Oftecvortaroj por Esperanto
Frequency dictionaries for Esperanto

Sabine Fiedler
sfiedler@uni-leipzig.de

Tria Interlingvistika Sympozio

Poznań 2014

Strukturo / structure

(1)Enkonduko / Introduction (What is a frequency dictionary?)

(2)Oftecaj analizoj jam ekzistantaj pri Esperanto / previous frequency lists and dictionaries for Esperanto

(3) La Leipzig-a projekto / The project at the University of Leipzig

Kio estas oftecvortaro? / What is a frequency dictionary?

- *Oftecvortaroj prezentas la leksikon de lingvo, de aŭtoro, ĝenro k.s. laŭ ĝia ofteco en teksto aŭ korpuso / F.d. present words of a language, an author, genre etc. according to their frequency in a corpus*
- *Ili tiel donas informojn, kiuj vortoj estu unue lernataj / F.d. provide information on the words that should be learnt first*

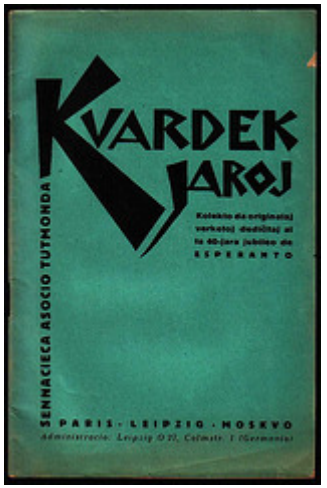
Superrigardo pri jam ekzistantaj review of oftecvortaroj / previous frequency lists for Esperanto

- Ĉefaj fontoj / main sources
- Haupenthal, R. (1991) “Lexikographie von Hilfssprachen und anderen Kommunikationssystem”. In: Hausmann, Franz Joseph / Reichmann, Oskar / Wiegand, Herbert Ernst /Zgusta, Ladislav (Hrsg.) *Wörterbücher Dictionaries Dictionnaires. Ein internationales Handbuch zur Lexikographie.* Band 5.3 Berlin et al.: de Gruyter, 3129-3137.
- Pabst, Bernhard: *La du plej oftaj vortoj: ‚la’ kaj ‚kaj’ – iom pri oftecvortaroj*.

  http://www.ipernity.com/blog/bernardo/204877 (2010) [10 Jan 2013].

S. Rublov (1927)

-100 000 vortoj (6 tekstoj: “Sennaciulo” N-ro 1, 30, 45; “ABC de Sennaciismo”, “Patroj kaj filoj”, “En malliberejo”)



Danneils (1928) : en 29 naciaj Esperanto-lernolibroj kaj komparas tiujn kun ekvivalentoj en oftecvortaroj en la germana, angla kaj la franca / compares the most frequent words in national languages

- Leo Blaas (1949-1955): 50 000 vortoj (en 200 hazarde elektitaj paĝoj el 110 verkoj de Esperanto (noveloj, ĵurnaloj, enciklopedioj)

- V. Sadler (1958): 3000 vortoj el prozo (datumoj pri ofteco, afiksoj vortklasoj ktp.) (konfirmita de Harry [1967])

- P. Bovet kaj H. Ith (1956): 577 internacie oftaj vortoj kompare al Esperanto

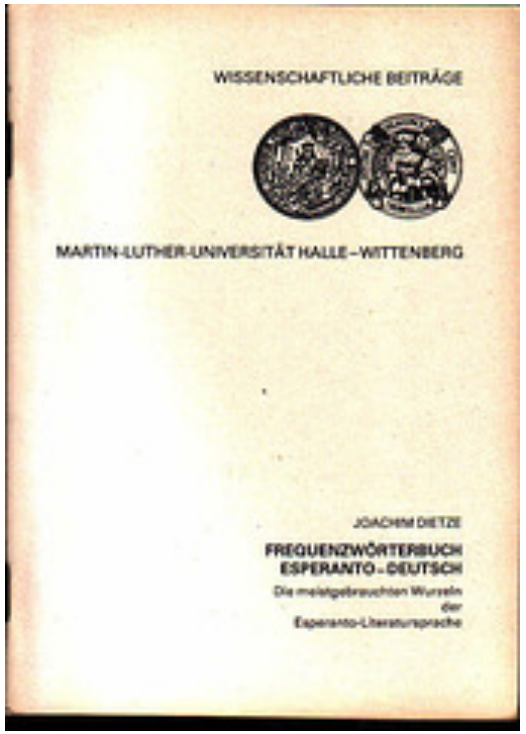
- Franca projekto *Le français fundamental* uzita por celoj de Esperanto (tradukita leksiko el parola kaj skriba komunikado)

