Wortschatz / Leipzig Corpora Collection

Working with under-resourced languages

Collecting smaller languages

Problems

- Often, only a small part of the Web in country is in the language under consideration.
- More resources can be found in .com domains, but they are difficult to identify because of the sice of the .com TLD.

Criteria

- More than 1.000.000 speakers
- Less than 1.000.000 sentences found so far

Search Engine based Bootstrapping

- Wordlist based approach using wordlists from
  - Universal Declaration of Human Rights (UDHR)
  - Watchtower documents
- Tuples of words for search engine queries
- Usage of Webservice-based APIs – if available - (Google, Microsoft Bing) to collect URLs

→ Can be iterated

Language	Code	Country	Speakers	Sentences
Somali	som	Somalia	13871700	121545
Swahili	swa	Tanzania	15458390	111927
Zulu	zul	South Africa	10349100	75869
Shona	sna	Zimbabwe	10759200	52060
Xhosa	xho	South Africa	7817300	41983
Amharic	amh	Ethiopia	17528500	24856
Northern Sotho	nso	South Africa	4101000	19000
Southern Sotho	sot	South Africa, Lesotho	6024000	18000
Yoruba	yor	Nigeria	19380800	12886
Ganda	lug	Uganda	4130000	11058
Southern Ndebele	nbl	South Africa	1090000	8000
Wolof	wol	Senegal	3976500	5715
Hausa	hau	Nigeria	24988000	5506
Tswana	tsn	South Africa, Botswana	4521700	5229
Tsonga	tso	South Africa	3669000	3264
Lingala	lin	Dem. Rep. of Congo	2141300	2783
Ndonga	ndo	Namibia	1070000	2659

Language	Code	Country	Speakers	Sentences
Venda	ven	South Africa	1064000	1893
Igbo	ibo	Nigeria	18000000	1716
Bamanankan	bam	Mali	2772340	1643
Oromo	orm	Ethiopia	17467900	1442
Éwé	ewe	Ghana	3112000	1000
Swati	ssw	South Africa, Swaziland	2034200	901
Rundi	run	Burundi	4851000	826
Tigrigna	tir	Ethiopia, Eritrea	5791710	805
Fulah	ful	Cameroon	22246400	519
Akan	aka	Ghana	8314600	505
Koongo	kng	Dem. Rep. of Congo	5000000	490
Rwanda	kin	Rwanda	7504900	469
Nyanja	nya	Malawi	8659700	446
Tumbuka	tum	Malawi	1142000	300
Acholi	ach	Uganda	1215000	192
Gikuyu	kik	Kenya	6623000	138
Pulaar	fuc	Senegal	3691000	18

Language	Code	Country	Speakers
Umbundu	umb	Angola	4002880
Kimbundu	kmb	Angola	4000000
Fon	fon	Benin	1435500
Mòoré	mos	Burkina Faso	5061700
Jula	dyu	Burkina Faso	1229000
Dan	dnj	Côte d'Ivoire	1610800
Baoulé	bci	Côte d'Ivoire	2130000
Zande	zne	Dem. Rep. of Congo	1142000
Yombe	yom	Dem. Rep. of Congo	1056400
Songe	sop	Dem. Rep. of Congo	1000000
Ngbaka	nga	Dem. Rep. of Congo	1016650
Luba-Katanga	lub	Dem. Rep. of Congo	1510000
Luba-Kasai	lua	Dem. Rep. of Congo	6300000
Kongo	kon	Dem. Rep. of Congo	5644100
Kituba	ktu	Dem. Rep. of Congo	4200000
Chokwe	cjk	Dem. Rep. of Congo	1009780
Alur	alz	Dem. Rep. of Congo	1367000
Tigré	tig	Eritrea	1050000
Wolaytta	wal	Ethiopia	1710000
Sidamo	sid	Ethiopia	2980000
Hadiyya	hdy	Ethiopia	1250000
Gamo	gmv	Ethiopia	1110000
Afar	aar	Ethiopia	1078200
Abron	abr	Ghana	1182000
Susu	sus	Guinea	1060280
Pular	fuf	Guinea	2929200
Kpelle	kpe	Guinea	1220000
Fang	fan	Guinea	1027900
Eastern Maninkakan	emk	Guinea	3531800
Oluluyia	luy	Kenya	5199727
Maasai	mas	Kenya	1045000
Lubukusu	bxk	Kenya	1433000
Kipsigis	sgc	Kenya	1916000
Kimîîru	mer	Kenya	1658000
Kamba	kam	Kenya	3893000
Kalenjin	kln	Kenya	5123400
Ekegusii	guz	Kenya	2120300
Dholuo	luo	Kenya	4410000
Yao	yao	Malawi	1916000
Soninke	snk	Mali	1250000
Maasina Fulfulde	ffm	Mali	1008500
Tswa	tsc	Mozambique	1180000

