Measuring the distance between comparable corpora between languages

Serge Sharoff

Centre for Translation Studies
School of Modern Languages
University of Leeds
LS2 9JT, Leeds, UK
s.sharoff@leeds.ac.uk

Abstract. The notion of comparable corpora rests on our ability to assess the difference between corpora which are claimed to be comparable, but this activity is still art rather than proper science. Here I will discuss attempts at approximating the content of corpora collected from the Web using various methods, also in comparison to traditional corpora, such as the BNC. The procedure for estimating the corpus composition is based on selecting keywords, followed by hard clustering or by building topic models. This can apply to corpora within the same language, e.g., the BNC against ukWac as well as to corpora in different languages, e.g., webpages collected using the same procedure for English and Russian.

1 Introduction

The British National Corpus (BNC) has been collected in the beginning of the 1990s using various written and spoken sources from the 1970s-1980s. It aims at representing modern English, but in many respects it is outdated and it does not cover many domains. The web is huge, it is easy to collect data from it either by using search engine queries [3,16] or crawling respective websites [9]. However, once we have a corpus, we still do not know its composition, e.g., the proportion of webpages for medical doctors and patients in the corpus of [3], or to what extent the BNC is similar to a web corpus.

The problem of not knowing the content gets another dimension when we use comparable corpora, i.e., two corpora which are claimed to be similar in one aspect or the other. It is possible that comparable corpora are collected in their own ways and are drawn from different distributions, e.g., from different website or for different languages. For example, within the TTC project [5] our aim is to explore the possibility of mining terminological resources from specialised comparable corpora. For this task, we used parallel seeds to collect corpora on comparable topics by retrieving webpages returned in response to queries containing identical or nearly identical terms in several languages (below shown for English, Chinese and Russian):

fossil fuel	化石燃料	ископаемое топливо
power station	发电厂	электростанция
hydroelectricity	水力发电	гидроэнергетика
photovoltaics	太阳能光伏	фотоэлектричество

...

However, the crucial question is: do we get comparable pages by sending comparable queries? One approach to comparing corpora across languages is by translating the features obtained from the documents, usually such features are the keywords [1]. Li and Gaussier produced a metric for assessing the comparability of multilingual collections [13], see also their paper in this volume. Their similarity measure is based on the proportion of the words for which a translation can be found in the opposite part of a comparable corpus. Their approach has been tested on a parallel corpus (Europarl) with added noise from other text types to decrease the degree of parallelism. Babych and Hartley also used parallel corpora from the TAUS Data Association[1] to determine the distance by measuring the difference in the frequencies of the translations of the top 500 words for each corpus [2].

The study reported in this paper is based on the same idea of a lexicon overlap measure via a dictionary, while the difference lies in (1) using the unsupervised methods for keyword selection, (2) clustering and topic modeling to assess and compare corpora with unknown composition, and (3) estimating the contents of really comparable rather than parallel corpora in realistic settings. The main contribution is that the procedure gives an answer to the question how similar large webcorpora such as ukWac or itWac [4] are to each other and to the BNC, as well as to the question about comparable corpora collected by using comparable queries, such as listed above.

2 Methodology

2.1 Corpora used for analysis

The BNC classifies its documents using a complex classification scheme [12], which includes such categories as

- domains (eight labels in total: natsci, appsci, socsci, belief, imaginative, leisure, business, world affairs);
- genre (seventy labels in total, W.advert, W.newsp.brdsheet.national.arts, etc)
- information about the audience, author, publication medium, etc.

In spite of the complexity of the classification system, it does not cover many substantial differences between the BNC texts, e.g., a text from socsci can be from the subdomain of linguistics or history, the domain of world affairs covers both local British and international politics. The composition of the BNC can be compared to ukWac and itWac [4], large corpora crawled from the .uk and .it domains to represent respectively British English and Italian (with subsequent

[1] http://www.tausdata.org/

language filtering). Another task is to compare the composition of specialised corpora collected using comparable seeds (Table 1).

In a somewhat similar study [17] the composition of I-En, I-Ru and ukWac was analysed in terms of their genres. Two separate sets of the genre categories of the BNC and RNC (Russian National Corpus) were mapped to metagenre labels, such as instruction or regulation, and these corpora were used to train supervised machine learning models for genre classification (using part-of-speech trigrams). The resulting classifiers were applied to the Web corpora. That study did show the similarities and differences between the metagenres of traditional corpora and those collected from the Web, but it was limited to a small set of eight metagenres. The composition of the topics on the Web remains to be investigated.

2.2 Keyword selection

The procedure for selecting the keywords per each document was based on the Log-likelihood (LL) score [15]. Like the commonly used tf*idf score the procedure is language-independent, but unlike tf*idf it takes into account the relative frequency of a term in a document against the reference corpus as well as the absolute number of its occurrences as the evidence of its statistical significance. The LL-score also allows setting a straightforward threshold using the value of 10.83 ($p = 0.001$) or 15.13 ($p = 0.0001$), for the justification see [15]. An example of keywords for a random webpage is shown below:[2]

LL-score	N	Keyword
126.07	7	Womble
101.47	9	neural
91.26	6	elegans
62.55	13	model
47.22	6	simulation
46.47	10	network
39.59	3	locomotion
36.66	3	biologically
26.71	3	constrain
21.95	3	nervous
21.78	3	cognitive
	...	

