Turkish Journal of Computer and Mathematics Education Vol.12 No.13(2021), 2465-2472

Research Article

Identification of Languages from The Text Document Using Natural Language Processing System

Manjula S[1], Dr. Shivamurthaiah M[2]

[1]Research Scholar, Department of Computer Science, Garden City University, Bangalore Karnataka
[2]Research Guide, Department of Computer Science, Garden City University, Bangalore Karnataka

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 4 June 2021

Abstract

One of the fundamental and significant tasks of data interpretation is language detection from textual data. The current effort is to detect the 22 distinct languages in a multilingual document using the Hybrid Isomap technique. Language identification research is becoming increasingly relevant in everyday life. Language identification tasks performed using the "European Parliament Proceedings Parallel Corpus 1996-2011." The corpus is a vast and systematic collection of machine treadle writings generated in a natural communication situation. This corpus is derived from the proceedings of the European Parliament, and it usually involves 21 European languages. The Natural Language Processing approach will facilitate in identifying the many languages included in the text document.

Key Words: European, Detection, Textual Data, Language

1. Introduction

Text corpora, or compilations of texts, have long piqued the curiosity of linguists. They were used by lexicographers to compile dictionaries, as well as linguists and historians to study language development throughout time. The phrase parallel corpus refers to increase the understanding, not compilations of writings that are just contentedly related to one another. Both are referred to as similar corpora because they represent issues linearly without the need for texts to be translations of each other (Plamada and Volk 2013).

Comprehensive dictionary often includes phase ” for various word senses, that frequently drawn from corpora. An approach for picking those sentences, known as excellent vocabulary instances (GDEX) (Kilgarriff, Husák, et al. 2008), scores similar paragraphs based on characteristics such as sentencing guidelines and the rarity of the terms included. Nevertheless, specialized equipment is required to evaluate each dictionary candidate and eliminate undesirable ones. In computer-assisted language acquisition systems, good example phrases are also important. A few of these help their customers, who are nonnative speakers, by offering use instances for a certain word or expression, then utilized in dynamically created lessons (Volodina et al. 2012; Pilan et al. 2016). In terms of gold standards of annotation, efforts are taken to enable comparison, which often entails doing the same annotation job by numerous validators and comparing the findings (Junczys- Dowmunt et al. 2016). The inter-annotator concordance (Artstein 2017) reflects that how people do in a given activity.

1.1 Parallel Corpora

The origin of linguistic study on parallel writings traced back to archaeological finds that included bilingual or trilingual inscriptions, such as the Rosetta Stone is a stele etched in three separate ancient languages with roughly the same text (Cysouw and Walchli 2007; Ziemski et al. 2016). The stone was unearthed at a period when all two languages had been dead for more than just a millennium, researchers were able to obtain an understanding of the

other two through their knowledge of Ancient Greek (Ostling 2015). More lately, worldwide alliances, either social, governmental, or commercial, have resulted in a vast set of interconnected sources, ranging from multilingual to multiple languages (Rafalovitch and Dale et al. 2009). Legal documents, recorded speech, training materials, medicine packaging inserts, and translation novels such as the Scripture, curricula, and fiction are examples of these resources (Eisele and Y. Chen 2010). Most of these tools have previously been used by linguistic applications, including word meaning disambiguation (Konig and Lezius 2000; Kazakov and Shahid 2013), establishing grammatical translating norms across language (Lavie et al. 2008), and computer lexicology (Tiedemann 2003; Volk et al., 2014; Volk et al., 2007).

1.2 Text Alignment

The process of determining minimum matching text elements in two languages in a set of translated is known as text alignment. Text alignment, despite phrase and phrase orientation, is heavily influenced by extra-linguistic factors including the topic, provenance, and technical structuring of the text data. If all character encodings are a close translation of one other, even sentences are seen as the smallest text AUs. For raw textbook translation, the evident organization often consists of chapters, and parallel texts will be proportionally huge. Text contact, on the other hand, does not have to be provided right away. "The initial alignment challenge while developing parallel corpora is to link related documents with everyone," writes Tiedemann (2009). He refers to this task as document alignment and document linkage (Gohring and Volk 2011).

2. Dataset

The world is communicating with each other with the help of different types of languages. This communication is either through speaking or document-level communication. In the official working area, communication will take place with the help of a document that is known as a text document.