Language	Code	Country	Speakers
Sena	seh	Mozambique	1340000
Makhuwa-Meetto	mgh	Mozambique	1348000
Makhuwa	vmw	Mozambique	3090000
Lomwe	ngl	Mozambique	1500000
Zarma	dje	Niger	2438400
Tamashek	tmh	Niger	1248200
Tiv	tiv	Nigeria	2210000
Nigerian Pidgin	pcm	Nigeria	30000000
Nigerian Fulfulde	fuv	Nigeria	11500000
Kanuri	kau	Nigeria	3760500
Izon	ijc	Nigeria	1000000
Ibibio	ibb	Nigeria	1750000
Edo	bin	Nigeria	1000000
Ebira	igb	Nigeria	1000000
Central Kanuri	knc	Nigeria	3240500
Berom	bom	Nigeria	1000000
Anaang	anw	Nigeria	1400000
Serer-Sine	srr	Senegal	1161900
Mandinka	mnk	Senegal	1346000
Mandingo	man	Senegal	6496300
Themne	tem	Sierra Leone	1230000
Mende	men	Sierra Leone	1499700
Dinka	din	Sudan	1365900
Bedawiyet	bej	Sudan	1186000
Sukuma	suk	Tanzania	5430000
Nyakyusa-Ngonde	nyy	Tanzania	1105000
Makonde	kde	Tanzania	1340000
Haya	hay	Tanzania	1300000
Gogo	gog	Tanzania	1440000
Teso	teo	Uganda	1849000
Soga	xog	Uganda	2060000
Nyankore	nyn	Uganda	2330000
Masaaba	myx	Uganda	1120000
Lugbara	lgg	Uganda	1637000
Lango	laj	Uganda	1490000
Chiga	cgg	Uganda	1580000
Bemba	bem	Zambia	3602000
Tonga	toi	Zambia, Zimbabwe	1127000
Ndebele	nde	Zimbabwe	1572800
Ndau	ndc	Zimbabwe	2380000
Manyika	mxc	Zimbabwe	1025000

Collecting smaller languages

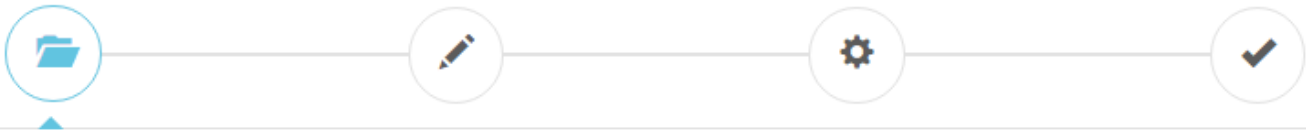
- Portal to allow external support
- User friendly interface, but work in progress
- Try it: http://curl.corpora.uni-leipzig.de

Collecting Web Pages for

Under-Resourced Languages

On this website you can contribute to corpus collection for under-resourced languages by simply entering a URL. The URLs or domains you provide will be crawled and reviewed for text data in the respective language. After processing you will be presented with statistics for the URLs you provided. The corpora created will be free for download.

Thank you for your Support!