They are all useful for describing this individual webpage, but some of them (*Womble, elegans*) are too specific for the purpose of clustering webpages of a general-purpose corpus: *Womble* is a keyword for only 22 pages in ukWac, so it is less useful for finding its clusters. Also a corpus of the size of ukWac generates 89,394 unique keywords for clustering its 2,541,926 documents, causing dimensionality problems. Restricting the keyword lexicon down to the most common words[3] limits the ability to select pages with specific topics, so a simple approach used in this study was to reduce the amount of keywords down to the

[2] http://www.comp.leeds.ac.uk/biosystems/neuroscience.shtml from the time it was collected for ukWac.

[3] As done, for example, in [10].

Table 1. Corpora used in this study (sizes given in tokens and documents)

Corpus	BNC	ukWac	itWac	EN-Energy	RU-Energy	ZH-Energy
Tokens	111246939	2119891296	179512658	7505765	7766462	12431752
Documents	4054	2541926	175646	5762	5126	3287

10,000 words most often selected as keywords in the entire corpus. The use of the complete set of keywords tested on the BNC was marginally detrimental to the results. The clusters remained essentially the same, so a smaller keyword set was used in all experiments reported below.

2.3 Clustering methods

For unsupervised detection of domains two scenarios were used, hard clustering and generation of topic models. Because of the need to cluster a fairly large number of webpages, the repeated bisection (RB) algorithm was used in Cluto [20], as it is quite efficient and tends to produce clearly interpretable results, cf also a study in [19]. A cluster is treated as a subcorpus and described in terms of its keywords against the entire corpus using the same LL-score as for obtaining the keywords of individual documents.

Generation of topic models is based on Latent Dirichlet Allocation (LDA), which estimates the distribution of probabilities of keywords belonging to different topics as well as the proportions of documents over the same set of topics (as in traditional clustering, the number of topics is set at the beginning of the experiment). This comes from an unsupervised procedure, in which the unknown distributions are derived in repeated Bayesian approximations from the distribution of hidden variables (like in Hidden Markov models). This setup helps in generalising inherent similarities between the keywords, see [6]. For each topic we get the degree of its association with documents and keywords. The advantage of topic models in comparison to RB clustering is that each document gets a proportion of its association with each topic, so that it can belong to several topics at once. However, this does not allow us to estimate the partitioning of the entire corpus into a set of identifiable segments in order to compare the relative size of each topic.

It is known from prior experience that there are more topics in the BNC than the set indicated in its classification scheme, so more than eight clusters need to be used. In all experiments with general-purpose corpora the option of 20 clusters was used. This number helped in producing interpretable models covering a diverse set of topics, while a larger number of clusters makes the comparison task inherently difficult. As discussed in [7], there is no correlation between human perception of the consistency of clusters and automatic measures (such as perplexity or I_2), so it is difficult to estimate the right amount of clusters automatically, while in the context of this study there was no scope for a proper human evaluation of the quality of each clustering solution for each corpus.

CLUTO [20] uses I_2 as the standard objective function, which maximises the cosine similarity between each node and the centroid of the clusters it belongs to. I_2 on the BNC grew slowly when the number of clusters varied from 10 to 27 (reflecting the reduced size of each cluster), so no definite estimate towards the desired number of clusters can be made:

N	I_2	**N**	I_2	**N**	I_2	**N**	I_2	**N**	I_2	**N**	I_2
10	1.38	13	1.44	16	1.49	19	1.53	22	1.57	25	1.60
11	1.40	14	1.46	17	1.50	20	1.55	23	1.58	26	1.61
12	1.42	15	1.47	18	1.52	21	1.56	24	1.59	27	1.62

2.4 Cross-linguistic comparison

The procedure described above generates the clusters (or topics) describing the composition of a corpus in a given language. The next step is to compare allegedly comparable corpora across languages. For this we need to match the clusters, either qualitatively by comparing their descriptive features or quantitatively by matching the translations of their features. In a similar study [2] the documents from the entire non-English parts of comparable corpora were translated into English using the Google MT server. For large corpora of the size of ukWac or I-RU, this approach is not feasible. So, the features need to be translated using a dictionary. However, there are also problems in using traditional bilingual dictionaries. First, comprehensive dictionaries are not available for many language pairs. Second, traditional dictionaries do not indicate the probabilities for choosing the translation equivalents. Finally, the procedure for mapping the features (keywords) of one languages into another one runs into problems in cases of natural variation in using terminology, so that some terms used in one document do not match its ostensibly parallel counterpart (e.g., very similar issues are discussed under the labels of *copyright* or *IPR*), or when there is a difference in the actual contents of the corpora in two languages, while the two corpora are close thematically (e.g., the recipes in English and Italian).