The size of the corpus is shown in below Table .1

Table 1. Size of the corpus

Language	Sentences	Words		Language	Sentences	Words
Bulgarian	411,636	-		Italian	2,081,669	50,259,169
Czech	668,595	13,195,311		Lithuanian	678,665	11,512,131
Danish	2,323,099	47,761,381		Latvian	666,026	12,085,228
German	2,176,537	47,236,849		Dutch	2,333,816	53,487,257
Greek	1,517,141	-		Polish	387,490	7,087,016
English	2,218,201	53,974,751		Portuguese	2,121,889	52,300,149
Spanish	2,123,835	54,806,927		Romanian	402,904	9,663,544
Estonian	692,210	11,358,009		Slovak	674,359	13,116,301
Finnish	2,119,515	33,708,706		Slovene	634,488	12,665,974
French	2,190,579	54,202,850		Swedish	2,241,386	45,665,947
Hungarian	658,824	12,606,986				

Table 2. Sizes of parallel corpora with word alignment and XML removal.

Parallel Corpus (L1-L2)	Sentences	L1 Words	English Words		Parallel Corpus (L1-L2)	Sentences	L1 Words	English Words
Bulgarian-English	406,934	-	9,886,291		Italian-English	1,909,115	47,402,927	49,666,692
Czech-English	646,605	12,999,455	15,625,264		Lithuanian-English	635,146	11,294,690	15,341,983
Danish-English	1,968,800	44,654,417	48,574,988		Latvian-English	637,599	11,928,716	15,411,980
German-English	1,920,209	44,548,491	47,818,827		Dutch-English	1,997,775	50,602,994	49,469,373
Greek-English	1,235,976	-	31,929,703		Polish-English	632,565	12,815,544	15,268,824
Spanish-English	1,965,734	51,575,748	49,093,806		Portuguese-English	1,960,407	49,147,826	49,216,896
Estonian-English	651,746	11,214,221	15,685,733		Romanian-English	399,375	9,628,010	9,710,331
Finnish-English	1,924,942	32,266,343	47,460,063		Slovak-English	640,715	12,942,434	15,442,233
French-English	2,007,723	51,388,643	50,196,035		Slovene-English	623,490	12,525,644	15,021,497
Hungarian-English	624,934	12,420,276	15,096,358		Swedish-English	1,862,234	41,508,712	45,703,795

Here parallel corpus can be defined as it is also a corpus that contains a collection of the original text in language L1 and their translations into a set of languages L2. Table 2 shows Sizes of parallel corpora with word alignment and XML removal.

3. Methodology

To perform the language identification process in this paperwork this corpus is considered as input values. The number of steps such as data acquisition, data preprocessing, tokenization, feature extraction, and classification; are the different steps that have been implemented in language identification processing.

