>>> revista tradumàtica
Traducció i Tecnologies de la Informació i la Comunicació
Número 07: L'aplicació dels corpus lingüístics a la traducció : desembre 2009 :
http://www.fti.uab.cat/tradumatica/revista : ISSN: 1578-7559

Catalogue of Free-Access Translation-Related Corpora

Viviane Possamai
PhD Student, Translator
Universidade Federal do Rio Grande do Sul

The use of corpora – and of computer tools that make possible to analyze them – have brought new insights to the field of translation over the last 10 years. Corpora have been used in different translation-related areas. In translation studies the application of corpus technologies and methodologies enable translation scholars to explore the translated text as a mediated communicative event (Baker, 2003: 233) by examining the linguistic features characteristic of translational language (Machniewsk, 2006: 237). In translation teaching corpora can help students to understand the source language text and to produce fluent target language texts (Machniewsk, 2006). In the practice of translation, translators can use corpora to confirm their decisions, to know the collocates of certain words, to look for equivalents, to extract terminology and to see words in their context. With so many different applications and target users it is understandable why each corpus available on the Internet today has its own unique features, with different types of information, tools or text types available.

Although most corpora available for free can be used for translation-related purposes, in this catalogue we would like to have included only corpora that have explicitly declared to be related to translation, however, that was not possible because we did not find exclusively translation-related corpora for some languages. We have then included the first corpora that resulted from a Google search using the words *corpora* and *corpus* plus the language in question (i.e. *corpora* + Russian) and some listed in websites that compile lists of corpora (for example, David Lee's compilation, available at http://personal.cityu.edu.hk/~davidlee/devotedtocorpora/CBLLinks.htm).

We made two major distinctions in the types of corpora listed below: single language corpora and parallel corpora, since parallel corpora are very useful for translation studies and in the translation practice. We also included the Translated English corpus compiled by researchers from the University of Manchester. All the information provided was extracted from the corpora websites and/or related research articles.

We are aware that this is not an extensive list, but we hope it brings some insight and some clues to the reader by introducing some of the most important well-established and freely available corpora over the Internet.

References

Baker, M. (1993). "Corpus linguistics and translation studies: implications and applications." In Baker, M. et al. (eds.) (1993), *Text and Technology. In Honour of John Sinclair*. Amsterdam: John Benjamins.

Machniewski, M. (2006). Analysing and teaching translation through corpora: Lexical convention and lexical use. *Poznań Studies in Contemporary Linguistics* 41: 237-255.

Arabic

Name	arabiCorpus
URL	http://arabicorpus.byu.edu/
Features	• Composed of: newspaper texts, novels, medieval medical and philosophical texts, plays • Untagged • Search for individual and multiple words
Number of words	+ 65 million words

Catalan

Name	Cucweb
URL	http://ramsesii.upf.es/cgi-bin/cucweb/search-form.pl
Features	• Composed of: documents extracted from the Web under the .es domain • Tagged • Allows queries by lemma, part of speech, and syntactic function
Number of words	+ 200 million words

Name	Corpus Textual Informatitzat de la Llengua Catalana
URL	http://ctilc.iec.cat/
Features	• Composed of: literature texts (narrative, theatre, poetry, essays); treaties, manuals, articles, legal text, newspapers • Tagged • Allows search by lemma and filtering by author/work, time and type
Number of words	+ 52 million words

Chinese

Name	Chinese Internet Corpus
URL	http://corpus.leeds.ac.uk/query-zh.html
Features	• Composed of: Internet texts • Concordance and collocation tools
Number of words	280 million words

Name	Chinese Business Corpus
URL	http://corpus.leeds.ac.uk/query-zh.html
Features	• Composed of: Internet texts • Concordance and collocation tools
Number of words	30 million words

English

Name	Corpus of Contemporary American English

URL	http://www.americancorpus.org/help/corpusName_e.asp
Features	• Composed of: spoken, fiction, popular magazines, newspapers, and academic texts • Allows search by exact words or phrases, wildcards, lemmas, part of speech
Number of words	+ 400 million words

Name	American National Corpus
URL	http://www.americannationalcorpus.org/OANC/index.html
Features	• Composed of: fiction texts, documents, newspaper texts • Words (tokens) with part of speech annotations
Number of words	14 million words

Name	Collins Wordbanks *Online* English corpus
URL	http://www.collins.co.uk/Corpus/CorpusSearch.aspx
Features	• Composed of: contemporary written and spoken texts • Allows search by word combinations, wildcards, part-of-speech tags
Number of words	56 million words

French

Name	Frantext
URL	http://www.cnrtl.fr/corpus/
Features	• Composed of: French literature texts • Downloadable corpus
Number of words	146,693,289 characters

Name	Corpus journalistique de l'Est Républicain
URL	http://www.cnrtl.fr/corpus/
Features	• Composed of: newspaper texts • Downloadable corpus
Number of words	n/a

German

Name	COSMAS II
URL	http://www.ids-mannheim.de/cosmas2/
Features	• Composed of: literary texts, national and regional newspapers, the works of Marx and Engels, spoken language in transcribed form, morphosyntactically transcribed texts • Registered users can submit queries online and obtain concordances and word frequency count
Number of words	1,1 billion word; invited guests have access to the whole COSMAS corpus collection (currently 1,85 billion words)

