Language Resources Extracted from Wikipedia

Denny Vrandečić
AIFB, Karlsruhe Institute
of Technology (KIT)
76128 Karlsruhe, Germany
and Wikimedia Deutschland
10777 Berlin, Germany
denny.vrandecic@kit.edu

Philipp Sorg
AIFB, Karlsruhe Institute
of Technology (KIT)
76128 Karlsruhe, Germany
philipp.sorg@kit.edu

Rudi Studer
AIFB, Karlsruhe Institute
of Technology (KIT)
76128 Karlsruhe, Germany
rudi.studer@kit.edu

ABSTRACT

Wikipedia provides an interesting amount of text for more than hundred languages. This also includes languages where no reference corpora or other linguistic resources are easily available. We have extracted background language models built from the content of Wikipedia in various languages. The models generated from Simple and English Wikipedia are compared to language models derived from other established corpora. The differences between the models in regard to term coverage, term distribution and correlation are described and discussed. We provide access to the full dataset and create visualizations of the language models that can be used exploratory. The paper describes the newly released dataset for 33 languages, and the services that we provide on top of them.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Language models; I.2.6 [**Learning**]: Knowledge acquisition

General Terms

Languages, Measurement

1. INTRODUCTION

Statistical natural language processing requires corpora of text written in the language that is going to be processed. Whereas widely studied languages like English and Chinese traditionally have excellent coverage with corpora of relevant sizes, for example the Brown corpus [4] or Modern Chinese Language Corpus, this is not true for many languages that have not been studied in such depth and breath. For some of these languages, viable corpora are still painfully lacking.

Wikipedia is a Web-based, collaboratively written encyclopedia [1] with official editions in more than 250 languages. Most of these language editions of Wikipedia exceed one million words, thus exceeding the well-known and widely-used Brown corpus in size.

We have taken the text of several Wikipedia language editions, cleansed it, and created corpora for 33 languages. In order to evaluate how viable these corpora are, we have calculated unigram language models for the English Wikipedia, and compared it to widely used corpora. Since the English Wikipedia edition is far larger than any other — and size of a corpus is a crucial factor for its viability — we have also taken the Simple English Wikipedia edition, being smaller than many other language editions, and compared it to the reference corpora as well. The results of this comparison show that the language models derived from the Simple English Wikipedia are strongly correlated with the models of much larger corpora. This gives support to our assumption that the language models created from the corpora of other language editions of Wikipedia have an acceptable quality, as long as they compare favorably to the Simple English Wikipedia.

We make the generated unigram language models and the corpora available. The full data sets can be downloaded.[1] The website also provides a novel, graphical corpus exploration tool – Corpex – not only over the newly created corpora that we report on here, but also usable for already established corpora like the Brown corpus.

The next section introduces some background information on Wikipedia and language corpora, followed by related work in Section 3. We then describe the language models in Section 4, including their properties and acquisition. Section 5 compares the Wikipedia-acquired language models with widely-used language models and points out the differences and commonalities. We finish with the conclusions and future work in Section 6.

2. BACKGROUND

2.1 Wikipedia

Wikipedia[2] [1] is a wiki-based collaboratively edited encyclopedia. It aims to *"gather the sum of human knowledge"*. Today Wikipedia provides more than 17 million articles in 279 languages, and further a small set of incubator languages. It is run on the MediaWiki software [3], which was developed specifically for Wikipedia. In general, every article is open to be edited by anyone, (mostly) through the browser. Even though this editing environment is very limited compared to rich text editing offered by desktop word processing systems, the continuous effort has led to a com-

K-CAP'11, June 26–29, 2011, Banff, Alberta, Canada.

[1] `http://km.aifb.kit.edu/sites/corpex`
[2] `http://www.wikipedia.org`

petitive, and widely used, encyclopedia. The content is offered under a free license, which allows us to process the text and publish the resulting data.

As stated, Wikipedia exists in many language editions. A special language edition is the so called *Simple English* Wikipedia.[3] The goal of the Simple English Wikipedia is to provide an encyclopedic resource for users without a full grasp of the English language, e.g. children learning to read and write, or non-native speakers learning the language. For our work this means that we have, besides the actual English Wikipedia,[4] a second Wikipedia edition in the English language that is much smaller in size.

2.2 Language Corpora

Language corpora are the main tool of research in statistical natural language processing. They are big samples of text that have the purpose to represent the usage of a specific natural language. Using a corpus in a specific language, different statistical characteristics of this language can be defined. For example the distribution of terms is often used in NLP applications. This distribution is either measured independently (unigram model) or in context of other terms (n-gram model).

