Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 95 (2013) 455 – 463

Procedia
Social and Behavioral Sciences

5th International Conference on Corpus Linguistics (CILC2013)

Automatic Access to Legal Terminology Applying two Different Automatic Term Recognition Methods

María José Marín Pérez*, Camino Rea Rizzo

Universidad de Murcia, Campus de La Merced, Murcia 30071, Spain

Abstract

Automatic term recognition (ATR) methods help to identify the most representative terms in a corpus automatically, saving time and allowing managing large amounts of data that could not be dealt with manually. This paper presents the evaluation of two ATR methods implemented on a 2.6 million-word legal corpus designed and compiled *ad hoc:* Keywords (Scott, 2008) and Chung's method (2003). Both techniques have been assessed as regards precision and recall. The results clearly show that Keywords is, by far, the most efficient one achieving to recognize 62% true terms out of the 2,000 items evaluated in this study.

Keywords: automatic term recognition; legal English; specialised corpora.

1. Introduction

The identification of the terms in a specialized corpus is a fundamental task for the researcher interested in understanding the nature of the language variety the corpus texts are encoded in. Terms, according to Spasic et al. (2005:240), are "textual realizations of a specialized concept". They encapsulate such knowledge and are employed by the members of the specialized community to communicate amongst themselves, as stated by Rea (2008). As a matter of fact, authors agree that the terms in a specialized language differ not only from general usage, but also from other varieties of language (Sager, 1980; Rondeau, 1983; Cabré, 1993; Alcaraz, 2000). Cabré (id.) highlights their univocal character as regards the relationship between their form and content, and considers them mono-referential since she believes terms to only designate one object. An insight into the nature of specialized languages

*Corresponding author: Tel.: +34-868-888-622 ; fax: +34-000-000-0000
E-mail address: mariajose.marin1@um.es

1877-0428
doi:10.1016/j.sbspro.2013.10.669

is provided by Rea (id.) whereas the relationship between terminology and Natural Language Processing is explored by Almela (2008).

The potential applications of term recognition are manifold: compilation of specialized glossaries and dictionaries, machine translation, or ontology building, to mention but a few. Thus, finding efficient methods that perform this function automatically in a reliable way might be the first step towards a detailed description of specialized languages, especially taking into consideration the large size of corpora nowadays.

This paper describes the evaluation of two state-of-the-art Automatic Term Recognition methods (henceforth ATR) implemented on a 2.6 million-word legal corpus designed and compiled by the authors, the United Kingdom Supreme Court Corpus (UKSCC). Both Chung's (2003) method and the Keywords tool included in the software pack designed by Scott (2008), Wordsmith 5.0, are validated as regards their precision and recall. The validation process was carried out automatically due to the subjectivity implied in human validation because of the lack of consensus amongst judges assessing candidate term lists. A 10,000 entry legal glossary compiled by the authors (legal English specialists) was employed as gold standard for comparison.

Section 2 presents the structure and characteristics of both the corpora employed in this study: UKSCC, a collection of 192 judicial decisions issued by the Supreme Court of Great Britain; and LACELL, a 20 million-word general English corpus designed and compiled by the LACELL research group at the English Department of the University of Murcia, which the authors belong to. Section 3 is devoted to a description of the methods selected and evaluated while section 4 describes the validation process and its results. Finally, section 5 presents the main conclusion reached after the evaluation of both methods. UKSCC and LACELL: the study and reference corpus.

2. UKSCC and LACELL: the study and reference corpus

UKSCC is an *ad hoc* legal corpus of law reports (collections of judicial decisions) which has been compiled according to corpus linguistics standards as stated in Sánchez et al. (1995) and Wynne (2005) for general corpora and its adaptation to specific corpora (Pearson 1998; Rea 2010). It is a 2.6 million-word specialized corpus which will be used as the study corpus for term identification comprising 192 texts all belonging to the category of law reports (written collections of judicial decisions).The full criteria governing the design and compilation of UKSCC are specified by Marín & Rea (2011).

The general corpus acting as reference for comparison is LACELL (*Lingüística Aplicada Computacional, Enseñanza de Lenguas y Lexicografía*), a 20 million-word general English corpus. It is a balanced synchronic corpus of general English including both written texts from diverse sources such as newspapers, books (academic, fiction, etc.), magazines, brochures, letters and so forth, and also oral language samples from conversation at different social levels and registers, debates and group discussions, TV and radio recordings, phone conversations, everyday life situations, classroom talk, etc. Its geographical scope ranges from USA, to Canada, UK and Ireland.