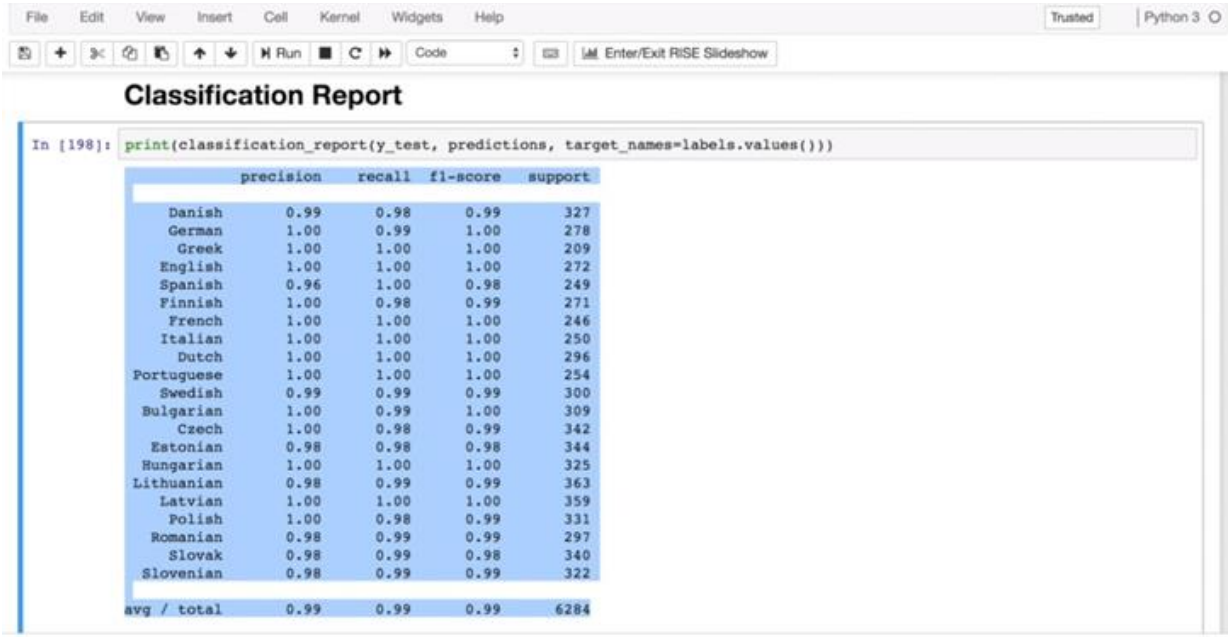
1. **Data Acquisition:** It is nothing but the collection of input data for language identification. In this research work **"European Parliament Proceedings Parallel Corpus 1996-2011"** is considered as input values. This document is an unstructured data format that has all types of symbols, punctuation marks, stopping words, articles, and so on. This is converted into structure format by using XML methodology.
2. **Data Preprocessing:** The preprocessing is defined as enhancing the quality of raw data into a standard format so that it can be further used for research work. During this enhancement process the noise present in the document has been removed, so in this text document noises are unnecessary spaces or white spaces, stopping words, articles, symbols, punctuation marks so these things will be removed from the input document so that it can further used for the next step.
3. **Tokenization:** It the process of segmenting sentences into individual words. While performing segmentation activity the first paragraph will undergo segmentation and it generated line by line segmentation result. Each line will be considered as an input for segmentation which performs word by word segmentation. Each word is considered as a Token and this complete process is known as Tokenization. Once this process is completed then each word will be considered as an input for the further process.
4. **Feature Extraction:** To classify different languages it needs feature values. The major requirement of this feature extraction is to get unique information from each word and this information can be further used for

language classification. Tokens are considered input values to perform the feature extraction process. The complete tokens are divided into two different sets such as training set and testing or evaluation set. Here we are using a training set for assessment and implementation purposes and a testing set will be used for evaluation and these two sets are independent of each other. A Hybrid Isomap algorithm has been implemented on training set data and feature values are extracted. These values are represented in a matrix format where rows of the matrix represent data items and columns of the matrix represent unique feature values. Further for each column calculate the mean value and subtract with each arrival, the attributes with higher variance are considered as more important than the lower variance. If the importance of the attribute is autonomous of the variance of that attribute then divide each such type of value with a standard deviation of that each column of the resultant feature extracted matrix. After this calculate the covariance of the standard deviation value. this can be done by taking reversing the standard deviation value and then multiply the reversed matrix of standard deviation with an original matrix of standard deviation. The numerical depiction is covariance of S = S'.S (where S is considered as standard deviation matrix). Once we get the covariance value then calculate the eigenvector and eigenvalue on S. Then categorize the Eigenvector in descending order. These categorized values are further used to calculate the new feature that is the quality grade version of the training set. Each inspection is a blend of authentic variables, where the weights are determined by the eigenvectors are autonomous of one another and each column of standard deviation is also autonomous of one another. In the last drop unimportant feature from the absolute resultant matrix set.

5. **Classification:** The extracted feature extraction values are considered as an input for the classification process. For this classification process SVM and Bayesian classification has been implemented. The classification rated for SVM classifier is 98.2% and Bayesian classification is 99.1% accuracy for classification of 21 different languages.

4. Experimental Results

The proposed methodology is implemented on "European Parliament Proceedings Parallel Corpus 1996-2011". The proposed operations are implemented by using Python Language on Jupiter Platform and results are shown in Figures 1 and 2.



	precision	recall	f1-score	support
Danish	0.99	0.98	0.99	327
German	1.00	0.99	1.00	278
Greek	1.00	1.00	1.00	209
English	1.00	1.00	1.00	272
Spanish	0.96	1.00	0.98	249
Finnish	1.00	0.98	0.99	271
French	1.00	1.00	1.00	246
Italian	1.00	1.00	1.00	250
Dutch	1.00	1.00	1.00	296
Portuguese	1.00	1.00	1.00	254
Swedish	0.99	0.99	0.99	300
Bulgarian	1.00	0.99	1.00	309
Czech	1.00	0.98	0.99	342
Estonian	0.98	0.98	0.98	344
Hungarian	1.00	1.00	1.00	325
Lithuanian	0.98	0.99	0.99	363
Latvian	1.00	1.00	1.00	359
Polish	1.00	0.98	0.99	331
Romanian	0.98	0.99	0.99	297
Slovak	0.98	0.99	0.98	340
Slovenian	0.98	0.99	0.99	322
avg / total	0.99	0.99	0.99	6284

Figure 1: The classification result along with the average accuracy rate.