Language models extracted from these language corpora are used in all application that *recognize* or *synthesize* natural language. Examples for applications that recognize natural language are:

Speech Recognition: Language models are used to identify the text sequence with the highest probability matching the speech input.

Spell-checking: In the context of previous terms, the most probable spelling of terms is identified based on the language model.

Syntax Parser: Syntax parsers depend on language models to build syntax trees of natural language sentences. If annotated with part-of-speech tags, language corpora are also used as training data.

The examples presented above describe applications of language corpora to recognition tasks. Further, language models are also applied in systems that synthesize text:

Auto-completion: The term distribution encoded in language models can be used to auto-complete input of users. In many cases, these are language models optimized for a specific task, for example language models of queries in search systems. For auto-completion, often the context of previous terms is used to compute the probability of the next term.

Machine Translation: Machine translation systems recognize text in one language and synthesize text in another language. To ensure grammatical correctness or at least readability of the synthesized text, language models can be used to identify the most probable word order in the output sentences.

Over the years, a number of corpora have been established. These corpora contain documents of high quality with little noise. Examples are news items or legislative documents. As these corpora mostly contain full sentences that are grammatical correct, they are often used as representative corpora of the according languages. This is also a main difference to automatically constructed corpora. Examples for such automatically constructed corpora are collections of Web documents that are crawled from the Web. In these corpora, the level of noise is much higher as they contain for example syntax elements or misspelled terms.

		Unique	
Corpus	*Docs*	*Terms*	*Tokens*
Brown	500	36,708	958,352
Reuters	806,791	369,099	171,805,788
TREC4+5	528,155	501,358	217,215,818
JRC-Acquis (EN)	7,745	229,618	41,163,635

Table 1: Size of reference copora measured by number of documents, number of unique terms (or types) and total number of tokens.

In our experiments, we use several English corpora for comparison. We focus on English due to the availability of English corpora. In other language such as Croatian or Slovenian, only few corpora are freely available. In detail, we use the following corpora:

Brown Corpus: [4] This corpus was published as a standard corpus of present-day edited American English. It has been manually tagged with part-of-speech information and has therefore often been used as training and testing corpus for deep analysis NLP applications.

TREC Text Research Collection V.4+5: [5] This corpus was used in the ad-hoc retrieval challenge at TREC. It contains a compilation of documents from the Financial Times Limited, the Congressional Record of the 103rd Congress, the Federal Register, the Foreign Broadcast Information Service, and the Los Angeles Times.

Reuters Corpus (Volume 1): [6] Collection of news stories in the English language that has often been used as real-world benchmarking corpus.

JRC-Acquis: [7] Legislative documents of the European Union that are translated in many languages. This corpus is often used as a parallel corpus, as the sentences are aligned across the translations.

Table 1 contains statistics about the size of the presented reference corpora in respect to the number of documents, unique terms and tokens.

3. RELATED WORK

In recent years, Wikipedia has often been used as language resource. An example is presented by Tan and Peng [12]. They use a Wikipedia based n-gram model for their approach to query segmentation. Using the model extracted

[3] http://simple.wikipedia.org

[4] http://en.wikipedia.org

[5] http://trec.nist.gov/data/docs_eng.html

[6] http://about.reuters.com/researchandstandards/corpus/

[7] http://langtech.jrc.it/JRC-Acquis.html

from the English Wikipedia, they achieve performance improvements of 24%. They therefore present a successful application of the language models derived from the English Wikipedia. In this paper we show that other language editions of Wikipedia can be exploited in the same way and are therefore valuable resources for language models in various languages.

Exploiting the multilingual aspects of Wikipedia, different approaches have been suggested to use the Wikipedia database in different languages for multilingual Information Retrieval (IR) [11, 10]. The language models presented in this paper have no dependencies across languages. For each language, we suggest to exploit the according Wikipedia edition to build a language resource that is specific for this language. However, these resources could also be applied in cross-lingual systems, as many of these systems also rely on language-specific background models.