→ Akademio de Esperanto *Baza Radikaro Oficiala* (Albaut 1975): Difinado de baza vortaro en 9 grupoj

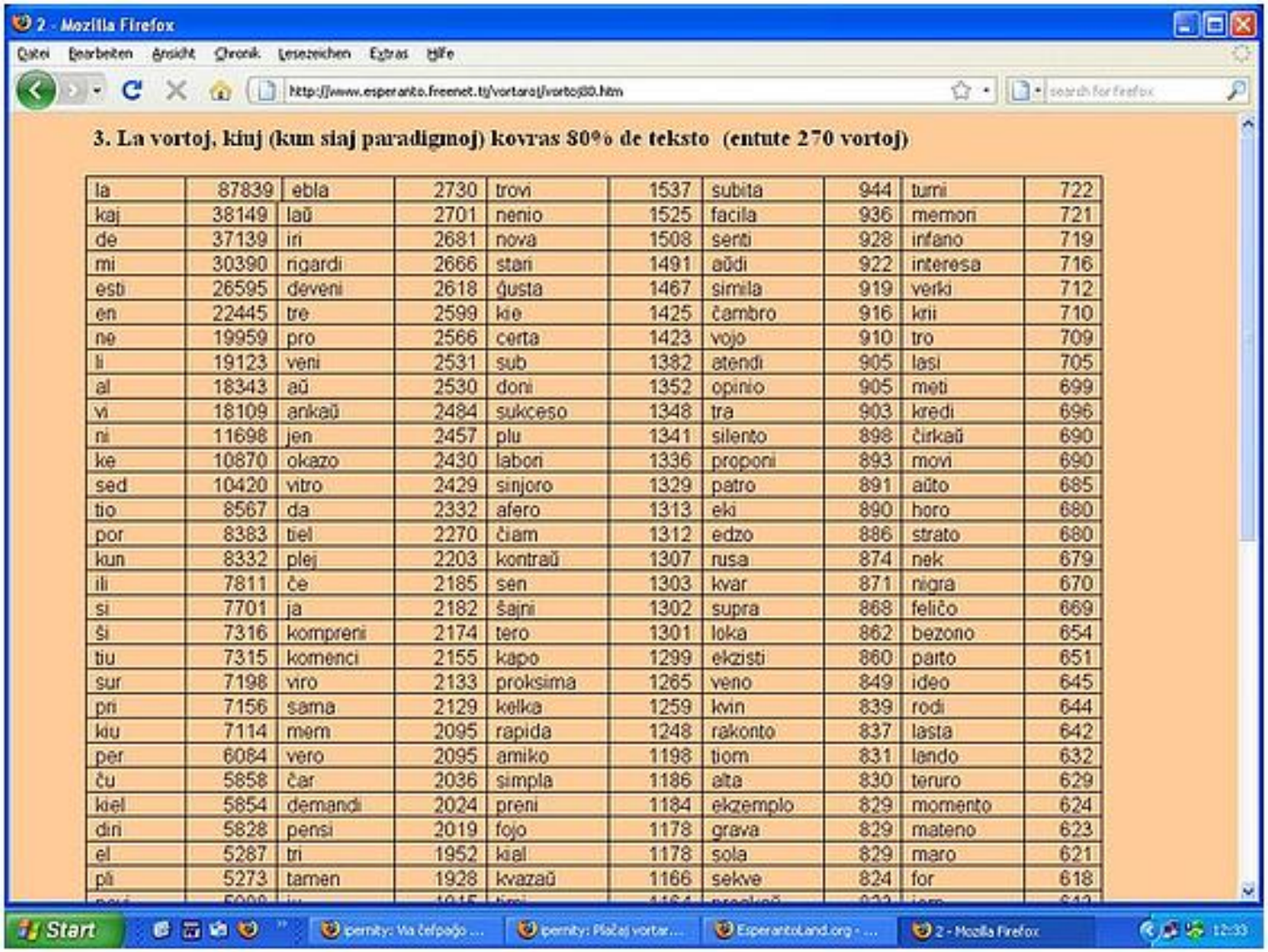
- H. Vatré (1986) “Vortstatistikaj esploroj” (Saarbrücken: Iltis) (bazita sur literaturo)