The first two problems can be mitigated by using statistical dictionaries. The dictionaries used in this study were extracted from parallel corpora by Giza++ [14], with the MultiUN corpus [8] used for the English-Russian pair (230 MW, 8.2 million sentences), and Europarl [11] for the English-Italian pair (47 MW, 1.6 million sentences). The corpora have been lemmatised before extracting the equivalents in order to improve the chances of matching the variety of forms for morphologically rich languages. This results in a sparse translation matrix Tr, which for a word lists its equivalents with their translation probabilities:

Tr(**лопасть**) blade:0.778, vane:0.107, rotor:0.052

Tr(**возможность**) opportunity:0.290, possibility:0.100, capacity:0.068, possible:0.050, ability:0.023, able:0.020 [4]

[4] The Oxford Russian Dictionary, for example, translates возможность by *possibility, opportunity, means, resources*, which does not cover the full range of probable translation equivalents and does not provide any frequency estimates.

The last problem can be mitigated by using the clusters, which provide a broader gist of the topics addressed and then mapping the gists rather than individual documents. The gist can be generated by either clustering or producing topic models. Even though the gists are based on the keywords and they are compared by translating them, their advantage over keyword mapping comes from the possibility of generalising the range of expressions used in individual text over the proximity to a more general field (such as property rights or food items). Formally speaking, the distance can be measured by the cosine similarity of the original English vectors E_i with the original feature vectors from other languages F_j which have been translated into English and weighted with the translation probabilities

$$\cos(E_i, F_j^t) = \frac{E_i \cdot F_j^t}{\|E_i\| \|F_j^t\|}$$

where F_j^t have been produced as

$$F_j^t = Tr \times F_j$$

3 Results

Even though corpus analysis via clustering is more important for web-derived corpora, because we do not know their exact composition, it makes sense to evaluate our methods on a corpus we know better, such as the BNC (Section 3.1). Then, in Section 3.2 I will compare I-RU and itWac, two general-purpose Internet corpora for Russian and Italian to ukWac [16,4]. Finally, I will present a study of specialised comparable corpora across different languages (Section 3.3).

3.1 Clusters in the BNC

Table 2 lists the clusters and topic models with their keywords. The clusters produced by Cluto are numbered according to their internal consistency (the smallest number for the greatest consistency). Also, at the bottom of the list of clusters I indicate the most frequent genre categories from the BNC associated with this genre. In case the genre is not informative (W.misc), the most common domain category is given after a slash. Obviously not all cluster members share the same code, but the pattern is consistent. The following is the distribution of the BNC genre codes of documents belonging to Cluster 18 in Table 2:

N	BNC codes	N	BNC codes
424	W.fict.prose	5	W.letters.personal
49	W.misc	5	S.brdcast.discussn
43	S.oral.history	4	W.religion
29	W.fict.poetry	4	W.non.ac.medicine
26	W.biography	4	S.speech.unscripted
19	W.non.ac.soc	4	S.classroom
5	W.pop.lore	3	W.non.ac.humanities
5	W.essay.school	3	W.newsp.brdsht.social

The documents outside of W.fict.prose are still reasonably similar, as they include ADG (a book for teenage girls), ADM (accounts of travelling through Ireland), AP7 (experiences of old age), etc.[5]

The LDA procedure estimates the distribution of the topics for each document, so it is not possible to obtain information about the relative size of the topics and their constituent documents. In the end, they cannot be directly matched to the labels, thus the labels in Table 3 are indicative, selected on the basis of their keywords.

The two clustering methods agreed on a number of domains, such as fiction, local British and international affairs. However, for some domains (underlined in Table 2) the clusters and topic models disagreed. One considerable difference stems from the ability of topic models to differentiate between the everyday language and professional discourse (the topics labelled as business and health care). In addition to this, hard clustering did not produce clusters related to history and research, which constitute relatively small (but consistent) topics in the BNC, e.g., documents HHY, HJ0, HPN, etc, containing research project applications and texts on project management. Also, LDA generated an inconsistent topic, which combined references to sport clubs with other local British texts, possibly generalising on the membership of their keywords to place names in the UK.

3.2 Comparing Internet corpora to the BNC

The same procedure was applied to ukWac and generated the clusters and topics reported in Table 4. The clusters offer a quick way into comparing the composition of ukWac to that of the BNC. For instance, there is a cluster without specific keywords (Cluster 0 in ukWac). This closely corresponds to the fiction cluster in the BNC (Cluster 18 in Table 2). However, unlike the BNC the ukWac cluster contains little fiction and mostly consists of diary-like blogs, tabloid columns on everyday topics and chats.

Domains that reasonably match both the BNC and ukWac include political news, sports, health, business and religion. The differences concern a broader collection of topics in ukWac within each domain, as well as the presence of more modern texts. For example, clustering of ukWac detects two clusters with keywords related to computing, one of which (Cluster 7) is quite similar to Cluster 0 in the BNC. Another computing cluster of ukWac (15) belongs to the field of web-based communications (the longer list of its keywords also includes *HTML, Internet, click, Google*, etc).

RB clustering in Cluto seems to be considerably more efficient on a large corpus than LDA. Clustering of the entire ukWac took 1494 sec, while LDA was not able to deal with a selection of more than 500,000 documents from ukWac (on a computer with 4GB memory), and producing topic models even for this subset took 4896 sec (clustering of the same subset with Cluto took 304 sec).