Apart from Wikipedia, language corpora have also been built from other Web resources. Recently, a number of huge Web-based corpora have been made available by Google[8] and Microsoft.[9] Baroni et al. [2] constructed large corpora of English, German and Italian Web documents that were also annotated based on linguistic processing. Ghani et al. [5] proposed to use the Web to create language corpora for minority languages. By creating and adapting queries for Internet search engines, they collect documents with a broad topic coverage. In this paper, we claim that the coverage is already given by Wikipedia for many languages. We show that Wikipedia based language models have similar properties than language models derived from traditionally used corpora. This is not known for the Web based corpora. Further, Wikipedia supports many more languages than the above mentioned Web based resources. Finally, given the lower effort to access Wikipedia compared to crawling the Web, we claim that using Wikipedia as a resource for language models is an appropriate choice in many application scenarios.

There are other multilingual corpora that are not based on Web documents. For example, Koehn [6] created a parallel corpus of the proceedings of the European Parliament that is mainly used to train machine translation systems. This corpus is similar to the JRC-Acquis corpus used in our experiments. However, the results of our experiments support the conclusion that these corpora can not be used to build representative language models. A possible explanation is that these multilingual corpora are much smaller compared to the other resources and that they are often focused on specific topic fields.

The application of language models are manifold. A prominent example are retrieval models used in IR. Zhai and Lafferty [13] suggest to use background language models for smoothing. These models are based on a collection of datasets that also includes the TREC4+5 corpus. This motivates the comparison of the Wikipedia language models to this corpus presented in Section 5. Most of the related work about the application of language models is based on tasks in English. The language models that we suggest in this paper could be used to apply the same approaches in various languages. The improvements achieved in specific tasks such as IR through the usage of background language models, could then be replicated and verified in experiments using corpora in other languages than English as well.

4. THE LANGUAGE MODELS

4.1 Acquistion

Articles in Wikipedia are written using the MediaWiki syntax, a wiki syntax offering a flexible, but very messy mix of some HTML elements and some simple markup. There exists no proper formal definition for the MediaWiki syntax. It is hard to discern which parts of the source text of an article is actual content, and which parts provide further functions, like navigation, images, layout, etc. This introduces a lot of noise to the text.

We have filtered the article source code quite strictly, throwing away roughly a fourth of the whole content. This includes most notably all template calls, which are often used, e.g., to create infoboxes and navigational elements. The actual script that provides the filtering is available on the Corpex website as open source, so that it can be reused and further refined. When exploring the corpus, one can easily see that quite some noise remains. We aim to further clean up the data and improve the corpora over time.

The content of the Wikipedia editions is provided as XML dumps.[10] We have selected only the actual encyclopedic articles, and not the numerous pages surrounding the project, including discussion pages for the articles, project management pages, user pages, etc., as we expect those to introduce quite some bias and idiosyncrasies. The table in Figure 1 contains the date when the XML dump was created, for reference and reproducibility of the results. Combined, we have processed around 75 gigabytes of data.

4.2 Statistical overview

Figure 1 offers an overview of some statistics on the acquired corpora and the generated language models.

The rank of the Wikipedia language edition, the depth, and the number of tokens are meant as indicators for the quality of the generated corpus. As discussed in Section 5, we estimate the quality of the Simple English Wikipedia compared to other corpora. The results indicate that any corpora with at least the depth, rank, and size (w.r.t number of tokens) should be at least as reliable as a language corpus as the Simple English Wikipedia is. The corpora created from other language editions, that do not fulfill these conditions, should be used with more care as their quality is not sufficiently demonstrated yet.

The depth in Figure 1 refers to a measure introduced by the Wikipedia community, called the *Wikipedia article depth*,[11] which is meant as a rough indicator for the strength of the collaboration within the wiki. It assumes that a high number of edits and support pages indicate a higher quality of the language edition overall, and is defined as

$$\text{Depth} = \frac{\#\text{Edits}}{\#\text{Total pages}}(\frac{\#\text{Non-Article pages}}{\#\text{Articlepages}})^2$$

4.3 Web site and service

We provide a web site that allows to easily explore the created frequency distributions, called Corpex. Corpex allows

[8] http://ngrams.googlelabs.com/

[9] http://research.microsoft.com/en-us/collaboration/focus/cs/web-ngram.aspx

[10] http://downloads.wikimedia.org

[11] http://meta.wikimedia.org/wiki/Depth

to select the corpus of interest and then to explore the probabilities within the given corpus, both on word completion and the next character.

Corpex also allows to compare two different corpora. Figure 2 and 3 are screenshot of Corpex for the letter t on the English Wikipedia and the Brown corpus respectively. This allows the user to exploratively acquire an understanding for and compare the frequency distribution of words in the explored corpora.