Name	DWDS Digitales Wörterbuch der deutschen Sprache

URL	http://www.dwds.de/textbasis
Features	• Composed of: specialized corpora (e.g. spoken language, language of the former German Democratic Republic GDR), and the DWDS core corpus, a balanced corpus of German texts from the 20th century. tagged • supports linguistic queries on several annotation levels (word forms, lemmas, STTS part-of-speech categories) and offers filtering (author, title, text type, time intervals)
Number of words	The core corpus consists of 100 million tokens

Name	NEGRA
URL	http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/
Features	• Composed of: German newspaper texts • Tagged with part-of-speech and completely annotated with syntactic structures
Number of words	355,096 words (20,602 sentences)

Name	TIGER Corpus
URL	http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/
Features	• Composed of: German newspaper texts • POS-tagged and annotated with syntactic structure
Number of words	900,000 words

Italian

Name	CORIS/CODIS
URL	http://corpora.dslo.unibo.it/coris_eng.html
Features	• Composed of: authentic and commonly occurring texts in electronic format • Concordance and collocation searches
Number of words	100 million words

Japanese

Name	Leipzig Corpora Collection
URL	http://corpora.informatik.uni-leipzig.de/download.html
Features	• Composed of: newspaper texts or texts randomly collected from the web • Corpora processed automatically • For each word, the most significant words appearing as immediate left neighbour, as immediate right neighbour, anywhere within the same sentence are given
Number of words	n/a

Portuguese (Brazilian)

Name	CETENFolha
URL	http://www.linguateca.pt/CETENFolha/

Features	• Composed of newspaper texts • Available in two versions (tagged and untagged) • Can be downloaded through FTP or HTTP or accessible through http://www.linguateca.pt/acesso/corpus.php?corpus=SAOCARLOS
Number of words	24 million words

Portuguese (European)

Name	CETEMPúblico
URL	http://www.linguateca.pt/CETEMPublico/
Features	• Composed of: newspaper texts • Can be downloaded through FTP or HTTP or accessible through the website
Number of words	180 million words

Name	Projecto AC/D
URL	http://www.linguateca.pt/ACDC/
Features	• Large collection of different corpora
Number of words	Total number of words 362,156,776

Russian

Name	BOKR (The Russian Reference Corpus)
URL	http://bokrcorpora.narod.ru/index-en.html
Features	Composed of: Pilot version
Number of words	100 million words

Name	Russian collection
URL	http://corpus.leeds.ac.uk/ruscorpora.html
Features	Composed of: • a pilot version of the Russian National Corpus (a representative collection of various genres) • Russian newspapers (several major Russian newspapers, 2001-2004) • Russian Internet Corpus (modern Russian language as used on the Internet) • Russian fiction (morphosyntactic features have been manually disambiguated)
Number of words	Russian National Corpus (50 million words) Russian newspapers (70 million words) Russian Internet Corpus (160 million words) Russian fiction (1,5 million words)

Spanish

Name	Corpus del Español
URL	http://www.corpusdelespanol.org/x.asp
Features	• Composed of: general texts from the 1200s to the 1900s • Search for exact words or phrases, wildcards, lemmas, part of speech,

	or any combination of these. You can search for surrounding words (collocates) within a ten-word window (e.g. all nouns somewhere near *cadena*, all adjectives near *mujer*, or all nouns near *girar*)
Number of words	100 million words

Name	Corpus lingüístico de referencia de la lengua española en Argentina
URL	http://www.lllf.uam.es/
Features	• Composed of: legal, technical, literary, newspaper, scientific, commercial and school texts • Texts are downloadable from the website
Number of words	2 million words

Name	Corpus lingüístico de referencia de la lengua española en Chile
URL	http://www.lllf.uam.es/~fmarcos/informes/corpus/cochile.html/
Features	• Composed of: legal, technical, literary, newspaper, scientific, commercial and school texts • Texts are downloadable from the website
Number of words	2 million words

Parallel corpora

Name	BOnonia Legal Corpus - BOLC
URL	http://corpora.dslo.unibo.it/bolc_eng.html
Features	• Composed of: legal texts
Number of words	Italian (33,5 million); English (21 million)
Languages	Italian and English

Name	Corpus Tècnic
URL	http://bwananet.iula.upf.edu/
Features	• Composed of: Law, Economy, Genomics, Medicine and Environment texts • The corpus can be explored through a tool named Bwananet
Number of words	100 million words
Languages	Catalan, Spanish, English, French and German

Name	E-C Concord
URL	http://ec-concord.ied.edu.hk/paraconc/index.htm
Features	• Composed of: novels, academic articles, essays; legal documents, literary texts • Contains a concordance tool
Number of words	Chinese (413 million words), English (1,4 million words)
Languages	Chinese, English

Name	COMPARA

URL	http://www.linguateca.pt/COMPARA/index.php
Features	• Composed of: literary texts • allows users to search for sentences that have been joined, split, added to, deleted from, and reordered in translation. • Other searchable features are translators' notes, foreign words, titles, emphasis and named entities
Number of words	Portuguese (1,4 million words), English (1,5 million words)
Languages	Portuguese, English

Translated English

Name	Translational English Corpus
URL	http://www.llc.manchester.ac.uk/ctis/research/english-corpus/
Features	• Composed of: translated texts from fiction, biography, news and inflight magazines • Documented in terms of extralinguistic features such as gender, nationality and occupation of the translator, direction of translation, source language, publisher of the translated text, etc.
Number of words	Around 10 million words