Comparable Web-Crawled Corpora as a Resource for Contrastive Studies

Parallel corpora are widely used as sources of language data to support contrastive research. However, they have a number of limitations that may or may not affect the outcomes of a contrastive study.

First, translated texts might not typify a language well. Translations were shown to have specific statistical properties (known as *translationese*) that distinguish them from originally-authored documents. Besides, in some parallel corpora the direction of translation can be unknown or mixed, making it difficult to account for translationese on either source or target side of the corpus.

Second, parallel corpora are prone to issues with their representativeness – they usually contain texts of only a few domains and registers (e.g. documents provided by international institutions such as UN or European Union, movie subtitles and open-source software documentation). As a small example, if we try to look for words like “word”, “eye”, or “love”, having a frequency of 378.5, 331.1, and 253.97 per million in the *British National Corpus*, whereas only 86.9, 22.0, and 0.26, respectively, in *EUR-Lex English*, which is one of the largest parallel corpora publicly available.

Third, and probably most important, for many language combinations there are simply not enough parallel texts available. The Czech *InterCorp*[1] Project, collecting parallel texts for 40+ languages, was able to produce a 100+ M token corpora only for two languages (English and Spanish), leaving some 15 languages below a 20 M token threshold.

This is why linguists working in the area of contrastive studies have to work with resources that do not contain mutual translations. Such corpora are usually referred to as “comparable”, i.e. having similar characteristics (text types, genres, domains and registers, time of origin, compatible annotation, and possibly also the size). We would like to argue that corpora created from texts acquired by “general crawling” the web, if a suitable methodology is applied, can relatively easily provide for such comparable language resources.

Though web-crawled corpora have been compiled during the last two decades by several research groups, “comparable methodology” have been applied only to some of the projects. A short summary of the projects we are aware is shown in Table 1:

Table 1: Multilingual Web-Crawled Corpora Projects

Project	**Langs**	**PoS tags & Lemmas**	**Corpus Manager**	**Access**
COW[2]	6	Yes	NoSketch Engine[3]	Registration[4]
CLARIN SI[5]	10	Yes	NoSketch Engine, KonText[6]	Open
Leeds Internet Corpora[7]	16	Yes	CQPweb[8]	Open
Aranea[9]	25	Yes	NoSketch Engine	Open/Registration[10]
Sketch Engine[11]	~ 45	Yes[12]	Sketch Engine[13]	Paid
Wortschatz Leipzig[14]	293	No	(A Custom Query System)	Open

[1] https://intercorp.korpus.cz/

[2] https://www.webcorpora.org/

[3] https://nlp.fi.muni.cz/trac/noske

[4] Accounts are not provided for graduate students

[5] http://www.clarin.si/info/concordances/

[6] https://github.com/czcorpus/kontext

[7] http://corpus.leeds.ac.uk/internet.html

[8] https://cwb.sourceforge.io/cqpweb.php

[9] http://unesco.uniba.sk/guest/

[10] Registration is required to work with larger corpora

[11] https://www.sketchengine.eu/

[12] For most languages

[13] https://www.sketchengine.eu/

[14] https://corpora.wortschatz-leipzig.de/

It is obvious that a primary selection criterion for each linguist is the presence of the respective languages they want to study. From this perspective, the Leipzig portal seems to be “unbeatable”. On the other hand, not a rather rudimentary query system but, above all, the absence of morphosyntactic annotation and lemmatization are the main drawbacks here, especially for languages with rich morphology.

The *Sketch Engine* is a wonderful option for those who have access to the respective license. A considerable number of languages is covered, and the quality of annotation is relatively high. The main advantage, however, is the presence of functionalities not included in *NoSketch Engine*, *Sketch Engine'*s open-source subset*:* collocation profiles (“word sketches”), distributional thesauri, sketch differences, calculation of multi-word terminology units, etc.

The *COW* and *CLARIN SI* portals are good choices to work with languages they provide.

Leeds Internet Corpora portal also includes some Asian Languages (Arabic, Chinese, Japanese and Georgian – if considered to be Asian ;-), and some less-resourced languages (such as Lithuanian). The *CQPweb* corpus manager is slightly less user-friendly and does not include some functionalities present in *NoSketch Engine* used by (almost) all other portals. It has, however, some unique functionalities of its own.

Aranea can be considered a “compromise” in many situations. It is well-suited for pedagogical purposes. The smaller corpora (125 M) can be used without registration, so that students can be offered hands-on tasks from the very beginning of a training session.

The two appendices to this annotation show relative frequencies (ipm=items per million) of twenty most frequent adjectives extracted from six *Aranea Minus* corpora for Russian, Ukrainian, Czech, French, Spanish, and Romanian, respectively[15]. Such lists can be conveniently used in teaching contrastive lexicology for the respective languages.