Corpex also provides a RESTful service that allows to gather the same data in JSON, so that it can be further processed. Corpex is implemented using dynamically materialized prefix trees on the file system, thus leading to response times of under 50 milliseconds in general.

Corpex also offers the complete frequency lists for download, so that it can be used for further processing or analysis.[12]

5. ANALYSIS AND DISCUSSION

In this paper, we propose to use Wikipedia in various languages as a language corpus. For many of these languages, no other corpora are available. However for English, established corpora as presented in Section 2.2 can be compared to the Wikipedia corpus. We intend to show that the Wikipedia corpus in English as well as Simple English have the same characteristics as the reference corpora. By extending this hypothesis, we claim that the Wikipedia corpus in other languages can also be used as a background language model. This is based on the fact that the language model derived from the Simple English Wikipedia, which is much smaller than the English Wikipedia, is very similar to the language model derived from the English Wikipedia. However this can not be verified due to the lack of appropriate reference corpora in these languages.

We will use several measures to compare the different corpora in our experiments:

Term Overlap: The percentage of common terms in two corpora in respect to all terms in each corpus. This also determines the *token overlap*, that is defined as the share of tokens of these common terms to all tokens. Both measures are non-symmetric for two corpora.

Pearson Product Moment Correlation: The number of occurrences of each term in two corpora is interpreted as drawings from two random variables X and Y, the Pearson product moment correlation coefficient computes the linear dependence of these variables [9]. A value of 1 implies perfect linear correlation and a value of 0 implies no linear correlation. However this does not exclude other non-linear correlations of X and Y.

Linear correlation of term occurrences in two corpora shows that terms are equally distributed and can therefore be applied as similarity measure of language corpora.

Jensen-Shannon Divergence: Considering the language models defined by two corpora, similarity measures defined on probability distributions can be used to compare these corpora. We use the Jensen-Shannon divergence [8] that is based on the Kullback-Leibler divergence [7]. These measures are rooted in information theory and measure the entropy of distributions given the information provided by another distribution. The Jensen-Shannon divergence has been established as a standard measure to compare two probability distributions.

Applied to the comparison of corpora, a low divergence value means that there is no information gain of one language model given the information of the other language model, which implies that the distributions are similar.

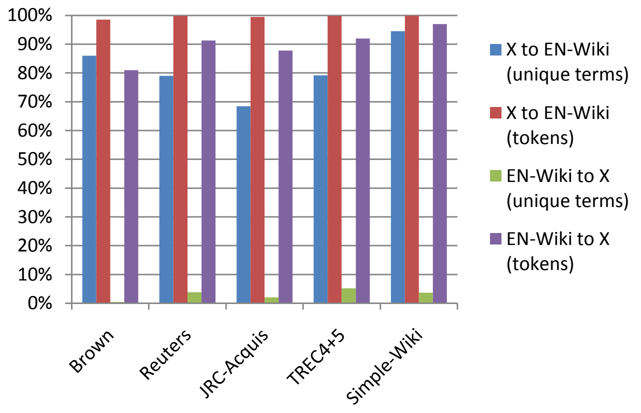


Figure 4: Common unique terms and overlapping tokens in the English Wikipedia and the reference corpora. The relative overlap values presented in this chart correspond to the percentage of common terms or overlapping tokens in respect to all terms or tokens in the respective corpus.

5.1 Term Overlap (English)

The term overlap in pairwise comparison of the reference corpora and Simple and English Wikipedia are presented in Table 2. The overlap between the reference corpora is astonishingly low — mostly below 50% with exception of the small Brown corpus. Even comparing the other corpora to the large English Wikipedia, the overlap is only between 79% and 86%.

A qualitative analysis of non-common terms shows that many of these terms are actually noise. While being written mainly in English, the used corpora also contain terms of other languages. Many of these foreign terms are not found in the English Wikipedia.

Despite of the low overlap of unique terms, the token overlap is much higher. In Figure 4 we visualize both term and token overlap of the English Wikipedia to all other corpora. Considering the overlap of the reference corpora to Wikipedia, the 80% common terms cover more than 99% of all tokens. In the other direction, the .4% to 5% of the Wikipedia terms that are also present in the reference corpora cover more than 80% of the tokens in Wikipedia. This clearly shows that the common terms are the most frequent terms and therefore important to characterize the language models.