- Z. Tišljar’s (1980) listo de proks. 1300 diversaj morfemoj el parola korpuso (24 000 vortoj)

- J. Dietze (1989): proks. 1 000 oftaj Esperanto-radikoj kun germanaj tradukoj (30 000 vortformoj el Esperanto-ĵurnaloj, kiel ekz. “Paco”, “Hungara vivo”, “El Popola Ĉinio”)



- Z.D. Usmanov, F. Šukurov kaj S. Jokubzoda (sen jaro, Duŝanbe) 1,450,000 vortformoj el tradukita kaj origina Esperanto-literaturo



la	87839	ebla	2730	trovi	1537	subita	944	turni	722
kaj	38149	laŭ	2701	nenio	1525	facila	936	memori	721
de	37139	iri	2681	nova	1508	senti	928	infano	719
mi	30390	rigardi	2666	stari	1491	aŭdi	922	interesa	716
esti	26595	deveni	2618	ĝusta	1467	simila	919	verki	712
en	22445	tre	2599	kie	1425	ĉambro	916	krii	710
ne	19959	pro	2566	certa	1423	vojo	910	tro	709
li	19123	veni	2531	sub	1382	atendi	905	lasi	705
al	18343	aŭ	2530	doni	1352	opinio	905	meti	699
vi	18109	ankaŭ	2484	sukceso	1348	tra	903	kredi	696
ni	11698	jen	2457	plu	1341	silento	898	ĉirkaŭ	690
ke	10870	okazo	2430	labori	1336	proponi	893	movi	690
sed	10420	vitro	2429	sinjoro	1329	patro	891	aŭto	685
tio	8567	da	2332	afero	1313	eki	890	horo	680
por	8383	tiel	2270	ĉiam	1312	edzo	886	strato	680
kun	8332	plej	2203	kontraŭ	1307	rusa	874	nek	679
ili	7811	ĉe	2185	sen	1303	kvar	871	nigra	670
si	7701	ja	2182	ŝajni	1302	supra	868	feliĉo	669
ŝi	7316	kompreni	2174	tero	1301	loka	862	bezono	654
tiu	7315	komenci	2155	kapo	1299	ekzisti	860	parto	651
sur	7198	viro	2133	proksima	1265	veno	849	ideo	645
pri	7156	sama	2129	kelka	1259	kvin	839	rodi	644
kiu	7114	mem	2095	rapida	1248	rakonto	837	lasta	642
per	6084	vero	2095	amiko	1198	tiom	831	lando	632
ĉu	5858	ĉar	2036	simpla	1186	alta	830	teruro	629
kiel	5854	demandi	2024	preni	1184	ekzemplo	829	momento	624
diri	5828	pensi	2019	fojo	1178	grava	829	mateno	623
el	5287	tri	1952	kiel	1178	sola	829	maro	621
pli	5273	tamen	1928	kvazaŭ	1166	sekve	824	for	618

Dietze (1989)	Ivanov (2003)	Usmanov k.a. (2010)
La	La	La
De	De	Kaj
Kaj	Kaj	De
En	En	Mi
Est	Estas	Esti
Al	Al	En

The dictionary is based
on the following sources:
{ newspaper texts collected from the Internet in 2011,
{ texts from the Esperanto edition of Wikipedia as it appeared in 2012,
{ literary texts collected from various web sources in 2012,
{ further randomly selected Internet texts gathered in 2011 and 2012.

[illegible]e words and their frequency were determined on the basis of five different sources:
{ newspaper texts collected from the Internet in 2011 (approx. 50,000 sentences),
{ texts from the Esperanto edition of Wikipedia (as it appeared in 2012) (approx. 650,000 different sentences),
{ literary texts collected from various web sources in 2012 (approx. 600,000 sentences),
{ randomly collected Internet texts, crawled in 2011 (approx. 280,000 sentences),
{ randomly collected Internet texts, crawled in 2012 (approx. 1.1 million different sentences).
[illegible]e material was crawled by the Natural

The words and their frequency were determined on the basis of five different
sources:
{ newspaper texts collected from the Internet in 2011 (approx. 50,000 sentences),
{ texts from the Esperanto edition of Wikipedia (as it appeared in 2012)
(approx. 650,000 different sentences),
{ literary texts collected from various web sources in 2012 (approx. 600,000
sentences),
{ randomly collected Internet texts, crawled in 2011 (approx. 280,000 sentences),
{ randomly collected Internet texts, crawled in 2012 (approx. 1.1 million
different sentences).
The material was crawled by the Natural