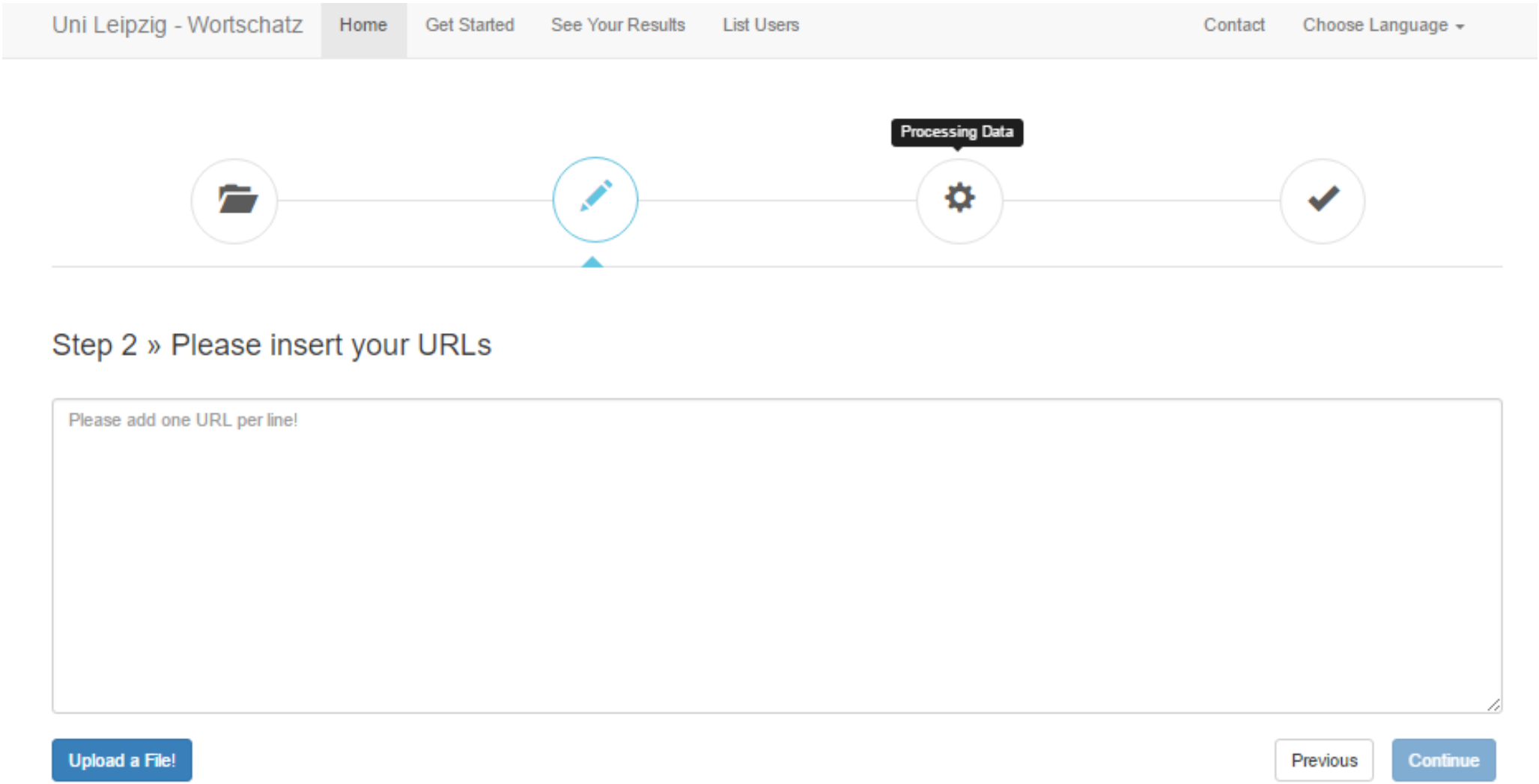
Step 1 » Please select a language.

Yombe [yom] - Dem. Rep. of Congo
Yoruba [yor] - Nigeria
Zande [zne] - Dem. Rep. of Congo
Zarma [dje] - Niger
Zaza [zza] - Turkey
Zhuang [zha] - China
Zulu [zul] - South Africa

Continue

Automatische Sprachverarbeitung

Collecting smaller languages



Job details for amh-2016-09-20-07-51-19

Job Name	amh-2016-09-20-07-51-19
Submitted on	2016-09-20 07:51:20
Processed	2016-09-20 14:04:04 / 2016-09-20 20:20:11
Language	Amharic [amh]
# URLs	8
Status	All the URLs have been processed. No details available.