The reason why the Supreme Court of the United Kingdom was selected as the source to obtain the texts to compile the study corpus from is its importance as a legal institution. It is at the top of the judicial pyramid and acts as the court of last resort in the whole territory. The Supreme Court hears cases which might have been dealt with at English, Northern Irish, Scottish or Welsh courts thus producing texts which, from the point of view of their lexicon, are varied and rich. Furthermore, their lexical richness also derives from the court's wide jurisdiction as it covers all branches of law.

The legal genre the texts belong to, that of law reports, represents one of the major sources of law in common law countries like the United Kingdom where law is 'judge-made', that is to say, law is not codified as it is in countries like Spain which belong to the continental law realm. In common law systems law is based on the existing jurisprudence and, although statutes (the laws passed at the parliament) have gained importance in the last 150 years (Orts, 2006), judicial decisions, as far as they interpret the law and set precedent, stand as the major basis barristers or judges resort to when having to argue or decide on a given case.

3. Description of the ATR methods implemented and tested

ATR methods date back to the late 1980s. They arise with the aim of extracting specialized terms from large document collections which could very difficultly be handled otherwise. Their nature and efficiency has been

profusely reviewed (Maynard & Ananiadou, 2000; Cabré et al., 2001; Drouin, 2003; Lemay et al., 2005; Kit & Liu, 2008 or Vivaldi et al., 2012, amongst many others).

The two methods employed in this study were singled out for two reasons. On the one hand, Chung (2003) records excellent results obtaining over 90% overlap between the automatic identification of terms and the lists produced by specialists in some of the word groups she organizes the token lists into after calculating their ratio of occurrence. As a matter of fact, she manages to reach 86% precision on average in mining single-word terms from her anatomy corpus.

On the other hand, Keywords, a popular tool for corpus analysis which extracts from a specialized corpus those word types which are "unusually frequent (or unusually infrequent) in comparison with what one would expect on the basis of the larger word-lists" (Scott, 2008: 184), also manages to identify the terms in UKSCC even more accurately than other methods designed to that purpose (Marín, 2014).

As far as Chung's technique is concerned, it discriminates terms from non-terms by establishing a threshold cut. In order to be considered as a term, a word's ratio value had to be over 50. Chung reaches this conclusion after validating her method by comparison with a qualitative one, the rating scale approach, with the purpose of assessing the degree of overlap between it and the quantitative technique employed by her. Thus, two experts are asked to classify the vocabulary in a 5,500 word text from her anatomy corpus. They classify the words into four different categories depending on their level of specialization.

In contrast, the quantitative method employed by the author consists in calculating the ratio of occurrence of the types in the anatomy text given to the experts. She normalizes the frequencies of the text types in both her anatomy corpus and a general one and calculates their ratio by dividing the former by the latter. Then, basing her classification on these results and on the absolute frequency figures obtained, she also produces different groups and compares them to the ones by the specialists. The results of the comparison yield 86% coincidence on average, especially regarding highly specialized words and non-terms.

She therefore concludes that it might be possible to identify terms using an automatic method although the last decision to include a word in a given category should be made by the researcher after either consulting the expert or the contexts of occurrence of a given word, since she believes that the most effective approach is the qualitative one.

As regards Keywords, Scott (2008) considers a word is key to a given set of texts because of its capability to point at a text's *aboutness*. This idea turns into a mathematical concept by applying different measures such as chi-square or Dunning's (1993) log-likelihood algorithm in order to identify those words which perform this function. In this study, keyness was calculated adjusting Wordsmith 5.0 to use the latter measure due to its greater accuracy.

4. Validation process and results

4.1Validation process

The evaluation of both methods was carried out in terms of precision and recall. Precision indicates the percentage of true terms a method manages to identify successfully with respect to the total amount of candidates extracted by it. Recall refers to the amount of true terms identified with respect to the total amount of true terms in the corpus. These parameters can be calculated both manually and automatically. Some authors like Chung resort to human validation by asking specialists to confirm which of the candidate terms identified are true ones. There is an objection to this assessment procedure which is related to the high degree of subjectivity that human validation adds to it, apart from the logical limitations imposed by the large size of corpora. Furthermore, agreement amongst judges is often hard to reach and some authors even decide to study their decisions separately precisely due to that fact. This is why we opted for automatic validation. A 10,000 entry legal glossary was compiled by merging four different online legal glossaries. The glossary was then completed after manually supervising the list of discarded elements that may have been left out of the validation results due to varied reasons (two paper dictionaries were employed in this case). Silence (undetected terms) was therefore compensated for by doing so. Precision improved by 3-4% owing to manual supervision.