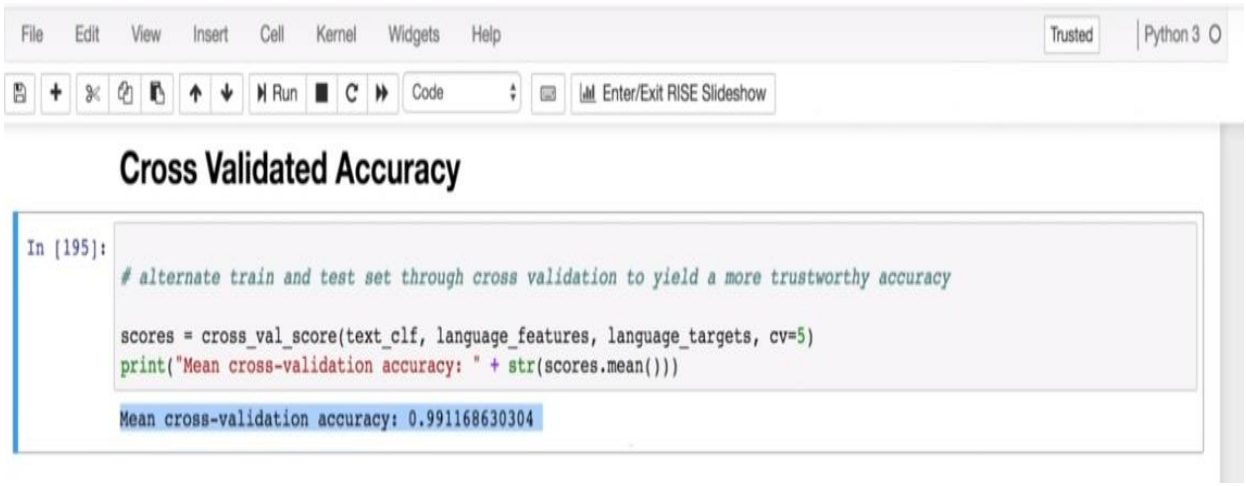


Figure 2: The mean accuracy value as 99.116%.

4.1 Input Text for Tokenization

```
text = "Natural language processing (NLP) is a field " + \
       "of computer science, artificial intelligence " + \
       "and computational linguistics concerned with " + \
       "the interactions between computers and human " + \
       "(natural) languages, and, in particular, " + \
       "concerned with programming computers to " + \
       "fruitfully process large natural language " + \
       "corpora. Challenges in natural language " + \
       "processing frequently involve natural " + \
       "language understanding, natural language" + \
       "generation frequently from formal, machine" + \
       "-readable logical forms), connecting language " + \
       "and machine perception, managing human-" + \
       "computer dialog systems, or some combination " + \
       "thereof."
```