[5] The ids are from the BNC index [12].

Even though the clusters and the topic models in Table 4 are drawn from two slightly distributions, there is a broad similarity between their results. The clusters of music, movies, health, computing, communication, food and gardening, religion, politics, travel and business have compatible keywords.

In the case of topic modeling, it is also possible to estimate the distance of each of the topics generated for ukWac to the BNC topics in order to estimate the relative distance between the topics of a web corpus against an established reference one. The following is the list of topics in ukWac with their closest match in the BNC, sorted by the distance score:

Label	**ukWac**	**BNC**	cos
Legal	1	14	0.746
Politics	4	19	0.740
Locations	7	7	0.682
Diary/fict	0	16	0.599
Religion	9	9	0.598
	...		
Academic	13	18	0.201
Sports	3	7	0.179
Music	12	16	0.157
Online	18	18	0.115
Forum	19	19	0.061

The match between the majority of the BNC topics and ukWac is reassuring for the prospects of using ukWac as a reference corpus for linguistic research using larger amount of more modern data. The two most dissimilar topics at the bottom (Online and Forum) reflect the genre changes which came with the Web. The changes in the musical cluster in ukWac against those in the BNC reflect the difference in the topics most discussed in the 1980s (BNC) vs 2000s (ukWac). There is also much greater variation in the amount of musical and sport subdomains covered in ukWac, which explains the divergence from the BNC clusters.

Table 6 lists the clusters generated for itWac, an Italian web-corpus collected using the same compilation procedure as ukWac. It again showed broad compatibility with ukWac. The domains of itWac pages include film (Cluster 0), food (1), blogs and chats (2 and 7), legal texts (3,9), music (4), education (5 and 17), computers (6), religion (8), sports (10), health (11), politics (12 and 13), business (14), culture (15), communication (16), travel (18) and local news (19). Some differences between the keywords in itWac and ukWac are related to the realities in individual countries (*Berlusconi* vs *Labour*). In ukWac there is a Do-It-Yourself cluster (4) and a cluster on NHS (10), while itWac contains two legal clusters (3 and 9), their longer keyword lists reveal that Cluster 9 is more about criminal cases, while Cluster 3 is about commercial and civil legislation.

3.3 Comparing comparable corpora

Energy corpora for English, Chinese and Russian have been collected by making comparable queries. They are considerably smaller than ukWac (less than 10 million words). It was expected that we can find fewer subdomains in them. The I_2 value was again not indicative, so the experiment was set for 15 clusters. As in the previous experiments both clustering and topic models were used. However, on a smaller corpus hard clustering produced less clearly interpretable results (also probably because of the difficulty in clear separation of relatively similar topics), while LDA produced fairly reasonable models listed in Table 7.

The analysis of topic models contradicted the original assessment made on the basis of term lists extracted from the corpora using the standard BootCat procedure [3]:[6]

7467 renewable energy	1508 возобновлять энергия
3127 greenhouse gas	1401 парниковый газ
3049 natural gas	2106 природный газ
2320 solar energy	2274 солнечный энергия
1994 carbon dioxide	1124 углекислый газ
1920 solar cell	2710 солнечный батарея
1529 fuel cell	862 топливный элемент

...

Given the overlap between the terms with similar frequencies and the fact that the corpora were collected with the set of keywords, it was expected that their content is sufficiently similar, and overall the corpora are good candidates for extracting and aligning term lists. However, LDA identified some topics present in both corpora, which are less ideal for term extraction, such as information for investors and news about utility companies, forums (we can expect less consistency in terminology used there) and legal texts. The Russian corpus was also contaminated with documents relating to computers (because of links to *power supply*), general news and student essays, which contain low-level introductions into a range of topics in this field. In general such lists of topics are useful in cleaning the corpus to achieve its greater consistency.

4 Conclusions

Three main outcomes of this study concern:

1. the possibility of rapid unsupervised assessment of the content of large corpora (clusters still need human interpretation, but this can be done quickly as long as they are consistent);
2. the possibility of comparing the content of corpora across languages and improving comparability;
3. the difference between hard clustering and LDA.

With respect to (1), we can use clustering to reveal more information even about a well-annotated corpus, such as the presence of a considerable cluster of texts for railway and aircraft enthusiasts in the BNC (otherwise obscurely

[6] Lemmatisation of term elements in Russian affects the syntactic pattern of the composite term [18].

coded as W.misc or W.leisure). It also shows broad classes of texts in a corpus of unknown composition such as ukWac or itWac.

With respect to (2), the clusters can be used to compare the content of individual corpora, but the interpretation needs to be done manually again. Intersection between the keywords of clusters in two corpora can be also done automatically.

With respect to (3), clustering is a better approach to web-sized corpora, as LDA cannot reasonably handle ukWac and it takes considerably more time on the BNC or itWac. At the same time, on smaller corpora, LDA detects models which are easier to interpret in comparison to clustering solutions.

Acknowledgements

This research was funded by European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement no 248005 (TTC).