URL List

Show 10 entries Search:

Status	Language	URL
finished	amh	http://www.ebc.et/web/news
finished	amh	http://ethiopiazare.com/
finished	amh	https://www.lds.org/general-conference/201 ...
finished	amh	http://www.formerethiopianairforce.com/
finished	amh	http://etntc.org/
finished	amh	http://amharic.voanews.com/a/ethiopians-de ...
finished	amh	http://www.ginbot7.org/
finished	amh	http://www.ethioshengo.org/

Corpus Information for Amharic [amh] Ethiopia

Language	Amharic
ISO Code	amh
Country	Ethiopia
Corpus Name	amh_community_2016
Tokens	637583
Types	133454
Sentences	43083
Sources (URLs)	7662
Build date	2016-09-20
URLs	List of URLs download List of Domains download
Download	amh_community_2016 2016-09-20
Contact	No contact person for this language. Use this Contact to add contact details.

CORPORA COLLECTION

UNIVERSITY LEIPZIG

Filter URLs ...

http://am.wikipedia.org/	33672
http://www.ginbot7.org/	2212
http://etntc.org/	1738
http://www.ethioshengo.org/	1392
http://www.ebc.et/	1063
https://www.lds.org/	972
http://www.formerethiopianairforce.com/	443
http://amharic.voanews.com/	104

Email Notifications

```
Statistics for user hau-2016-05-23-07-43-24
-------------------------------------------

We used Heritrix to crawl your URLs.
You submitted 1 seed URLs. 1 seeds could be crawled but 0 couldn't for various reasons.
Heritrix downloaded 6814 URLs from which 6589 URLs were valid.
In total we downloaded 129 MiB in 6h4s10ms. We had a download speed of 0.3 URI/sec.

We then extracted the text content from the webpages. Following this, we separated 3144 documents
by language using linguistic methods like trigramms and stop words, and the encoding.
For your language 'hau' we found 1015 documents, 32.28 % of the total. Furthermore, 6 other
languages were found but discarded.

After the preprocessing, the documents were split in sentences and the sentences classified by
language.
For the language 'hau' we found 8800 sentences from a total of 9057.

We successfully re-created the corpus 'hau_community_2016' with your additional language data.
The URLs you provided added 805 source URLs, 8431 sentences and 4845 words. The old corpus had
1687 sources, 33158 sentences and 35387 words.
```