The lists invite a discussion of interesting cross-linguistic observations:

- Spanish is the only language where the adjective for language/country/nation (“Spanish”) did not make it into the top of the list. Spanish is spoken in many countries, and, unlike in other cases, the use of the adjective to refer to national entities is limited.
- Romanian is the only language having “European” among the most used adjectives.
- The Ukrainian list contains two ordinal numerals that are annotated as a subset of adjectives in all languages.
- The Russian language has two adjectives for the country name – one for the respective nationality and one for the ethnic group and language.
- The double appearance of “Romanian” in the respective list is most likely due to flaws in text filtration (some texts seem to lack diacritics). This is a typical artifact of corpus per-processing that linguists should be aware of.
- Variation in the rank of semantically similar adjectives (new, good, important, different) can indicate cross-linguistic differences in idiomatic patterns of these most-frequent items.
- The comparisons of the lists reveals differences in part-of-speech annotation across languages: in the Czech corpus “každý” (each) is tagged as an adjective while in other corpora similar words are treated as pronouns or pronominal determiners. Romanian “şi” (and) stands out as a word with conjunctive functions that is referred to adjectives when using a universal tag set. Spanish “nuestro” (our) is more often annotated as an adjective, although it has a clearly pronominal functionality.

This contribution aims to showcase *Aranea*, a family of comparable web-corpora, as a resource for contrastive studies. We will demonstrate research and pedagogical potential of corpus-derived bi- and multilingual data for comparative studies of collocations and keyword lists, as well as idioms by means of the Context query functionality of *NoSketch Engine*.

[15] For all languages, the lists have been created by a CQL expression [atag="Aj"]

Appendix 1: Adjectives in Three Slavic Languages (with English equivalents)

ru	> en	ipm	uk	> en	ipm	cs	> en	ipm
новый	new	1,304.4	український	Ukrainian	1,366.3	velký	big/large	1,665.9
должный	due	1,013.4	новий	new	1,233.9	další	next	1,583.8
большой	big/large	793.5	перший	first	1,147.8	nový	new	1,542.3
основной	basic	542.3	державний	state	1,130.6	dobrý	good	1,416.9
российский	Russian (nationality)	509.6	великий	big/large	901.4	celý	whole	1,198.3
последний	last	466.2	різний	different	719.8	každý	each	1,185.9
главный	main	456.8	національний	national	670.0	jiný	other	1,098.9
различный	various	453.3	основний	basic	643.8	český	Czech	1,081.0
высокий	tall	447.7	міський	urban	572.6	malý	small	744.6
хороший	good	429.4	головний	main	571.5	vysoký	high	700.4
общий	general	426.9	повинний	due	569.8	možný	possible	571.3
государственный	state	426.1	соціальний	social	566.5	poslední	last	568.3
современный	modern	420.5	міжнародний	international	519.5	různý	different	543.6
разный	different, other	419.8	місцевий	local	505.4	rád	order	469.2
необходимый	necessary	405.9	загальний	general	493.7	starý	old	464.5
нужный	required	392.0	другий	second	471.6	vlastní	own	464.1
следующий	next	384.9	сучасний	modern	462.7	hlavní	main	443.9
важный	important	371.1	останній	last	461.7	ostatní	other	365.6
полный	full	366.7	російський	Russian (nationality)	461.4	dlouhý	long	365.5
русский	Russian (ethnonym)	352.4	навчальний	educational	432.8	důležitý	important	362.7

Appendix 2: Adjectives in Three Romance Languages (with English equivalents)

fr	> en	ipm	es	> en	ipm	ro	> en	ipm
autre	other	1,738.8	grande	big/large	1,501.1	şi	*and	2,166.1
grand	big/great	1,139.1	bueno	good	1,197.9	mare	big/large	1,596.3
nouveau	new	1,111.7	nuevo	new	1,169.2	nou	new	1,213.6
bon	good	912.0	mismo	same	738.2	bun	good	836.9
petit	little	872.8	nuestro	our	618.8	european	European	564.6
même	same	783.5	pequeño	little	571.9	mic	little	543.9
dernier	last	576.9	último	latest	533.1	public	public	506.1
seul	alone	462.2	social	social	473.8	singur	single	467.5
beau	beautiful	452.1	propio	own	468.2	important	important	443.8
social	social	439.6	público	public	465.7	general	general	394.4
français	French	436.8	importante	important	424.4	local	local	381.1
différent	different	434.9	general	general	420.6	social	social	338.1
nombreux	numerous	362.3	nacional	national	411.2	propriu	own	337.4
national	national	329.0	solo	only	355.7	român	Romanian	312.9
général	general	328.5	diferente	different	335.2	roman	Romanian	307.7
important	important	325.6	político	political	330.3	special	special	303.6
politique	political	320.9	alto	high	326.6	necesar	necessary	292.4
jeune	young	310.0	internacional	international	325.1	politic	political	282.7
possible	possible	305.2	posible	possible	321.8	economic	economic	277.7
meilleur	better	299.0	humano	human	313.6	diferit	different	275.4