5.2 Correlation of Language Models

We visualize the results of the correlation analysis using net charts. Each dataset is represented by an axis and a

[12] http://km.aifb.kit.edu/sites/corpex/data/

	Brown	Reuters	JRC-Acquis	TREC4+5	Simple-Wiki	English-Wiki
Brown	-	77%	55%	83%	72%	86%
Reuters	8%	-	18%	50%	29%	79%
JRC-Acquis	9%	30%	-	34%	28%	69%
TREC4+5	6%	37%	15%	-	26%	79%
Simple	9%	37%	22%	45%	-	94%
English	.4%	4%	2%	5%	3%	-

Table 2: Common terms in the reference corpora and Simple and English Wikipedia.

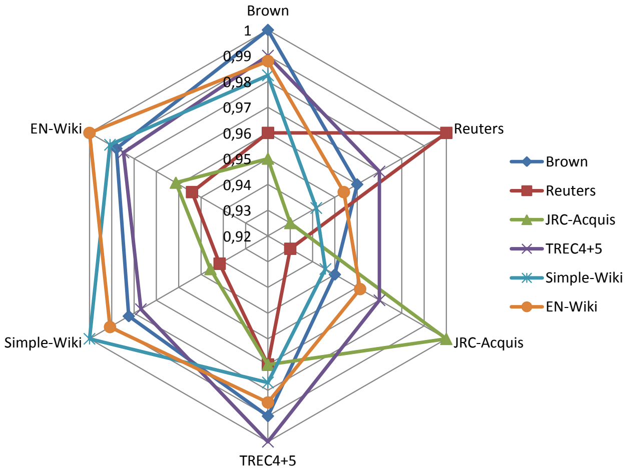


Figure 5: Net chart visualizing Pearson product moment correlation coefficient between any pair of the English corpora. For each corpus pair, only common terms are considered for the correlation value.

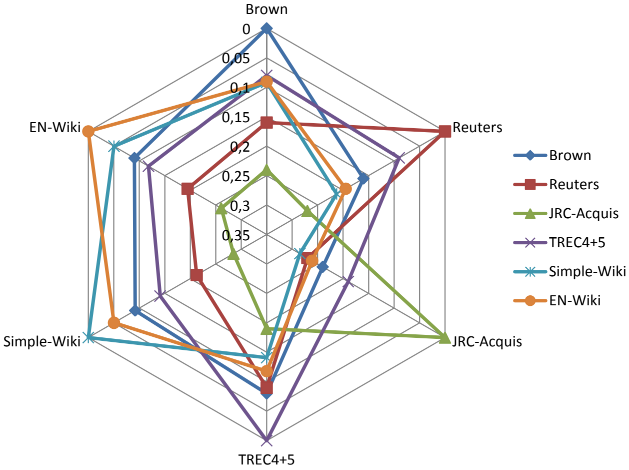


Figure 6: Net chart visualizing the Jensen-Shannon divergence between any pair of the English corpora.

data line. This allows to draw the pairwise comparison of any two corpora into one chart.

The values of the Pearson product moment correlation coefficient are presented in Figure 5, the values for the Jensen–Shannon divergence in Figure 6. The patterns found in both charts are very similar. This shows that both measures – motivated by a probabilistic model as well as by an information theoretic model – identify the same level of correlation of all corpus pairs. We draw the following conclusions of the results presented in Figure 5 and 6:

- Overall the correlation between any pair of corpus is very high. The correlation coefficient is always above .93 and the divergence value below .29.

  To get an idea of the range of these measures, we also compared the English Wikipedia to the German Wikipedia. Actually, the term overlap is approx. 25% in both directions covering more than 94% of all tokens in both Wikipedias. However the correlation coefficient is only .13 and the divergence .72. This shows that the correlation of the common terms is low for these corpora.

- The Brown, TREC4+5, Simple Wikipedia and English Wikipedia corpora have the highest correlation. Actually, the correlation coefficient between the Brown corpus and the English Wikipedia is .99 meaning almost perfect linear correlation. This supports our claim that Wikipedia can be used as representative language corpus like the Brown corpus for English.

  This also shows that the term distributions in Simple and English Wikipedia are very similar. While only using a fraction of the vocabulary, the content in the Simple Wikipedia is based on a similar language usage as the English Wikipedia.

- The Reuters corpus is correlated to the TREC4+5 corpus. Our hypothesis is that the large share of newspaper articles found in both corpora is the reason for this correlation.