The corpus texts included in UKSCC and LACELL were both processed using Scott's (2008) Wordsmith 5.0.No cut-off lists were employed to filter the information obtained except for the function wordlist provided by Haley and

Nation (1986) in the software package Range. Keywords was automatically applied to UKSCC, the study corpus, resorting to LACELL as the reference corpus. Chung's (2003) method had to be implemented by copying the data on a spreadsheet and then resorting to the information provided by Wordsmith (normalized frequency counts in the study and reference corpus respectively).

Then, after implementing both methods, the candidates were put on an excel spreadsheet and compared to the gold standard using the 'search function' provided by the software. This method allowed us to establish the degree of coincidence between both lists (candidate term list and glossary list) and thus calculate the precision and recall levels achieved by each method.

4.2 Results

Figure 1 shows the average precision and recall values attained by each method and Figure 2 illustrates cumulative precision after applying both methods. In order to show how this parameter varies as the number of candidate terms increases, it was calculated progressively in groups of 200 which were ranked in descending order according to their respective weight.

As shown by both graphs, the results for *Keywords* are more accurate. It reaches 62% precision and 31% recall on average whereas Chung's method stands at 20 points below for both parameters. However, there are slight differences if we take into consideration cumulative precision, although, in general, Keywords outperforms Chung's method. Starting at almost 30 points below Keywords for the first 200 candidates (it only reaches 48.5% precision as opposed to 85%), Chung's method remains steady for the first 800 candidates reaching a peak of 58% from candidates 800 to 1000. On the other hand, Keywords appears to be much more efficient within this range descending progressively from 85% to 68% precision. From that point on, Chung's method drops sharply to 20% from candidates 1000 to 1200 climbing up progressively back to 58% within items 1600 to 1800, the only point at which it excels Keywords. Finally, it falls dramatically to 3.5% for the bottom 200 candidate terms. In contrast, Keywords performs more constantly within this range and decreases its efficiency steadily practically to the end of the graph, climbing up slightly from items 1600 to 2000 to 50% and 47.5% precision respectively.

It should be highlighted that Chung's method produces high noise levels due to it considering those elements not in the reference corpus as terms automatically. Owing to the character of the type of texts included in UKSCC, judicial decisions, proper names appear everywhere (as shown in the appendix table) and qualify as terms precisely due to that fact. Probably, the results would improve if these elements were not considered for evaluation although we have tried to be as faithful as possible to the author's original method trying not to manipulate the texts before their processing.

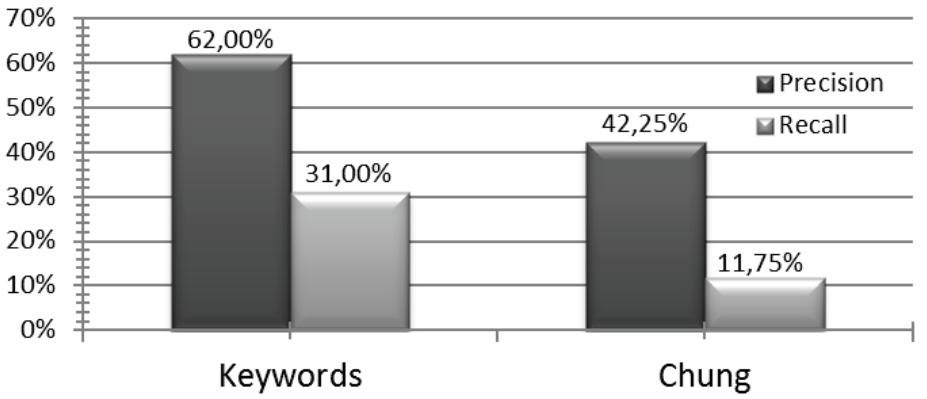

Fig. 1.Overall precision and recall.