References

1. Adafre, S., de Rijke, M.: Finding similar sentences across multiple languages in Wikipedia. In: Proc. 11^{th} EACL. pp. 62–69. Trento (2006)
2. Babych, B., Hartley, A.: Meta-evaluation of comparability metrics using parallel corpora. In: Proc. CICLING 2011 (2011)
3. Baroni, M., Bernardini, S.: Bootcat: Bootstrapping corpora and terms from the web. In: Proc. of LREC2004. Lisbon (2004), http://sslmit.unibo.it/~baroni/publications/lrec2004/bootcat_lrec_2004.pdf
4. Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E.: The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. Language Resources and Evaluation 43(3), 209–226 (2009)
5. Blancafort, H., Daille, B., Gornostay, T., Heid, U., Mechoulam, C., Sharoff, S.: TTC: Terminology extraction, translation tools and comparable corpora. In: Proc. EURALEX2010. Leeuwarden (5-6 July 2010)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
7. Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., Blei, D.M.: Reading tea leaves: How humans interpret topic models. In: Proc. Neural Information Processing Systems (2009)
8. Eisele, A., Chen, Y.: MultiUN: A multilingual corpus from United Nations documents. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). Valletta, Malta (2010), http://www.euromatrixplus.net/multi-un/
9. Joho, H., Sanderson, M.: The SPIRIT collection: an overview of a large web collection. SIGIR Forum 38(2), 57–61 (2004)
10. Kilgarriff, A.: Comparing corpora. International Journal of Corpus Linguistics 6(1), 1–37 (2001)
11. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proc. MT Summit 2005 (2005), http://www.iccs.inf.ed.ac.uk/~pkoehn/publications/europarl-mtsummit05.pdf

12. Lee, D.: Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle. Language Learning and Technology 5(3), 37–72 (2001), `http://llt.msu.edu/vol5num3/pdf/lee.pdf`
13. Li, B., Gaussier, E.: Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In: Proc. 23rd International Conference on Computational Linguistics (Coling 2010). Beijing (2010)
14. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics 29(1), 19–51 (2003)
15. Rayson, P., Berridge, D., Francis, B.: Extending the Cochran rule for the comparison of word frequencies between corpora. In: Proc 7th International Conference on Statistical analysis of textual data (JADT 2004). pp. 926–936. Louvain-la-Neuve (2004)
16. Sharoff, S.: Creating general-purpose corpora using automated search engine queries. In: Baroni, M., Bernardini, S. (eds.) WaCky! Working papers on the Web as Corpus. Gedit, Bologna (2006), http://wackybook.sslmit.unibo.it
17. Sharoff, S.: In the garden and in the jungle: Comparing genres in the BNC and Internet. In: Mehler, A., Sharoff, S., Santini, M. (eds.) Genres on the Web: Computational Models and Empirical Studies, pp. 149–166. Springer, Berlin/New York (2010)
18. Sharoff, S., Kopotev, M., Erjavec, T., Feldman, A., Divjak, D.: Designing and evaluating a Russian tagset. In: Proceedings of the Sixth Language Resources and Evaluation Conference, LREC 2008. Marrakech (2008)
19. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: KDD Workshop on Text Mining (2000)
20. Zhao, Y., Karypis, G.: Empirical and theoretical comparisons of selected criterion functions for document clustering. Machine Learning 55(3), 311–331 (2004)

Table 2. Clusters in the BNC

0	3.3%	Inc, Corp, software, user, Unix, system, lifespan, IBM, module, application, version, package, file
1	1.5%	studio, video-tape, speaker, voice, read, over, report, say, yesterday, male, police, Oxford, Swindon
2	1.7%	pollution, environmental, conservation, nuclear, emission, waste, energy, forest, ozonosphere
3	7.6%	Yeah, oh, get, yeah, well, know, yes, go, No, na, think, there, cos, gon, what, no, just, like, say
4	3.7%	player, win, game, Cup, season, club, team, play, match, goal, championship, score, League, ball
5	1.7%	aircraft, engine, railway, station, car, fly, train, pilot, squadron, locomotive, steam, line, crew
6	2.0%	patient, disease, treatment, study, cell, infection, gastric, health, acid, clinical, concentration, care
7	2.8%	God, Jesus, church, Christian, Christ, faith, king, bishop, prayer, gospel, spirit, pope, Holy
8	3.4%	school, award, teacher, student, education, pupil, course, curriculum, research, study, ref, subject
9	6.6%	okay, right, yeah, get, what, so, yes, if, just, well, go, think, know, actually, mean, there, oh
10	9.9%	government, Minister, party, political, election, Soviet, labour, state, country, president, Prime
11	4.1%	court, case, Act, law, defendant, plaintiff, contract, section, person, appeal, any, solicitor, under
12	4.9%	music, film, guitar, band, play, song, album, Eliot, bass, movie, sound, musical, pop, actor
13	3.5%	cell, gene, DNA, protein, energy, equation, sequence, molecule, temperature, surface, particle
14	6.5%	company, market, rate, price, share, cost, profit, tax, business, firm, UK, investment, bank
15	4.6%	Darlington, local, Council, council, housing, pension, councillor, authority, scheme, service, area
16	5.4%	art, painting, century, artist, exhibition, Edward, building, William, museum, town, king, church
17	5.5%	social, language, theory, word, may, text, behaviour, process, individual, information, meaning
18	15.5%	her, him, me, my, say, look, eye, smile, back, go, feel, door, tell, hand, know, think, like, could
19	5.6%	fish, water, your, plant, knit, bird, food, colour, breed, tank, can, horse, leaf, use, dry, egg, specie