- The JRC-Acquis corpus is the only outlier of our reference corpora. This corpus contains official legislative documents. The language usage in these documents is probably different to the language usage in Wikipedia and in newspaper articles as found in the Reuters and TREC4+5 corpus, which is supported by our findings.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we suggested to exploit Wikipedia for language models in various languages. We also presented a novel visualization that allows to interactively explore the term distributions of different corpora. In summary, our main contributions are the following:

- We acquired and published unigram language models from Wikipedia for 33 languages, where some of them did not have yet any language models available.

- We compared the language models retrieved from the Wikipedia corpus with other language models based on reference corpora. This comparison allows some understanding about the quality of the acquired language models. In particular, we show that the language model derived from the Simple English Wikipedia is very similar to the language model derived from the English Wikipedia. As the Simple English Wikipedia is much smaller, we argue that other Wikipedia versions of comparable size or bigger have similar properties and can therefore be used as appropriate language resources.

- We proposed a novel visualization for exploring frequency distributions, which is available on the Web.

- We published a web service to programmatically use the frequency distributions, which can be for example used for auto-completion.

With the experience gathered on creating the first few corpora, we plan on turning Corpex into a pipeline that will be able to offer the service for the full set of language editions of Wikipedia, dynamically updating as new dumps are provided. This will offer access to a growing and up to date set of corpora for over 200 languages, even though for some of them the usage scenarios may be restricted due to insufficient size or quality.

We will also explore the creation of further language models besides frequency distributions, like n-grams. We hope that the provisioning of the full datasets and of all the means to create them will enable further scenarios and services beyond those described in this paper. The further analysis of the data may lead to measures for quality of the Wikipedia language editions. Beyond Wikipedia, the corpora and language models can provide much needed resources for NLP researchers in many languages.

Acknowledgements

Research reported in this paper was funded by the European Commission through the FP7 project RENDER[13] and the German Research Foundation (DFG) through the Multipla project.[14]

7. REFERENCES

[1] P. Ayers, C. Matthews, and B. Yates. *How Wikipedia works*. No Starch Press, 2008.

[2] M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226, 2009.

[3] D. J. Barrett. *MediaWiki*. O'Reilly, 2008.

[4] W. N. Francis and H. Kucera. Brown corpus manual. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, US, 1979.

[5] R. Ghani, R. Jones, and D. Mladenić. Mining the web to create minority language corpora. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, page 279–286, New York, NY, USA, 2001. ACM.

[6] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Machine Translation Summit*, volume 5, 2005.

[7] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[8] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[9] K. Pearson. Notes on the history of correlation. *Biometrika*, 13(1):25–45, Oct. 1920.

[10] M. Potthast, B. Stein, and M. Anderka. A Wikipedia-Based multilingual retrieval model. In *Proceedings of the 30th European Conference on Information Retrieval (ECIR)*, pages 522—530, Glasgow, 2008.

[11] P. Schönhofen, A. Benczúr, I. Bíró, and K. Csalogány. Cross-Language retrieval with wikipedia. In *Advances in Multilingual and Multimodal Information Retrieval*, pages 72—79. 2008.

[12] B. Tan and F. Peng. Unsupervised query segmentation using generative language models and wikipedia. In *Proceeding of the 17th International Conference on World Wide Web*, WWW '08, page 347–356, New York, NY, USA, 2008. ACM.

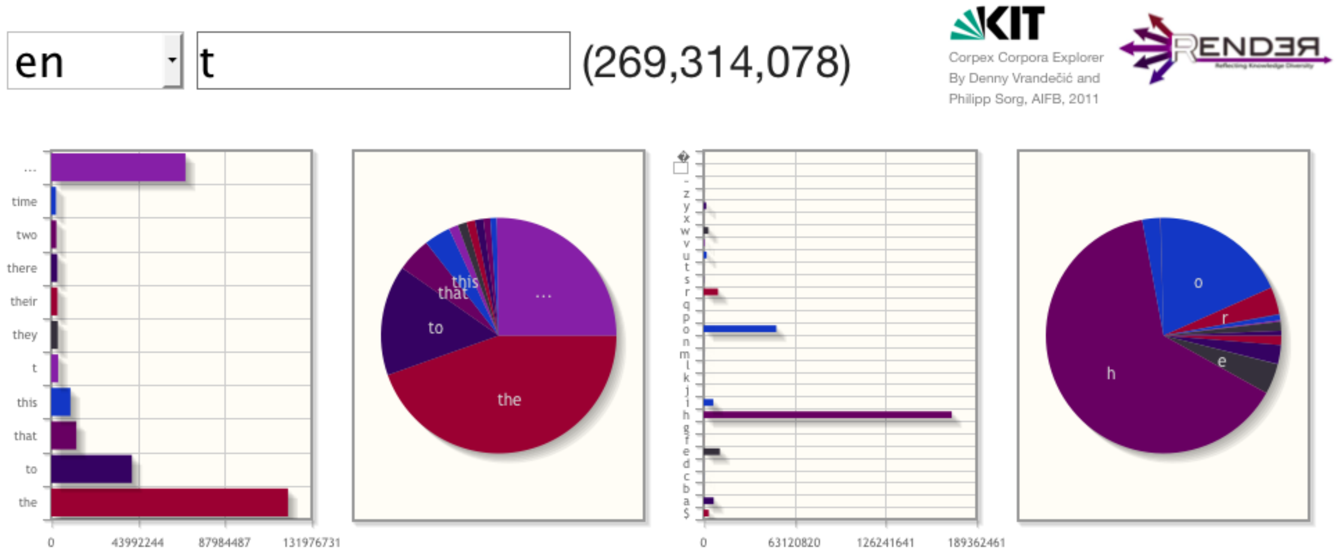
[13] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22:179–214, Apr. 2004.