Table 3.5: Most frequent words with one hyphen

Position in Wordlist	Frequency Class	Word
408	9	*s-ro*
762	9	*S-ro*
1,438	10	*Esperanto-Asocio*
1,573	10	*Esperanto-movado*
2,126	11	*D-ro*

Table 3.6: Most frequent words with two hyphens

Position in Wordlist	Frequency Class	Word
12,213	14	*TEJO-aktuale-n*
14,685	14	*Rio-de-Ĵanejro*
15,631	14	*Ku-i-o*
17,604	14	*vid-al-vide*
21,420	15	*Chaux-de-Fonds*

Table 3.10: Longest words in various frequency ranges

N	Longest word (Translation)	Length
1	*la*	2
	the	
10	*estas*	5
	is/are	
100	*Esperanto*	9
	Esperanto	
1,000	*esperantistoj*	13
	Esperantists	
10,000	*kvadratkilometroj*	17
	square kilometres	
100,000	*Esperanto-parolantoj*	20
	Esperanto speakers	
	kontraŭrevoluciuloj	17
	counterrevolutionaries	
1,000,000	*töredezettségmentesítőtleníttethetet(-)lenségtelenítőtlenkedhetnétek*	65
	(Hungarian)	

gogogoĥo is the Esperanto version of the famous Welsh place

Table 3.11: Longest words without hyphen

Word (Translation)	Length	FC
töredezettségmentesítőtleníttethetet(-)lenségtelenítőtlenkedhetnétek (Hungarian)	65	22
Rindfleischetikettierungsüberwachungs(-)aufgabenübertragungsgesetz (German)	63	20
Llanfairpwllgwyngyllgogerychwyrndrobwllllantysilio(-)gogogoch (Welsh)	58	20
Lanvajrpulgvingilgogeriĥrindrobllantisiliogogogoĥo (Esperantized name of the Welsh location)	50	22
Pneumoultramicroscopicossilicovulcanoconiótico (Portuguese)	46	21
Chargoggagoggmanchauggagoggchaubunagungamaugg Lake in the USA	45	22
Hottentottenstottertrottelmutterattentäter (German)	42	22
dinitropolisakarozidoputinidometilenoido (name of a substance; invented term - cf. p. 27)	40	22

Table 3.12: Longest words with hyphen

Word (Translation)	Length	FC
dumvintra-perkamiono-surglaciit-tundralagoj-vojaĝo	50	22
‘a journey during winter by means of a lorry on frozen tundra lakes’		
la-vorto-kiun-plej-vi-ŝatas-kaj-vi-mem-elpensis	47	22
’the word that you like best and that you invented yourself ’		
kupro-indiumo-galiumo-sulfuro-seleno-interligoj	47	22
‘Copper-indium-gallium-sulphur-selenium-combination’		
a-lalla-lalla-rumba-kamanda-lind-or-burúmë (Entish)	42	22
redaktoro-eldonanto-presanto-administranto	42	20
‘editor-publisher-printer-manager’		
alternativa-alpropriga-peranta-kunvivema	40	22
‘alternative-appropriate-intermediate-convivial’		
angla-germana-hispana-Esperanto-hungara	39	22
‘English-German-Spanish-Esperanto-Hungarian’		
kosmopolitismo-homaranismo-sennaciismo	38	22
‘cosmopolitism-homaranism-anationalism’		

t Favoriten Extras ?

WORTSCHATZ
UNIVERSITÄT LEIPZIG

Willkommen beim Wortschatz-Portal.

Wort:

Suche! ?

Beachte Groß-/Kleinschreibung

Die Daten werden aus sorgfältig ausgewählten öffentlich zugänglichen Quellen automatisch erhoben. Die Beispielsätze werden automatisch ausgewählt und stellen keine Meinungsäußerung des Projektes Deutscher Wortschatz dar. Für die darin enthaltenen Inhalte und Meinungen sind ausschließlich die Autoren verantwortlich. Auch ohne besondere Kennzeichnung unterliegen im Wortschatz wiedergegebene Marken wie Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. den gesetzlichen Bestimmungen. Die synonyme Verwendung eines Trademarks beschreibt nicht notwendigerweise produktspezifische Eigenschaften sondern kennzeichnet stattdessen die Verwendung des Begriffs im allgemeinsprachlichen Kontext.