0: W.non.ac.tech.engin; 1: W.news.script; 2: W.misc/W.app.science; 3: S.conv; 4: W.pop.lore/W.newsp.brdsht.nat.sports; 5: W.misc/W.leisure (texts for railway/aircraft enthusiasts); 6: W.ac.medicine; 7: W.religion; 8: W.misc/W.soc.science; 9: S.consult; 10: W.non.ac.polit.law.edu,W.newsp.brdsht.nat.reports; 11: W.ac.polit.law.edu; 12: W.newsp.brdsht.nat.arts,W.pop.lore; 13: W.non.ac.nat.science,W.ac.nat.science; 14: W.commerce; 15: S.meeting; 16: W.non.ac.humanities.arts,W.biography; 17: W.ac.soc.science; 18: W.fict.prose; 19: W.pop.lore,W.instructional/W.leisure

Table 3. Topic models in the BNC

news	police yesterday voice court read john hospital studio darlington speaker street road officer today claim britain crime council male murder charge
sport1	game club team player season match england league goal score race ball minute manager final john football championship sport test champion
fict1	door light smile across foot voice watch moment himself body sound stop dark walk mind black pull wall happen stare catch
pol1	political labour economic class society worker britain organisation population union individual community argue industrial economy process thus period influence production authority
med1	patient care health disease treatment hospital medical cell risk drug cause doctor test blood normal rate infection symptom factor parent cancer
ac	research education award student teacher course staff training project university department subject management pupil skill develop college experience date organisation library
biz1	price rate share bank firm sale profit financial investment account income product industry sell value customer capital rise management benefit director
rail	road building hotel town railway build village mile station north train walk street aircraft engine route centre site travel south park
lang	language sense experience kind human theory particular individual behaviour subject example knowledge english nature understand reason person process text itself suggest
hist	church century john lord christian england death himself english king royal father william jesus edward land roman bishop french amongst
leis1	music film artist record painting exhibition museum sound band american collection gallery song theatre picture press audience john director style television
comp	computer user software file application product corp datum technology module version unix window network machine package database design access available disk
leis2	colour food fish plant design garden light easy cover white flower size range available shape machine knit pattern piece wall grow
spok	yeah actually right hundred okay twenty pound sort alright nice thank fifty eight probably thirty nine anyway half seven thousand remember
legal	court person section order contract legal rule apply require authority appeal decision term claim property shall duty matter action provision application
pol3	council committee authority minister secretary labour community meeting office house national health scheme chairman county agree government scotland proposal matter district
fict2	mother love girl father friend herself remember wife door walk miss mind husband morning stay evening someone speak sister baby dress
envt	animal plant bird specie forest energy land environmental environment produce food million fish grow waste pollution soil scientist nuclear chemical site
med2	cell figure model datum value thus method function contain section sequence type produce structure test process surface solution occur position describe
pol2	minister president soviet national election foreign force political leader international military party union march economic europe april june official prime european

Table 4. Keywords clusters and topic models in ukWac

0	12.2%	I, her, he, she, my, his, me, him, i, it, do, say, PM, post, man, go, have, love, think, but
1	7.5%	road, mile, town, Canal, park, Road, walk, route, fn, Park, Street, Museum, village, north
2	2.5%	game, poker, player, ball, play, Sudoku, puzzle, sudoku, score, Games, win, goal, casino, match, tournament
3	4.6%	school, teacher, education, pupil, learn, skill, learning, child, student, training, teaching, learner
4	5.8%	water, cleared, colour, dive, light, surface, diver, valve, boat, wheel, wall, battery, bike, engine
5	10.0%	sector, organisation, local, business, development, management, sustainable, environmental
6	5.6%	student, University, module, research, study, course, academic, Studies, degree, science, graduate
7	6.3%	datum, system, software, model, use, computer, network, data, technology, user, NUN, solution
8	1.4%	insurance, mortgage, loan, property, Insurance, auto, quot, lender, rate, insurer, Agents
9	2.1%	God, Jesus, Christ, Christian, church, Lord, he, Church, Bible, his, sin, faith, prayer, verse, Him
10	3.1%	patient, care, NHS, health, nurse, hospital, clinical, nursing, mental, medical, service, Trust, Nursing
11	3.5%	plant, fish, bird, garden, flower, tree, food, specie, fruit, wine, seed, vegetable, cook, sauce, soil, sugar
12	3.1%	disease, drug, treatment, cancer, patient, cell, blood, infection, symptom, therapy, animal, gene, risk
13	3.2%	club, race, season, win, League, Cup, team, player, championship, match, lap, football, Championship
14	7.0%	pension, government, Committee, Labour, union, that, political, shall, Minister, vote, member
15	5.4%	file, search, user, text, server, use, Windows, web, page, site, library, directory, Web, browser
16	5.9%	customer, company, your, you, payment, business, sale, product, card, any, charge, market, service, price
17	3.5%	hotel, bedroom, room, holiday, beach, accommodation, apartment, restaurant, bathroom, cottage
18	4.8%	music, song, band, album, guitar, sound, musical, dance, vocal, gig, jazz, bass, track, play, concert
19	2.4%	film, movie, cinema, camera, DVD, comedy, actor, star, Jolen, scene, Hollywood, character