Live Demo

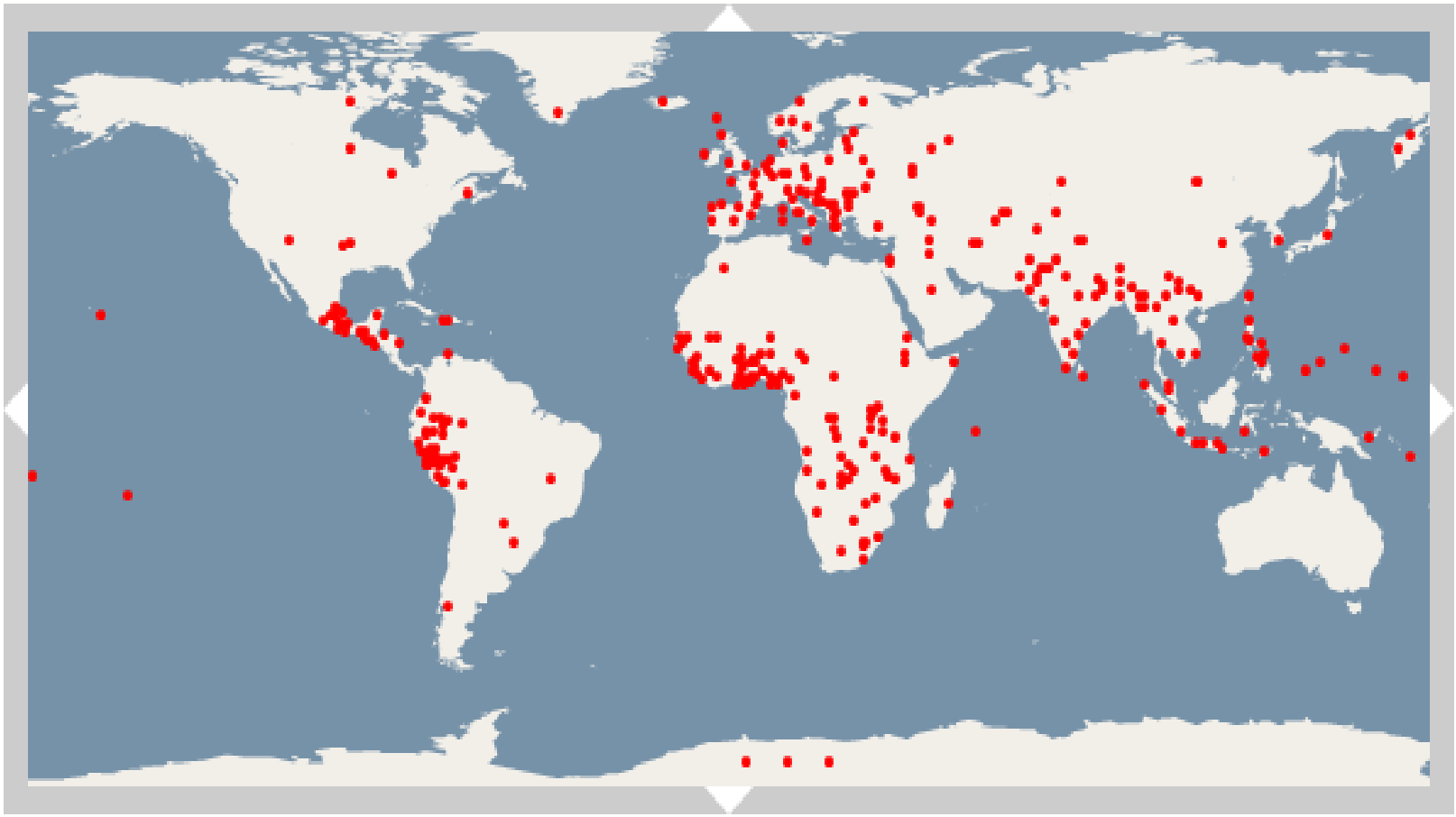
Language Identification

In two steps:

1. step: Identification of the document language using the following algorithm:
   Use lists of the 50 most frequent words for many languages. The language having by far the most hits wins. Documents are stored according to their language.
2. step: The same for each sentence using 5.000 words per sentence. Sentences not in the expected language are rejected.

How to find stopwords 1: UDHR

- The **Universal Declaration of Human Rights** is available in more than 360 languages in UNICODE
- Size of the German version: approx. 1700 words



How to find stopwords 2: Bibles

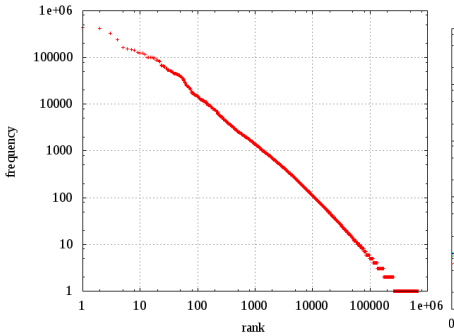
Bibles (or parts of the Bible)

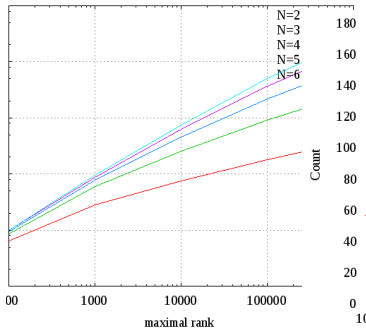
- in nearly 1000 languages
- With typically around 7000 verses
- Sources: bible.is and other websites

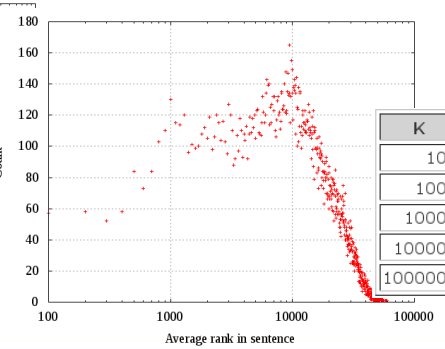
Statistical Evaluation

- Enrichment of corpora with statistical annotations
  - Word frequencies
  - Co-occurrence frequencies and significance (based on left or right neighbours or whole sentences)
- Creation of further corpora statistics
  - 230 statistical properties based on letter, word, sentence, sources level etc.

