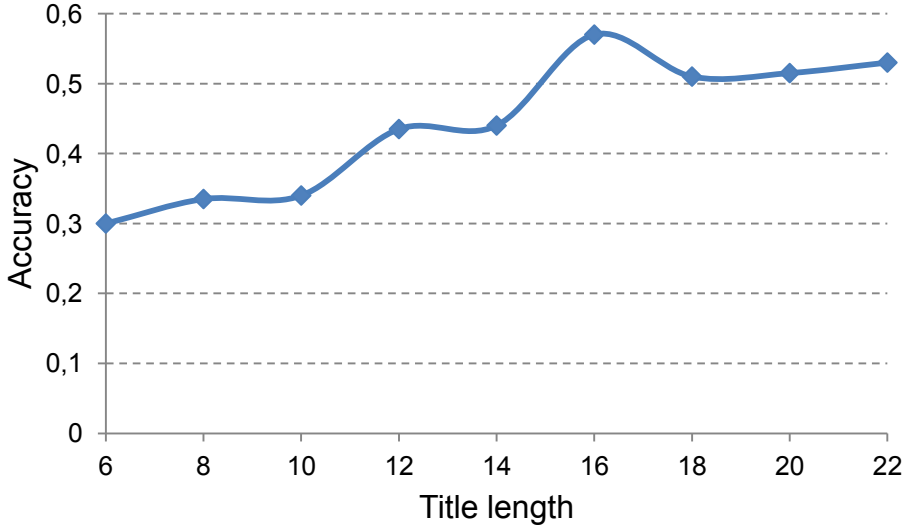


Variable title length, constant size of a training set

training set = 3469 titles,
classified set = 865 titles



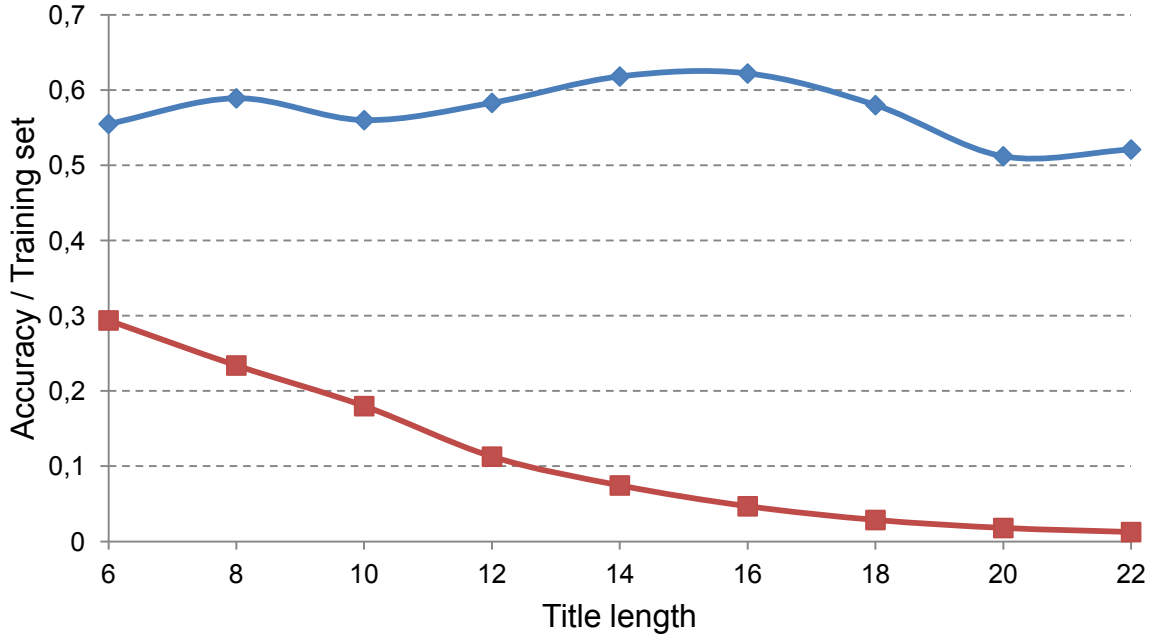
training set = 600 titles,
classified set = 200 titles



Variable title length, variable size of a training set

training set: variable,

classified set: variable

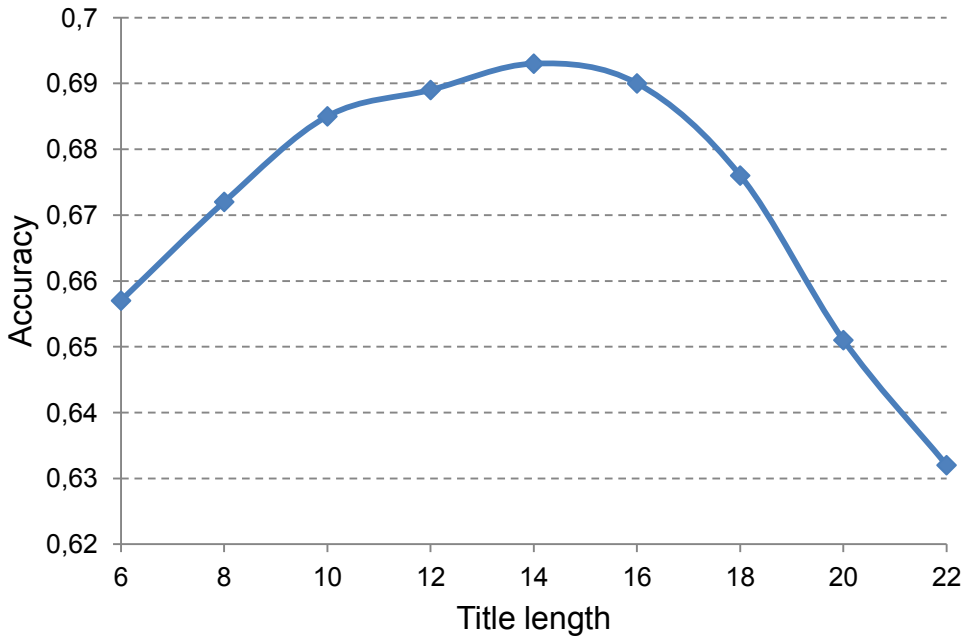


Length	Recognition rate	Number of available titles
6	0,555	3415
8	0,589	2719
10	0,56	2091
12	0,583	1310
14	0,618	865
16	0,622	543
18	0,58	331
20	0,512	207
22	0,521	144