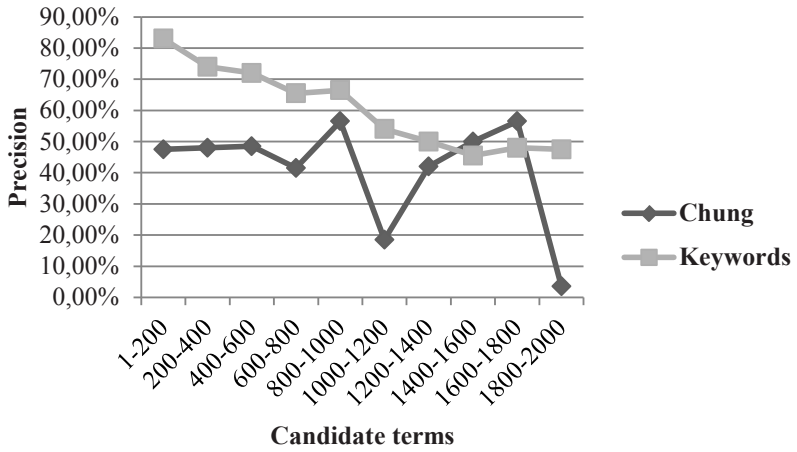

Fig. 2.Cumulative precision on top 2000 candidates.

5. Conclusions

This study has shown a comparison between two automatic term recognition methods, Chung's (2003) and Keywords (2008), as regards the precision and recall levels achieved by each of them when implemented on a 2.6 million-word legal English corpus.

The graphs above show how Keywords excels Chung's method both in terms of precision and recall. As a matter of fact, Chung's technique produces a high degree of noise basically because of its automatic inclusion of words not in the reference corpus within the term list. The results would definitely improve if these words did not qualify as such. Probably, using a specific cut-off list as a filter to eliminate such elements as proper names prior to the implementation of the method would increase precision and recall.

Carrying out a comparison of this kind might be a helpful source of information for the researcher interested in working with the terminology of a given English variety. This way, they can find the most effective method of automatically accessing the terms of the art of a given sublanguage that could be employed for varied purposes such as designing and compiling specialized glossaries and dictionaries, extracting lists of terms for automatic translation, building ontologies, etc.

Finally, other ATR methods could also be tested and compared to find out which one of them is more effective in identifying legal terms although, as stated by Chung, when in doubt, it is the specialist who will make the final decision to include a given word within the list of specialized terms since automatic methods cannot account for such complex and often subjective phenomena as synonymy or polysemy.

Appendix A.

Table 1.Top 100 candidate terms extracted by both methods (true terms are highlighted in bold).

Chung'smethod	Ratio	Keyword	Keyness
EHRR	∞(*)	**COURT**	28955.793
EWCA	∞	**SECTION**	27627.5586
UKHL	∞	**PARA (**paragraph**)**	25311.1152
MANCE	∞	LORD	25155.4434
SIAC	∞	**V** (versus)	22486.0918

Chung'smethod	Ratio	Keyword	Keyness
ECHR	∞	**APPEAL**	21236.8652
EWHC	∞	**ARTICLE**	19301.6328
BAILII	∞	**ACT**	18577.8652
GESTINGTHORPE	∞	**CASE**	18328.9512
FOSCOTE	∞	**LAW**	10458.0918
EARLSFERRY	∞	**JUDGMENT**	9297.75
JFS	∞	**APPELLANT**	8048.33496
ECTHR	∞	**PROCEEDINGS**	7787.61963
STOJEVIC	∞	**CONVENTION**	7764.64355
TURPI	∞	WHETHER	7716.16992
LJ'S	∞	**LJ**	7707.0918
DALLAH	∞	**RIGHTS**	7023.53613
SUMPTION	∞	**DECISION**	6950.50488
SEISED	∞	**ORDER**	6632.18164
BANKOVIC	∞	**JURISDICTION**	6374.33105
MATRAVERS	∞	RELEVANT	6263.90625
TULLIS	∞	**CLAIM**	5832.43506
PANNICK	∞	**AC**	5830.41455
CHAGOSSIANS	∞	**PARAS**	5472.78809
IMGS	∞	APPLICATION	5029.07129
HRA	∞	KINGDOM	4896.52197
ASCO	∞	CIRCUMSTANCES	4704.37988
SKEINI	∞	OPINION	4641.26611
AQO	∞	**STATUTORY**	4629.16748
GOURDE	∞	**PROVISIONS**	4533.56982
IPP	∞	**CASES**	4428.68115
CHARTERERS	∞	**BREACH**	4419.3208
HSMP	∞	**JUDGE**	4404.61816
ARBITRAL	∞	**APPELLANTS**	4372.99072
ECRC	∞	**PRINCIPLE**	4212.11963
GHALANOS	∞	**CRIMINAL**	4197.90332
ALLDECH	∞	QUESTION	4077.19238
HLR	∞	**EHRR**	4068.38989
OTHMAN	∞	LORDS	4008.65527
TAXOL	∞	**PARAGRAPH**	3910.81714
STEART	∞	**WLR**	3907.81592
BCLC	∞	**POSSESSION**	3887.15381
BIOGEN	∞	STATE	3822.15796
APSA	∞	DUTY	3790.97339
SIAC'S	∞	AGREEMENT	3788.57275
ASHLEYS	∞	**DEFENDANT**	3774.44824