WORTSCHATZ
Corpora and Language Statistics





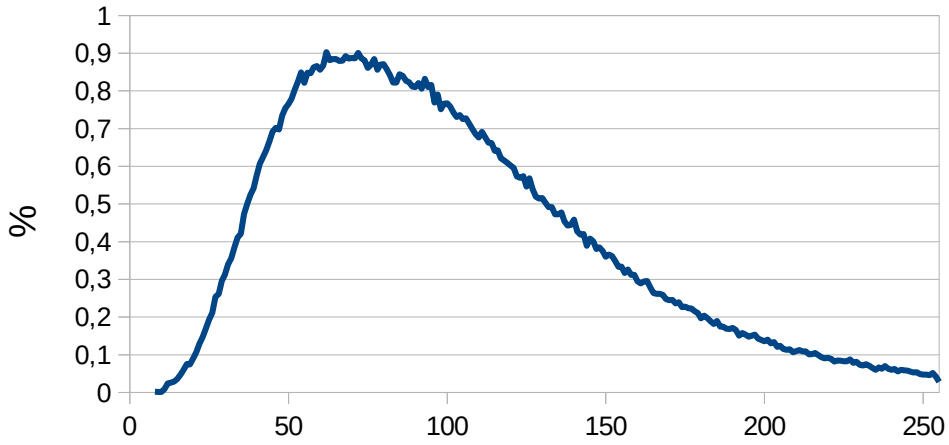


K	# of bigrams	# of trigrams	# of 4-grams	# of 5-grams	# of 6-grams
100	64	85	92	98	99
1000	279	592	767	877	944
10000	746	2529	4575	6199	7430
100000	1811	9025	22062	37161	50546
1000000	3938	27928	85267	168673	253078

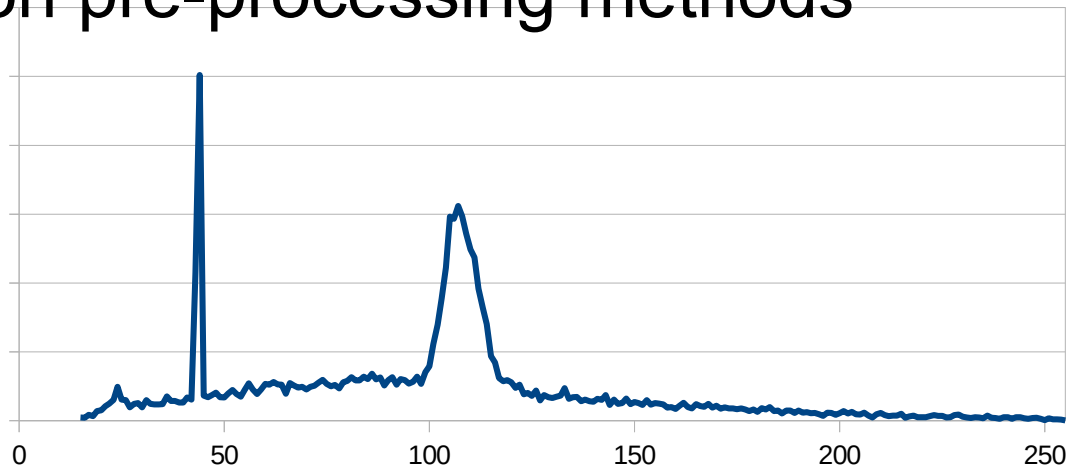
Applications of Corpora Statistics

Statistics-based Quality Assurance

- Automatic analysis of statistical properties for quality enhancements
  - Usage of typical distributions of linguistic properties
    - Length distributions (word, sentence, paragraph)
    - Character and n-gram distributions etc.
  - Basis for further cleaning procedures
  - Basis for adaptions on pre-processing methods



Hindi newspaper 2010



Sundanese Wikipedia 2007

Corpus-Based Quantitative Typology

- Correlations between parameters

corpus-based properties (measured)	typological parameters

- Determine correlations:
  - Between different measurements
  - Betwen measured properties and typological parameters
  - Typological parameters:
    - Refer e.g. to morphology or syntax
    - Source: World Atlas of Language Structures (WALS)
  - Usage of quantitativee Methods:
    - Correlation Analysis
    - Tests of Significance