After the implementation of the tokenization method

4.2 The output of the tokenization process

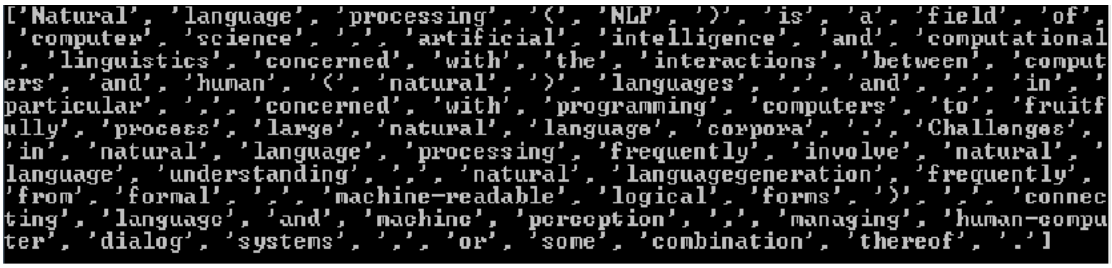
Tokenization on either a paragraph basis. Some tokenizers for splitting a paragraph into sentences are listed below. The results returned from each may change slightly, therefore you should select a suitable tokenizer that will perform best. Sent tokenize is a wrapping method for the Punkt Phrase Tokenizer's tokenize method. This tokenizer converts a text into a list of sentences by employing an unsupervised approach to train a model for abbreviated words, words, and phrases that begin lines. This takes us to the conclusion of this post, wherein we learned about tokenization and numerous implementation methods.

```
['Natural language processing (NLP) is a field of computer science, artificial i
ntelligence and computational linguistics concerned with the interactions betwee
n computers and human (natural) languages, and, in particular, concerned with pr
ogramming computers to fruitfully process large natural language corpora.', 'Cha
llenges in natural language processing frequently involve natural language under
standing, natural languagegeneration frequently from formal, machine-readable lo
gical forms), connecting language and machine perception, managing human-compute
r dialog systems, or some combination thereof.']
```

4.3 Word by word tokenization

Such characters are usually referred to as terms or phrases, however, it is occasionally necessary to distinguish between types and tokens. A token is a specific instance of a series of letters in a text that are grouped as a meaningful semantic system for processing. A category is indeed the class that all tokens with much the same

feature vector belong to. A term is a (possibly normalized) type that's also contained in the dictionary of the IR system. The collection of indexing words might be completely independent of the tokens, for example, semantic identifiers in a taxonomy, however in fact, in current IR systems, these were closely tied to tokens throughout the text.



5. Conclusion

In 22 languages, we created a corpus containing text, phrase, and word alignments. Restoration of flaws in the initial corpus tokenization, sentence segmentation, element labeling, lemmatization, syntactic dependency parsing, and matching on different structural stages are among the preliminary stages. Text categorization is an essential problem in language identification; because each country has a variety of languages, it is critical to recognize languages. Diverse sorts of social media use a variety of languages on the web. This complicates distinguishing different types throughout the country. This article paper discusses several feature extraction methodologies as well as a unique hybrid method for recognizing various languages in the nation. Language is categorized into classes based on the extracted features.

References

1. Artstein, R. (2017). “Inter-annotator Agreement”. In: Handbook of Linguistic Annotation. Ed. by N. Ide and J. Pustejovsky. Springer, pp. 297–313.
2. Cysouw, M. and B. Wälchli (2007). “Parallel Texts: Using Translational Equivalents in Linguistic Typology”. In: Sprachtypologie & niversalienforschung (STUF) 60 (2).
3. Eisele, A., and Y. Chen (2010). “MultiUN: A Multilingual Corpus from United Nation Documents.” In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC). (Valletta). European Language Resources Association (ELRA), pp. 2868–2872.
4. Gohring, A. and M. Volk (2011). “The Text+Berg Corpus An Alpine French- German Parallel Resource”. In: Traitement Automatique des Langues Naturelles, pp. 63–68.
5. Junczys-Dowmunt, M., B. Pouliquen and C. Mazenc (2016). “COPPA V2.0: Corpus Of Parallel Patent Applications. Building Large Parallel Corpora with GNU Make”. In: Proceedings of the 4th Workshop on Challenges in the Management of Large Corpora (CMLC). (Portorož). Ed. by P. Bański, M. Kupietz et al. Portorož, Slovenia.
6. Kazakov, D. and A. R. Shahid (2013). “Using Parallel Corpora for Word Sense Disambiguation”. In: Proceedings of Recent Advances in Natural Language Processing (RANLP), pp. 336–341.
7. Kilgarriff, A., M. Husák, K. McAdam, M. Rundell and P. Rychlỳ (2008). “GDEX: Automatically Finding Good Dictionary Examples in a Corpus”. In: Proceedings of the 13th EURALEX International Congress. Ed. by J. D. Elisenda Bernal. Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra.