Table 5. Topic models in ukWac

fict	being something never always quite away sure though thought best enough night ever found rather anything left someone fact actually love
legal	shall court section legal person committee whether required office general agreement following authority notice decision apply being appropriate period personal contract
biz	management technology industry market business marketing network customer mobile quality working software manager training media corporate digital client leading commercial security
sports	game club season football race poker match league player sport ball final england best side goal racing playing world half city
pol1	international political british labour party state european united britain europe iraq national trade union states minister world military security american power
med1	care training health young education working social national advice safety trust learning service practice centre community police mental disabled hospital charity
envt	planning government environmental environment management national transport waste review energy sector funding strategy future economic sustainable impact consultation quality housing development
loc	road north water south river park west along species east left bridge station railway town land fish near route village found
comm	insurance credit property money rate financial income bank market price card loan pension cent mortgage value investment account house capital cash
relig	church jesus christ christian lord being human word faith upon love spirit shall bible nature true religious power holy truth death
design	water light power energy surface colour equipment fire control space speed size engine standard model quality front unit type temperature black
travel	hotel room garden house travel food holiday accommodation city centre airport wine parking bedroom kitchen town park property best floor restaurant
music	music film band theatre album sound series dance review story radio song video best festival love performance movie artist stage musical
acad	learning education ac teaching science research language college school academic knowledge social practice assessment degree english department studies analysis international understanding
med2	treatment disease food medical cancer risk clinical blood body patient animal drug health medicine pain hospital human control found skin heart
news	news june july march september october centre april november august meeting january february city december friday conference monday sunday saturday west
hist	john house century church royal king william history england james english thomas museum later george early british robert mary david henry
comp	data file software user computer server text windows program version code type value image screen click select control html object function
online	online website search email click html library internet guide information news download please address content index text resource section info send
forum	php forum id date message post asp view index member joined location author subject page thread from quote

Table 6. Keywords in clusters of itWac

0	2.59%	film, cinema, regista, attore, personaggio, regia, scena, cinematografico, protagonista, pellicola, cast, storia
1	1.07%	vino, olio, cucchiaio, uovo, burro, cuocere, pasta, piatto, latte, zucchero, formaggio, pepe, farina, cucina
2	11.53%	mio, mi, che, suo, non, essere, dire, avere, io, uomo, tuo, fare, lui, quello, ti, ma, occhio, cosa, amore
3	13.99%	articolo, numero, comma, decreto, legge, cui, previsto, lavoro, relativo, presente, servizio, lavoratore
4	2.37%	musica, disco, canzone, album, musicale, brano, band, concerto, rock, suonare, chitarra, musicista, cd
5	3.19%	scuola, scolastico, docente, insegnante, alunno, classe, formativo, didattico, formazione, istruzione, educativo
6	1.71%	file, server, Windows, utente, software, Microsoft, versione, programma, web, browser, utilizzare, Linux
7	7.39%	mi, ciao, io, non, messaggio, ti, mio, avere, fare, se, inviare, ma, me, scrivere, dire, sapere, cosa, tuo, Posted
8	3.86%	chiesa, Dio, Gesù, Cristo, cristiano, santo, papa, fede, suo, cattolico, vescovo, religioso, uomo, padre
9	6.79%	articolo, emendamento, commissione, legge, esame, numero, relatore, corte, sentenza, comma, giudice
10	4.27%	squadra, gioco, giocatore, gara, partita, giocare, atleta, campionato, calcio, vincere, gol, atletico, tifoso, sport
11	3.01%	paziente, malattia, cellula, medico, farmaco, terapia, clinico, medicina, tumore, patologia, salute, sanitario
12	5.87%	guerra, Iraq, americano, pace, Bush, contro, palestinese, militare, terrorismo, popolo, israeliano, iracheno
13	7.56%	politico, governo, partito, europeo, paese, presidente, politica, Berlusconi, parlamento, maggioranza, voto
14	3.91%	banca, euro, mercato, consumatore, milione, prezzo, aumento, miliardo, sindacato, paese, lavoratore, Cgil
15	3.89%	teatro, arte, mostra, artista, opera, libro, spettacolo, romanzo, poesia, museo, autore, storia, scena, artistico
16	3.20%	Internet, rete, utente, software, prodotto, tecnologia, sito, servizio, cliente, web, elettronico, azienda, digitale
17	6.67%	università, ricerca, scientifico, studio, corso, universitario, laurea, professore, scienza, concorso, biblioteca
18	3.14%	Hotel, mare, hotel, acqua, albergo, isola, zona, situare, vacanza, spiaggia, metro, antico, città, km, ristorante
19	3.96%	Provincia, assessore, comune, sindaco, provinciale, parco, regione, comunale, territorio, area, cittadino