Automatische Sprachverarbeitung

Comparison of Language Parameters

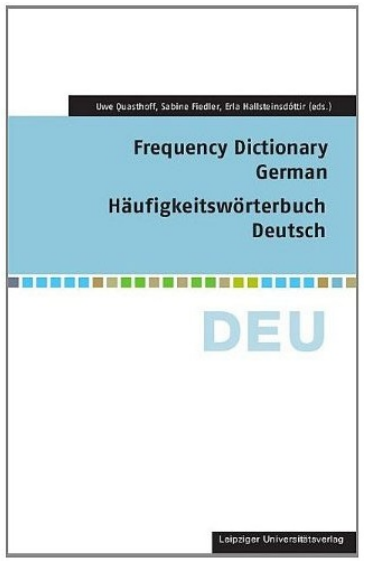
Lang	Alphabet	VC Ratio	Cap in text	Cap in top100K	Zipf slope	Cov: 10	Cov: 100	Cov: 1000	Len in text	Len in top100K
CES	**42**	**42:58**	**13.9%**	**24,3%**	**-0.96**	**13.5%**	**28.6%**	**47.0%**	**5.37**	**9.08**
DEU	31	38:62	36.0 %	68,3%	-1.07	15.5 %	39.1 %	59.2 %	5.86	9.73
ENG	26	40:60	17,6%	48,6 %	-1.10	21.4%	41.5%	63.2%	4.96	7.80
FRA	ca. 41	44:56	12.9%	40,6 %	-1.12	22.1%	46.8%	66.1%	4.82	7.80
HUN	ca. 38	42:58	12.9%	16,9 %	-0.95	19,7%	31,0%	47,9%	6,27	8.96
IND	26	42:58	19.5%	49.9 %	-1.08	13.5%	31.3%	58.8%	6.08	9.28
ISL	ca. 34	39:61	13,5%	30,0 %	-1.06	21,5%	40.3%	60.5%	5.08	8.47
UKR	34	46:54	16.3%	38.3%	-1.14	23.1%	46.1%	69.1%	4.36	8.48

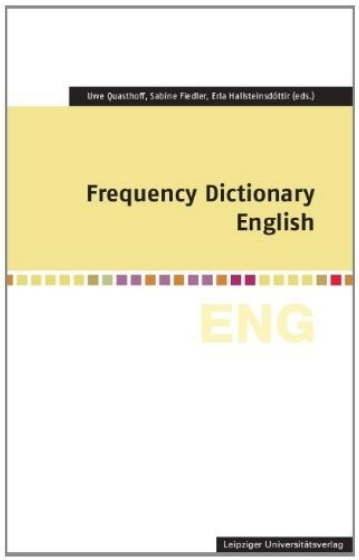
Results

- Correlations between measurements:
  - Negative Correlation between length of words in characters and sentence length in words (Cor=-0,55, p<0,001%)
  - The more syllables per word the fewer words per sentence (Cor=-0,49, p<0,001%)
  - The more syllables per word, the more syllables per sentence (Cor=0,47, p<0,001%)
- Correlations between measurements and typological parameters:
  - Morphological type: concatenative vs. isolating
    - Isolating languages have shorter words: p<1%, average of 8.43 and 6.95
  - Word order: SOV vs. SVO
    - SOV languages typically use more syllables per word: p<0,005%, average of 2.21 and 1.89

Frequency Dictionaries

- Book series, published by Leipzig University Press since 2011
- Most frequent words of a language with frequency information
  - Printed: Top 10,000 words
  - Electronically (CD): Up to top 1,000,000 words





- Searching for language experts for further languages
- quasthoff@informatik.uni-leipzig.de

Series: Frequency Dictionaries

Additional statistical information on word list about:

- alphabet, distribution of vowels and consonants
- word length (average and for different frequency ranges)
- Number of syllables per word
- Longest words for different frequency ranges
- Corpus timeline

Previous Volumes

In preparation:

- **NLD** – Dutch
- **VIE** – Vietnamese
- **CES** – Czech
- **POL** – Polish
- **RUS** – Russian
- **GLG** – Galician
- **DAN** – Danish
- **SWE** – Swedish
- **AFR** – Afrikaans
- **ZUL** – Zulu