Variable title length, variable size of a training set

training set: variable

classified set: variable



Length	Recognition rate	Number of available titles
6	0,657	20412
8	0,672	14050
10	0,685	9515
12	0,689	6165
14	0,693	4004
16	0,69	2602
18	0,676	1726
20	0,651	1168
22	0,632	806

Titles: possible research

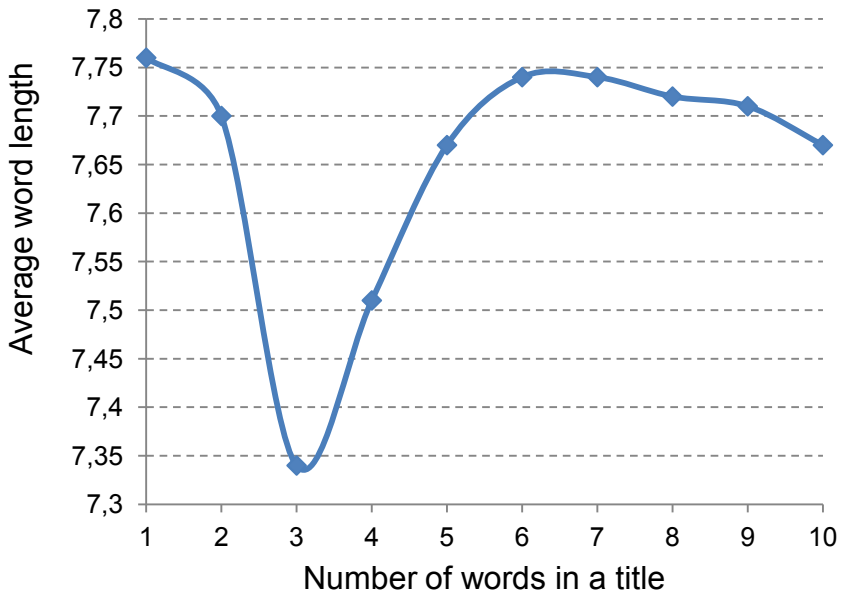
Length of title in words	Number of titles	Av. length of title in chars	Av. length of a word in title (in chars)
1	28888	7,76	7,76
2	69068	15,40	7,70
3	73137	22,02	7,34
4	62765	30,05	7,51
5	54517	38,37	7,67
6	48964	46,46	7,74
7	42270	54,21	7,74
8	35809	61,70	7,72
9	29199	69,40	7,71
10	23391	76,68	7,67

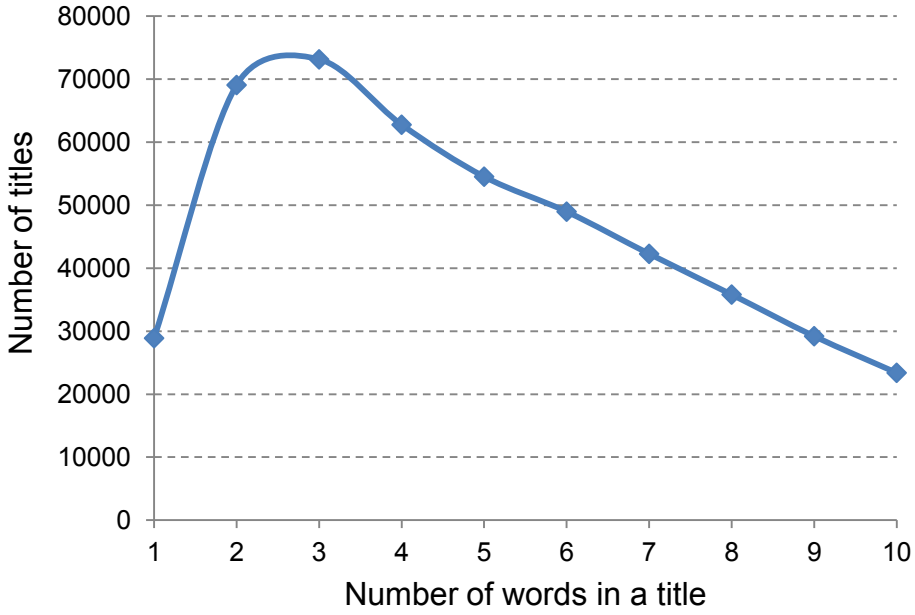
Titles: possible research

Possible relationships:

number of words in a title *vs* average word length

number of words in a title *vs* number of titles





CONCLUSIONS

Conclusions

1. Great bibliographies are specific sets of textual data and can be processed with quantitative tools like any other corpora.
2. MARC format is not appropriate for straightforward automatic processing (redundancy, opaque structure of fields).
3. Bibliography corpus has specific characteristics when compared with natural language corpora (more nominal and less verbal units).
4. Word spectra generated from a bibliography corpus and from a general language corpus are similar in shape but statistically different.
5. Titles are a good material for testing classification methods (evaluation using metadata).
6. Satisfactory results (accuracy 70%) can be obtained with titles of 12-14 words of length (should this be valid for other genres?).

Uniwersytet Wrocławski

Thank you