Table 7. English renewable energy topics

1	Table, Equipment, Market, Consumption, Capacity, Production, Industry, Generation, Distribution, Transformers, Sector
2	earth, atmosphere, dioxide, surface, cool, cause, warming, fluid, radiation, methane, reservoir, human, rock, warm
3	reactor, Nuclear, uranium, radioactive, barrel, mine, Uranium, fission, cent, Petroleum, reserve, billion, mining, safety
4	ocean, wave, OTEC, Ocean, tide, Intel, Tidal, Wave, marine, Hawaii, offshore, Conversion, device, surface, conversion
5	Commission, Public, shall, bill, Utility, credit, Federal, contract, FERC, eligible, District, regulation, federal, County
6	distribute, consumer, distribution, network, peak, meter, period, investment, value, datum, average, sector, reliability, factor
7	speed, rotor, blade, field, magnetic, shaft, circuit, wire, engine, transformer, phase, connect, rotate, frequency, torque
8	cogeneration, hydrogen, ethanol, engine, wood, combustion, Biomass, boiler, burn, crop, convert, landfill, residue, gasoline
9	Green, News, Hydro, Business, India, Stock, Development, Sustainable, Geothermal, Alternative, Environmental
10	river, hydroelectric, hydro, reservoir, fish, River, head, blade, hydroelectricity, Hydro, Hydropower, height, stream
11	post, read, bill, want, look, article, problem, money, green, really, question, link, warming, news, storey, idea, What, talk
12	announce, billion, Tags, investment, megawatt, Kansas, expect, sign, April, News, green, California, feed-in, release, Farm
13	sustainable, policy, economic, sector, reduction, Development, management, national, international, community
14	module, light, silicon, inverter, watt, sunlight, hour, roof, saving, device, appliance, save, film, array, tower, house, electron
15	Program, National, Department, California, Center, Association, Resources, Public, Information, Efficiency, Service, page

Table 8. Russian renewable energy topics

1	новость (news), ноутбук (laptop), процессор (processor), комментарий (comment), компьютер (computer), телефон (phone), ученый (scientist), рубрика (heading), память (memory), мобильный (mobile), intel, energy
2	финансовый (financial), государство (state), инвестиция (investment), мера (measures), доля (proportion), политика (politics/strategy), правительство (government), национальный (national), потребность (demand)
3	ооо (ltd), электронный (electronic), бесперебойный (uninterruptible), продажа (sale), дизельный (diesel), купить (buy), ремонт (repair), чертеж (drawing), ибп (UPS), фирма (firm), каталог (catalogue), товар (goods)
4	вот (this/yeah), кто (who), да (yes), сделать (do), сейчас (now), говорить (say), там (there), ни (neither), ли (would), надо (should), просто (simply), сам (-self), знать (know)
5	коммунальный (utilities), жилой (resident), федеральный (federal), водоснабжение (water supply), отопление (heating), оплата (payment), жкх (utilities), помещение (room), энергосбережение (energy saving)
6	рф (Russia), рао еэс (Unified Energy System), подробно (details), инвестиционный (investment), директор (director), правительство (government), президент (president), глава (chairman), новость (news), январь (January)
7	конференция (conference), выставка (exhibition), устойчивый (sustainable), научный (scientific), наука (science), охрана (protection), энергосбережение (energy saving), энергоэффективность (energy efficiency)
8	геотермальный (geothermal), биомасса (biofuel), водород (hydrogen), отходы (waste), возобновляемый (renewable), ветроэнергетика (wind generation), топливный (fuel), сжигание (burning), вэу (wind farm), солнце (sun)
9	море (sea), глобальный (global), океан (ocean), растение (plant), лес (forest), загрязнение (pollution), потепление (warming), природа (nature), парниковый (greenhouse), животное (animal), организм (organism), планета (planet)
10	плотина (dam), добыча (production), сооружение (installation), водохранилище (reservoir), месторождение (deposit), запас (resource), очистка (purification), сырье (raw materials), сточный (sewage)
11	аккумулятор (battery), воздушный (air), насос (pump), емкость (capacity), нагрузка (load), рисунок (drawing), ротор (rotor), конструкция (design), вал (shaft), корпус (body), вращение (rotation), мм (mm), охлаждение (cooling)
12	тэц (CHP), энергосистема (grid), газотурбинный (gas turbine), когенерация (cogeneration), киловольт (kV), нагрузка (load), котельная (boiler station), генерация (generation), теплоснабжение (heating supply), выработка (production)
13	паровой (steam), котел (boiler), силовой (power), трансформатор (transformer), гост (GOST), частота (frequency), пар (steam), линия (line), power, преобразователь (converter), передача (transmission), реактивный (reactive), переменный (alternating), провод (wires)
14	русский (Russian), улица (street), война (war), язык (language), московский (Moscow), республика (republic), школа (school), век (century), ребенок (child), франция (France), текст (text), германия (Germany), игра (game), километр, книга, транспорт, история, культура,
15	реферат (essay), реактор (reactor), движение (movement), ядро (nucleus), наука (science), поле (field), атом (atom), реакция (reaction), нейтрон (neutron), физика (physics), планета (planet), магнитный (magnetic)