[13] `http://render-project.eu`

[14] `http://www.multipla-project.org/`

Language	**C**	**R**	**D**	**Date**	**Tokens**	**Terms**	**10+**	**Top50**	**l1**	**l2**
Albanian	sq	64	27	2011/01/29	8,584,477	388,278	48,906	427	8.1	4.9
Bulgarian	bg	33	24	2011/02/02	32,518,959	844,892	124,321	618	8.4	5.3
Bosnian	bs	67	97	2011/02/02	7,392,085	453,828	54,998	1,301	8.4	5.5
Croatian	hr	37	21	2010/02/04	27,837,889	994,770	140,458	1,314	8.5	5.5
Czech	cs	17	35	2011/02/05	63,538,097	1,548,013	239,974	1,404	8.7	5.5
Danish	da	24	35	2011/01/30	36,497,359	1,006,141	126,626	404	10.3	5.3
Dutch	nl	9	30	2011/01/26	175,499,078	2,308,280	336,349	261	10.2	5.3
English	en	1	584	2011/01/15	2,014,858,488	7,659,102	1,110,470	295	8.8	4.9
– *Simple*	simple	41	55	2011/04/15	15,292,826	297,040	43,333	238	7.7	4.8
– *Brown*	brown	—	—	1960s	958,352	36,708	7,088	103	8.5	4.7
Estonian	et	38	26	2011/02/02	14,106,418	983,406	102,875	2,568	10.2	6.6
Finnish	fi	16	41	2011/01/31	57,456,478	2,686,562	292,152	3,136	12.0	7.3
French	fr	3	140	2011/01/12	486,524,274	3,107,353	497,189	243	8.4	4.9
German	de	2	88	2011/01/11	590,886,656	6,691,421	955,950	456	11.9	6.0
Greek	el	47	39	2011/02/03	25,208,880	756,738	104,184	480	8.7	5.6
Hungarian	hu	18	87	2011/02/03	77,661,090	2,641,225	304,770	1,474	10.3	6.1
Irish	ga	93	25	2011/02/06	2,819,777	145,031	16,528	183	8.1	4.8
Italian	it	5	78	2011/01/30	317,582,265	2,480,869	387,894	385	8.6	5.2
Latvian	lv	59	76	2011/02/05	8,141,029	446,366	59,215	1,667	8.6	6.1
Lithuanian	lt	28	17	2011/02/02	19,924,938	939,624	119,148	1,880	8.6	6.5
Maltese	mt	152	118	2011/01/30	1,357,178	81,034	10,808	337	7.7	5.1
Polish	pl	4	12	2011/01/27	167,946,149	2,694,814	446,576	1,687	9.0	6.0
Portuguese	pt	8	76	2011/01/24	177,010,439	1,711,936	259,032	362	8.4	5.0
Romanian	ro	20	88	2011/02/03	39,230,386	902,309	127,789	591	8.3	5.3
Serbian	sr	27	44	2010/01/30	39,351,179	1,182,685	160,632	710	8.4	5.5
Serbocroatian	sh	56	17	2010/02/07	14,644,455	731,093	93,955	1,579	8.5	5.5
Sinhalese	si	131	133	2011/02/03	4,220,958	287,042	30,722	1,170	7.9	5.2
Slovak	sk	29	21	2011/01/29	24,784,192	925,677	130,164	1,132	8.8	5.6
Slovenian	sl	35	17	2011/01/29	23,859,807	847,990	112,180	664	8.6	5.4
Spanish	es	7	171	2011/01/14	354,499,700	2,741,188	418,910	273	8.6	5.0
Swedish	sv	11	46	2011/01/29	82,785,880	1,953,939	249,579	527	10.8	5.6
Vietnamese	vi	19	44	2011/02/06	56,479,754	700,090	68,117	239	7.9	3.8
Waray-Waray	war	36	0	2011/02/05	2,589,896	130,047	7,974	15	7.6	4.8
Zulu	zu	247	0	2011/01/30	14,029	6,051	152	681	7.5	6.3