8. Konig, E. and W. Lezius (2000). “A Description Language for Syntactically Annotated Corpora”. In: Proceedings of the 18th Conference on Computational Linguistics. Vol. 2. Association for Computational Linguistics (ACL), pp. 1056–1060.
9. Lavie, A., A. Parlikar and V. Ambati (2008). “Syntax-driven learning of subsentential translation equivalents and translation rules from parsed parallel corpora”. In: Proceedings of the 2nd Workshop on Syntax and Structure in Statistical Translation. (Ohio). Association for Computational Linguistics (ACL), pp. 87–95.
10. Ostling, R. (2012). “Stagger: A modern POS tagger for Swedish”. In: Proceedings of the 4th Swedish Language Technology Conference (SLTC).
11. Pilan, I., E. Volodina and L. Borin (2016). “Candidate sentence selection for language learning exercises: from a comprehensive framework to an empirical evaluation”. In: Traitement Automatique des Langues 57.3, pp. 67–91.
12. Plamada, M. and M. Volk (Aug. 2013). “Mining for Domain-specific Parallel Text from Wikipedia”. In: Proceedings of the 6th Workshop on Building and Using Comparable Corpora (BUCC). (Sofia). Association for Computational Linguistics (ACL), pp. 112–120.
13. Rafalovitch, A., R. Dale et al. (2009). “United Nations General Assembly Resolutions: A Six-Language Parallel Corpus”. In: Proceedings of the Machine Translation Summit. Vol. 12, pp. 292–299.
14. Tiedemann, J. and G. Kotze (2009). “A Discriminative Approach to Tree Alignment”. In: Proceedings of the Workshop on Natural Language Processing Methods and Corpora in Translation, Lexicography, and Language Learning. Association for Computational Linguistics (ACL), pp. 33–39.
15. Tiedemann, J. and G. Kotze (2009). “A Discriminative Approach to Tree Alignment”.(2009b). “Building a Large Machine-Aligned Parallel Treebank”. In: Proceedings of the 8th International Workshop on Treebanks and Linguistic Theories (TLT), pp. 197–208.
16. Volk, M., J. Lundborg and M. Mettler (2007). “A Search Tool for Parallel Treebanks”. In: Proceedings of the Linguistic Annotation Workshop (LAW). Prague: Association for Computational Linguistics (ACL), pp. 85–92.
17. Volk, M., S. Clematide, J. Graën and P. Ströbel (2016). “Bi-particle Adverbs, PoS-Tagging and the Recognition of German Separable Prefix Verbs”. In: Proceedings of the Conference on Natural Language Processing (KONVENS).
18. Volodina, E., R. Johansson and S. J. Kokkinakis (2012). "Semi-automatic selection of best corpus examples for Swedish: Initial algorithm evaluation". In: Proceedings of the Workshop on NLP for Computer-Assisted Language Learning. 080. Linköping University Electronic Press, pp. 59–70.
19. Ziemski, M., M. Junczys-Dowmunt and B. Pouliquen (2016). “The United Nations Parallel Corpus v1.0”. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC). Portorož, Slovenia.
20. Download corpus link
21. source release (text files), 1.5 GB

22. tools (preprocessing tools and sentence aligner only), 8.6 KB
23. parallel corpus Bulgarian-English, 41 MB, 01/2007-11/2011
24. parallel corpus Czech-English, 60 MB, 01/2007-11/2011
25. parallel corpus Danish-English, 179 MB, 04/1996-11/2011
26. parallel corpus German-English, 189 MB, 04/1996-11/2011
27. parallel corpus Greek-English, 145 MB, 04/1996-11/2011
28. parallel corpus Spanish-English, 187 MB, 04/1996-11/2011
29. parallel corpus Estonian-English, 57 MB, 01/2007-11/2011
30. parallel corpus Finnish-English, 179 MB, 01/1997-11/2011
31. parallel corpus French-English, 194 MB, 04/1996-11/2011
32. parallel corpus Hungarian-English, 59 MB, 01/2007-11/2011
33. parallel corpus Italian-English, 188 MB, 04/1996-11/2011
34. parallel corpus Lithuanian-English, 57 MB, 01/2007-11/2011
35. parallel corpus Latvian-English, 57 MB, 01/2007-11/2011
36. parallel corpus Dutch-English, 190 MB, 04/1996-11/2011
37. parallel corpus Polish-English, 59 MB, 01/2007-11/2011
38. parallel corpus Portuguese-English, 189 MB, 04/1996-11/2011
39. parallel corpus Romanian-English, 37 MB, 01/2007-11/2011
40. parallel corpus Slovak-English, 59 MB, 01/2007-11/2011
41. parallel corpus Slovene-English, 54 MB, 01/2007-11/2011
42. parallel corpus Swedish-English, 171 MB, 01/1997-11/2011