Chung'smethod	Ratio	Keyword	Keyness
MUNBY	∞	**QC**	3683.96655
IMM	∞	MR	3676.36938
CHAHAL	∞	**APPLICANT**	3642.31787
APMSD	∞	**AUTHORITY**	3546.46826
BANCOULT	∞	SECRETARY	3524.19922
LPP	∞	FACTS	3516.82251
RUNA	∞	LTD	3420.80054
INTERVENER	∞	CONCLUSION	3411.19263
SALDUZ	∞	PURPOSES	3405.35791
MISFEASANCE	∞	PERSON	3389.80518
AVERMENTS	∞	**ISSUE**	3374.36133
SEISIN	∞	**EVIDENCE**	3299.42969
ENANTIOMER	∞	**RESPONDENT**	3252.15674
LUBA	∞	**TRIBUNAL**	3209.26611
CHARTBROOK	∞	**RULE**	3033.38306
IPT	∞	SUBSECTION	3029.07031
CHAGOS	∞	HOFFMANN	2972.30322
MANDLA	∞	PARLIAMENT	2966.60986
OFULUES	∞	STRASBOURG	2945.81201
EXCEPTIONALITY	∞	**ENTITLED**	2941.0791
JOINDER	∞	RESPECT	2833.23706
CHARTERPARTY	∞	REASONS	2803.30688
CORPN	∞	RELATION	2796.54736
HMRC	∞	**EXTRADITION**	2781.37329
HYDRODAM	∞	**TENANT**	2751.19141
UKPC	∞	UNITED	2719.12964
STOJEVIC'S	∞	LORDSHIPS	2699.44385
FSMA	∞	EFFECT	2693.64722
JFS'S	∞	**JUDICIAL**	2677.0271
WTR	∞	**COURTS**	2676.51587
MOORGARTH	∞	**PARTIES**	2654.61938
CORR'S	∞	**APPELLANT'S**	2563.87329
ENVIROCO	∞	PURPOSE	2540.96729
LONGMORE	∞	OBLIGATION	2539.6167
OCEANBULK	∞	APPLY	2511.0437
BURNTON	∞	**CONTRACT**	2492.60645
TMT	∞	R	2478.07495
EGLR	∞	**PROVISION**	2477.90698
REDLAW	∞	**EWCA**	2451.50879
REINSURERS	∞	BASIS	2411.77612
BOCARDO'S	∞	BINGHAM	2389.14355

Chung'smethod	Ratio	Keyword	Keyness
DALLAH'S	∞	**ARGUMENT**	2386.76733
EURCTHR	∞	**OFFENCE**	2385.9021
INCAPAX	∞	ASYLUM	2349.10596
OBITER	∞	**UNLAWFUL**	2339.28613
TRM	∞	NOBLE	2333.81494
VATA	∞	**DAMAGES**	2309.95361
BHRC	∞	**LIABILITY**	2284.2793
MÜLLER'S	∞	REASONABLE	2278.59497
OFFEREN	∞	J	2231.50293
OUSELEY	∞	HELD	2141.46338
TRAVAUX	∞	**REGULATION**	2119.13867
AQMAU	∞	**SENTENCE**	2086.39966

(*) These candidate terms display a ratio value that equals infinity due to them not being in the reference corpus and thus being divided by 0.