Wörter des Tages

e tagesaktuellen Begriffe. Ausgewählt aus ageszeitungen und Newsdiensten. Täglich um Uhr früh. Seit April 2002 auf Deutsch - und it März 2006 auch auf Norwegisch. Im Test: tzt auch als RSS 2.0!

Surfhilfe NextLinks und Crawler FindLinks

Die neue Version von NextLinks unterstützt Sie beim Surfen mit den meisten aktuellen Browsern. FindLinks nutzt freie Kapazitäten Ihres PCs, um das Internet zu analysieren und neue Daten für den Surfguide Nextlinks bereitzustellen. Helfen Sie mit bei der Analyse des Internet!

Webservices

Mit den Webservices ist ein direkter Zugriff auf die Daten des Projektes Deutscher Wortschatz aus einer beliebigen Software heraus möglich.

Int. Wortschatz Portal

Auf unserem internationalen Wortschatzportal in englischer Sprache können Sie derzeit in Korpora 229 verschiedener Sprachen Wörter nachschlagen. Um das Suchen für Sie einfacher zu

Wörterbuch

Über 100.000 Wörter und Wendungen auf Deutsch und Englisch. Die Besonderheit: Häufigkeitsangaben verraten Ihnen, wie oft die einzelnen Wörter verwendet werden. Seit Januar

Feedback

Feedback der Nutzer zu den verschiedenen Services des Wortschatz Portals.

WORTSCHATZ
UNIVERSITÄT LEIPZIG

Search in 229 Corpus-Based Monolingual Dictionaries

Newest Dictionaries

Word: Find! ?

Active dictionary: English case sensitive search

Random words:

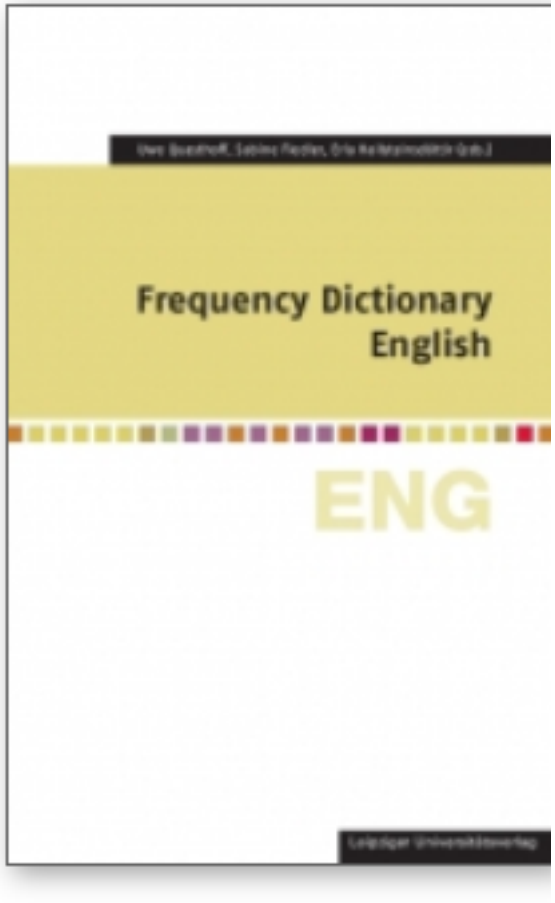
gap adult fence beer deployed

Change Dictionary:

Abkhaz	Acholi	Afrikaans	Akan	Albanian	Amharic
Arabic	Aragonese	Armenian	Aromanian	Assamese	Assyrian Neo-Aramaic
Asturian	Avar	Azerbaijanian	Balkar	Bamanankan	Banjar
Bashkir	Basque	Bavarian	Belarusan	Bengali	Bicolano
Bishnupriya	Bosnian	Breton	Bulgarian	Buriat	Catalan
Cebuano	Chavacano	Chechen	Cherokee	Chinese (simplified)	Chinese, Min Dong
Chinese, Min Nan	Chuvash	Cornish	Corsican	Crimean Tatar	Croatian
Czech	Danish	Dimli	Dutch	Emiliano-Romagnolo	**English**
English (AU)	English (CA)	English (NZ)	English (UK)	Esperanto	Estonian
Faroese	Fijian	Finnish	French	Frisian, Northern	Frisian, Western
Friulian	Fulah	Gaelic, Irish	Gaelic, Scottish	Gagauz	Galician
Ganda	Georgian	German	German (CH)	German, Swiss	Gilaki
Goan Konkani	Greek	Greenlandic	Guarani	Gujarati	Haitian
Hausa	Hebrew	Hindi	Hindi, Fiji	Hungarian	Icelandic
Ido	Ilocano	Indonesian	Interlingua	Interlingue	Italian
Japanese	Javanese	Kölsch	Kabardian	Kabyle	Kalmyk-Oirat
Kannada	Karakalpak	Kashubian	Kazakh	Khmer, Central	Kiswahili
Klingon	Komi	Konkani	Korean	Kurdish	Kyrgyz
Ladino	Lak	Lao	Latgalian	Latin	Latvian
Ligurian	Limburgish	Lingala	Lithuanian	Lushai	Luxemburgian
Macedonian	Malagasy	Malay	Malayalam	Maldivian	Maltese
Manx	Maori	Marathi	Mari, Meadow	Mari, Western	Mingrelian
Mirandese	Moksha	Mongolian (Cyrillic)	Mongolian (traditional)	Nahuatl	Navajo
Nepali	Newari	Norse, Old	Norwegian (Bokmål)	Norwegian (Nynorsk)	Novial
Occitan	Old English	Oriya	Oromo	Oromo, West Central	Ossetian
Pampanga	Pangasinan	Panjabi	Panjabi, Western	Papiamento	Pashto
Pennsylvanian Dutch	Persian	Picard	Piemontese	Polish	Portuguese (Brazil)
Portuguese (Macao)	Portuguese (Portugal)	Romanian	Romansch	Romany	Russian
Rusyn	Saami, North	Sami	Samoan	Samogitian	Sanskrit
Sardinian	Scots	Serbian	Shona	Sicilian	Silesian
Sindhi	Sinhala	Slovak	Slovenian	Somali	Sorbian (Lower)
Sorbian (Upper)	Sotho, Northern	Sotho, Southern	Spanish	Spanish (Mexico)	Sundanesian
Swahili	Swedish	Tagalog	Tajik	Tama	Tamil
Tatar	Telugu	Tetun	Thai	Tibetan, Central	Tigrigna
Tok Pisin	Tongan	Tswana	Turkish	Turkmen, Latin	Udmurt
Ukrainian	Urdu	Uyghur	Uzbek	Uzbek, Latin	Venda
Venetian	Vietnamese	Vlaams	Walloon	Waray	Welsh
Wolof	Yakut	Yiddish	Yoruba	Zeeuws	Zhuang
Zulu					

Download our Corpora browser and databases

http://univerlag-leipzig.de/catalog/category/142-Frequency_Dictionaries_Haeufigkeitswoerterbuecher



Einige erste Ergebnisse:

(1) Häufigste Wörter

1	la	3583018
2	de	1964614
3	kaj	1440028
4	en	960386
5	estas	666435
6	al	591090
7	La	468346
8	ne	465698
9	ke	392919
10	por	371602
11	mi	341218
12	estis	334510
13	li	325448
14	pri	272278
15	kun	248589

(2) Frequenz von Buchstaben (diakrit. Zeichen)

(3) Längste Wörter, Silbenzahl ...

.... Häufigste Farbe, häufigste Monat ....

Table 3.14: Summary of statistical data

Parameter	**Value**
Range of the alphabet (small letters)	31
Vowel-consonant ratio in the texts	38:62
Proportion of words with capital letters in the texts	36.0 %
Proportion of words with capital letters among the most frequent 100,000 words	68.8 %
Mandelbrot-slope for Zipf ’s law	-1.046
Text coverage by most frequent 10 words	15.5 %
Text coverage by most frequent 100 words	39.1 %
Text coverage by most frequent 1000 words	59.2 %
Average word length (with repetitions)	5.86
Average word length of the most frequent 100,000 words	9.73
Growth rate of word length in the frequency list	1.99
Average number of syllables (with repetition)	1.92
Average number of syllables of the most frequent 100,000 words	3.10
Number of letter trigrams among the most frequent 1,000 words	1,159
Number of letter trigrams among the most frequent 100,000 words	11,007
Percentage of words with at least one hyphen among the most frequent 100,000 words	1.93 %