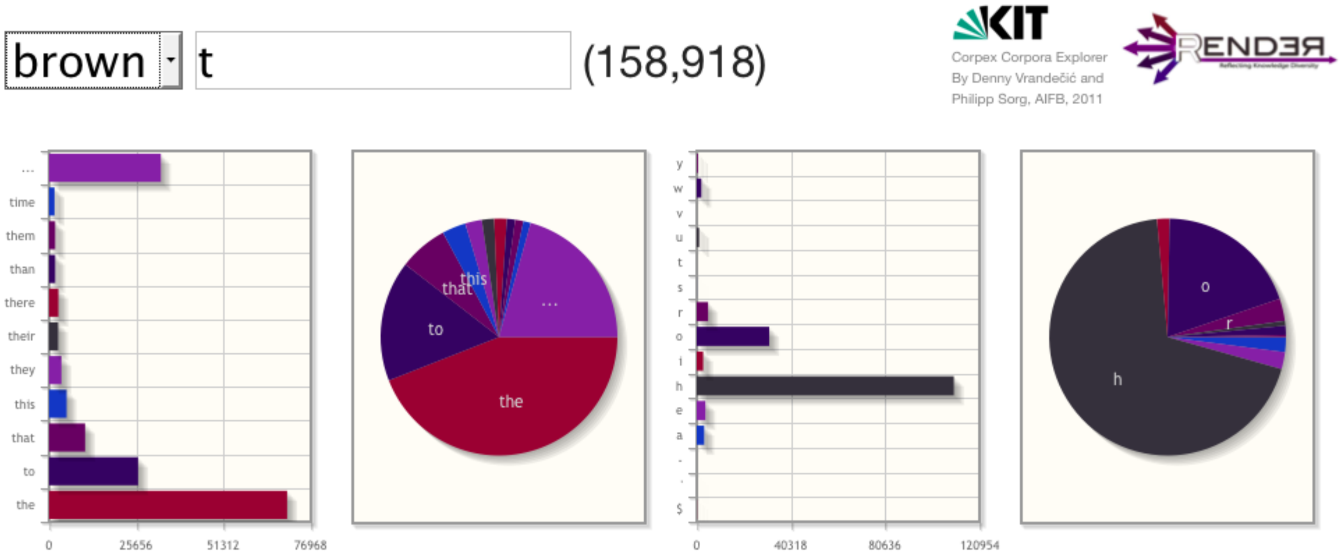
Figure 1: A few statistics on the language data. C is the language code used in the Corpex explorer (and usually also the Wikipedia language code). R is the rank of the Wikipedia language edition by number of articles (as of February 15, 2011), and D the Depth (see Section 4.2), providing a rough measure of collaboration. Date is the date when the database dump was created, that was used for the corpus creation. Tokens is the number of tokens in the given Wikipedia language edition, Terms the number of unique terms. 10+ is the number of terms that have appeared more than ten times. Top50 is the smallest number of terms that account for more than 50% of all the tokens. l1 is the average length of terms in the corpus. l2 is the average length of tokens in the corpus.



en | t | (269,314,078)

Corpex Corpora Explorer
By Denny Vrandečić and
Philipp Sorg, AIFB, 2011

Figure 2: Corpex website screenshot, showing the frequency distribution model created from the English Wikipedia corpus for the letter 't'. The two charts to the left show the distribution of word completions, the two charts on the right show the probability for the next character. Both times the barchart and the piechart show the same data.



brown | t | (158,918)

Corpex Corpora Explorer
By Denny Vrandečić and
Philipp Sorg, AIFB, 2011

Figure 3: For comparison with the Simple English Wikipedia corpus, this screenshot of Corpex displays the frequencies for the 't' over the Brown corpus.