References

Alcaraz, E. (2000). *El inglés profesional y académico*. Madrid: Alianza Editorial.

Almela, A. (2008). *Evaluating multiword automatic term recognition techniques on a veterinary medicine corpus.* MA Thesis. Murcia: Universidad de Murcia.

Cabré, M.T. (1993). *La terminología. Teoría, metodología, aplicaciones.* Barcelona: Antártida/Empúries.

Cabré, M.T. (2000).Terminologie et linguistique*:* la théorie des portes. *Terminologies nouvelles. Terminologie et diversité culturelle*, *21*, 10-15.

Cabré, M. T., Estopà, R., and Vivaldi, J. (2001). Automatic term detection: a review of current systems. In D. Bourigault, C. Jacquemin, and M.C. L'Homme (Eds.), *Recent advances in computational terminology 2*, (pp. 53–87). Amsterdam: John Benjamins, Natural Language Processing.

Chung, T. M. (2003). A corpus comparison approach for terminology extraction. *Terminology, 9(2),* 221-246.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19(1)*, 61-74.

Heatley, A. and Nation, I.S.P. (1996). *Range* (computer software). Wellington: Victoria University of Wellington. URL: http://www.victoria.ac.nz/lals/about/staff/paul-nation

Kit, C., and Liu, X. (2008). Measuring mono-word termhood by rank difference via corpus comparison. *Terminology*, *14(2)*, 204-229.

Lemay, C., L'Homme, M.C., and Drouin, P. (2005). Two methods for extracting specific single-word terms from specialized corpora: experimentation and evaluation. *International Journal of Corpus Linguistics, 10 (2)*, 227-255.

Marín, M.J. (forthcoming, 2014). Evaluation of five single-word terms recognition methods on a legal corpus. *Corpora, 9(1)*. Edinburgh: Edinburgh University Press.

Marín, M.J. and Rea, C. (2011). Design and compilation of a legal English corpus based on UK law reports: the process of making decisions. In M.L. Carrió Pastor and M.A. Candel Mora (Eds.), *Las tecnologías de la información y las comunicaciones: Presente y futuro en el análisis de corpora*. *Actas del III congreso internacional de lingüística de corpus* (pp. 101-110). Valencia: Universitat Politècnica de València. URL: http://webs.um.es/mariajose.marin1/miwiki/doku.php?id=scientific_production

Maynard, D. and Ananiadou, S. (2000). TRUCKS: A model for automatic multi-word term recognition. *Journal of Natural Language Processing, 8(1)*, 101–125.

Pearson, J. (1998). *Terms in Context*. Amsterdam: John Benjamins Publishing.

Rea, C. (2008). *El inglés de las telecomunicaciones: Estudio léxico basado en un corpus específico.* PhD Thesis. Murcia: Universidad de Murcia. URL: http://www.tdx.cat/handle/10803/10819

Rea, C. (2010). Getting on with corpus compilation: from theory to practice. *ESP World, 1 (27).*

Rondeau, G. (1983). *Introduction à la terminologie.* Québec: Gaëtan Morin Editeur.

Sager, J., Dungworth, D., and McDonald, P. (1980). *English special languages. Principles and practice in science and technology*. Wiesbaden: Brandstetter Verlag KG.

Sánchez, A., Cantos, P., Sarmiento R. and Simón, J. (1995). *Cumbre. Corpus lingüístico del español contemporáneo. Fundamentos, metodología y análisis*. Madrid: SGEL.

Scott, M. (2008). *WordSmith Tools 5.0.* Liverpool: Lexical Analysis Software.

Spasic, I., Ananiadou, S., McNaught, J. and Kumar, A. (2005). Text mining and ontologies in biomedicine: Making sense of raw text. *Brief Bioinform*, *6(3)*, 239-251.

Vivaldi, J., Cabrera-Diego, L.A., Sierra, G. and Pozzi, M. (2012). Using wikipedia to validate the terminology found in a corpus of basic textbooks. In *Proceedings of the eight international conference on language resources and evaluation (LREC'12).* Istanbul, May 2012. URL: http://www.lrec-conf.org/proceedings/lrec2012/index.html

West, M. (1953). *A general service list of English words*. London: Longman.

Wynne, M. (Ed.) (2005). *Developing linguistic corpora: a guide to good practice.* Oxford: ASDS Literature, Languages and Linguistics.