# **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. Combined query, compare functionality Sketch engine for lexical query and <http://www.ccl.kuleuven.be/~liesbeth/talks/LREC2016-Poly-GrETEL.pdf> for syntactic query

*Literature e.g:*

<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; EDGeS bible corpus

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 monolingual corpora engines

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

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

**Lead + PMs:** INT

**Participants + PMs:** INT

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Taak** | **Type personeel** | **Available** | **2019** | **2020** | **2021** | **2022** | **2023** | **Summed** | **Task total** |
| T124 Parallel corpora | Software developer | 7 |  |  | 1 | 3 | 3 | **7** | 8 |
| T124 Parallel corpora | (Computational) linguist | 1 |  |  | 1 |  |  | **1** |  |

**Deliverables:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Task | ID | D or M | Type | Description |
| T124 Parallel corpora | D124M1 | Milestone | Document | Document defining the requirements for the BlackLab-based parallel corpus search, based on input from a set of use cases. |
|  | D124D1 | Deliverable | Software | update pages\_kb set subissue=issue\_info.subissue from issue\_info where issue\_info.page=page\_id |
|  | D124D2 | Deliverable | Document | Document describing D124D1 and its application to use cases. |