# **T124** Search Engine Upgrades: Parallel Corpora

**IDs in Google Sheets: T124**

**Author:** INT

**What exists:** Corpus search engine for single corpora (BlackLab); Poly-GrETEL, Dutch Parallel Corpus with limited functionality; functionality in the commercial Sketch Engine...

**What must be adapted / extended / created anew: Corpus search engine for parallel corpora based on BlackLab search engine**

Functionality

1. retrieve and display matching “part” from other corpus for any single-corpus query
2. Alignments on different levels (document, paragraph, sentence, word)
3. Possible one to many word alignment (appeltaart> apple pie)
4. Gecombineerde bevraging, zie voorbeeld in de [Sketch engine](https://the.sketchengine.co.uk/bonito/corpus/first?corpname=preloaded%2Fopus2_nl&reload=&iquery=hond&queryselector=iqueryrow&lemma=&lpos=&phrase=&word=&wpos=&char=&cql=&default_attr=word&wpos_opus2_af=&wpos_opus2_sq=&wpos_opus2_ar=&wpos_opus2_bs=&lpos_opus2_pt_BR=&wpos_opus2_pt_BR=&lpos_opus2_bg=&wpos_opus2_bg=&wpos_opus2_zh=&wpos_opus2_zh_TW=&wpos_opus2_hr=&wpos_opus2_cs=&wpos_opus2_da=&sel_aligned=opus2_en&pcq_pos_neg_opus2_en=pos&iquery_opus2_en=cat&queryselector_opus2_en=iqueryrow&lemma_opus2_en=&lpos_opus2_en=&phrase_opus2_en=&word_opus2_en=&wpos_opus2_en=&char_opus2_en=&cql_opus2_en=&filter_nonempty_opus2_en=on&wpos_opus2_et=&wpos_opus2_fi=&lpos_opus2_fr=&wpos_opus2_fr=&lpos_opus2_de=&wpos_opus2_de=&wpos_opus2_el=&wpos_opus2_he=&wpos_opus2_hi=&wpos_opus2_hu=&lpos_opus2_it=&wpos_opus2_it=&lpos_opus2_ja=&wpos_opus2_ja=&lpos_opus2_ko=&wpos_opus2_ko=&wpos_opus2_lv=&wpos_opus2_lt=&wpos_opus2_mk=&wpos_opus2_no=&wpos_opus2_fa=&wpos_opus2_pl=&lpos_opus2_pt=&wpos_opus2_pt=&wpos_opus2_ro=&lpos_opus2_ru=&wpos_opus2_ru=&wpos_opus2_sr=&wpos_opus2_sk=&wpos_opus2_sl=&lpos_opus2_es=&wpos_opus2_es=&wpos_opus2_sv=&wpos_opus2_tr=&wpos_opus2_uk=&fc_lemword_window_type=both&fc_lemword_wsize=5&fc_lemword=&fc_lemword_type=all&fc_pos_window_type=both&fc_pos_wsize=5&fc_pos_type=all&usesubcorp=) voor lexicale query en <http://www.ccl.kuleuven.be/~liesbeth/talks/LREC2016-Poly-GrETEL.pdf> voor syntactische query; [TreeAligner](http://www.lrec-conf.org/proceedings/lrec2014/pdf/504_Paper.pdf)
5. Relation with treebank querying (aligned treebanks)?
6. use cases from translation experts?

*Literature:*

<http://www.lrec-conf.org/proceedings/lrec2014/pdf/504_Paper.pdf>

<http://www.usc.es/congresos/pacor/abstracts1.html>

<http://martinweisser.org/corpora_site/concordancers.html>

**Why important for CLARIAH (scientific impact):** requests from Flanders and many others; Leiden bilingual corpora

Parallel corpora are useful for a range of users:

1. language learners, who want to get translation suggestions for a certain source language word or phrase.
2. translators, who similar to language learners want to translation suggestions for a word or phrase, in particular for genre and domain-specific translation variants.
3. linguists, who would like to formulate precise (contrastive) queries with reference to linguistic annotations and alignments. E.g. zero article use in English compared to Dutch/German.

For the last group a powerful query system is needed. The system should allow queries with conditions over both sides of a parallel text plus conditions on word and phrase alignment and allow negation constraints.

The availability of a powerful corpus query system for parallel corpora will allow for a much broader user group.

**Targeted/Actual users:** linguistic researchers; translators; language learners; developers of language learning tools

**Actual use (quantify!):** none (facility still nonexistent); many for single corpora engines

**Social Impact** **(concrete examples):** accelerated linguistic research; language learners

**Estimate in PMs (try to justify):** 8pm

**Lead + PMs:** INT

**Participants + PMs:** INT

**Deliverables:**

* T124 Parallel corpora D124M1 Milestone Document Requirements and technical design
* T124 Parallel corpora D124D1 Deliverable Software Parallel corpus search engine: backend and frontend
* T124 Parallel corpora D124D2 Deliverable Document Parallel corpus search engine: documentation