# SEARCHING MULTI-WORDS SIMULTANEOUSLY IN MULTIPARALLEL CORPORA

**Johannes Graën** **Martin Volk** **Simon Clematide**
Institute of Computational Linguistics
University of Zurich
Zurich, Switzerland
{graen|volk|siclemat}@cl.uzh.ch

PARSEME Note: *This paper is related to WG 1 (Lexicon-Grammar Interface) and to WG 3 (Statistical, Hybrid and Multilingual Processing of MWEs).*

## Abstract

We describe a web-based system for searching translations of multi-word units in large multiparallel corpora. The system covers the debates of the European Parliament in five languages and offers a unique resource for linguists, terminologists and translators.

Our search tool provides a simple and intuitive user interface, which supports content-oriented queries while relieving the user from specifying complicated search expressions in a complex query language[1]. We describe the automatic preprocessing of the linguistic data, the retrieval component, and the techniques needed for offering a zero-configuration search.

## Introduction

Large collections of multiparallel texts, i.e. multilingual documents with aligned paragraphs or sentences across all languages, are openly available. These corpora are highly useful and valuable for translators, terminologists, and contrastive corpus linguists if they can be exploited effectively (see Volk et al. (2014) for a more detailed discussion about their usability, the covered languages, and the amount of integrated parallel data).

We have built *Multilingwis*[2] *(Multilingual Word Information System)*, a web-based search tool for multiparallel word-aligned corpora. It provides a simple search interface for translations[3] of multi-word units in the available languages, currently English, French, German, Italian and Spanish. The tool is optimized for quick ad-hoc searches and explorations of translation variants and supports content-oriented access to translated multi-word units across multiple languages.

Our automatically computed rich and multi-layered annotation of the multiparallel corpus provides the basis for the search requests. This annotation includes the automatic alignment of text units (e.g. speech turns in debates), sentences, and words.

Figure 1 shows the output of a sample query. We queried for the German support verb unit "*in Betracht ziehen*" (literally: to draw into consideration). Multilingwis returns the translation variants in four languages with their respective frequencies in the corpus. Function words are ignored in both the query and the result set which enables the system to pool

---

[1] The query corresponding to the free form input "violaciones de los derechos humanos" into our system would be "`[lemma="violación"] [upos!="ADJ|ADV|VERB|NOUN"]{0,3} [lemma="derecho"] [upos!="ADJ|ADV|VERB|NOUN"]{0,3} [lemma="humano"] within s`" for the well-known CQP query language.

[2] http://pub.cl.uzh.ch/purl/multilingwis

[3] We use the term *translation* to refer to words that express the same content in parallel texts.

similar variants and allows the user to navigate through the corpus by querying for one of those variants shown. The aligned words are highlighted.


*Figure 1: Screenshot of Multilingwis with translation patterns for a German support verb unit.*

## Data preparation

We extracted parallel text units from the *Corrected & Structured Europarl Corpus*[4] (Graën et al. 2014), to each of which we subsequently applied the TreeTagger for tokenization, part-of-speech tagging and lemmatization. Tagging was done with the language models available from the TreeTagger's web page[5]. We adapted the TreeTagger's tokenizer (abbreviation lexicons, punctuation) and extended its tagging lexicon (especially the German one) with lemmas and PoS tags for frequent words unknown to the TreeTagger's language models.

We assigned universal part-of-speech tags to each token using the mapping for language-specific tagsets defined by Petrov et al. (2012). Universal part-of-speech tags helped us to easily separate content words from function words across all languages. Each language comprises about 22 million content words (out of about 41 million tokens) in our data set.

For word alignment, we applied *GIZA++* (Och & Ney 2003) for each language pair and each direction, resulting in 20 sets of directed 1:n alignments of content words (only adjectives, adverbs, nouns and verbs were aligned). We symmetrized these sets by constructing the union of alignments, thus favoring recall over precision for our application.

The linguistic data obtained (tokens with lemmas and part-of-speech tags, sentence segments with their pairwise alignments and word alignments for content words for each language pair) is stored in a relational database. Database features such as *multi-column indexes*, *materialized views* and *stored procedures* allow for an efficient search and retrieval of the corpus data.

## Discussion

Every preprocessing step for our corpus data can be improved further: Our corpus still contains some misaligned text units. Subsequent sentence and word alignment cannot work for these cases since it depends on correct alignment on the text level. In sentence pairs where we align non-corresponding text, our statistical word alignment tool GIZA++ will

[4] Altogether 146,652 speech turns are available in all these five languages in CoStEP, which is based on Europarl release v7 (Koehn 2005).

[5] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/#Linux

nonetheless return the most probable alignments, which results in a long tail of incorrect translation variants that occur only once. Therefore, we currently work on filtering out sentences that are not parallel.

Our formula for ranking the matches of each corpus query currently only considers the consistent shortness of sentences across languages. Although frequent translation patterns will be shown more often than rare ones for obvious reasons, we plan to integrate the frequency of translation patterns into the ranking of the examples.

A different *Multilingwis* edition, for instance, one based on the United Nations corpus with Arabic, Chinese, English, French, Russian, and Spanish would connect less related languages in a single view. A *Multilingwis* edition with movie subtitles would be interesting for language learners.

## References


Graën, J., Batinic, D. & Volk, M., 2014. Cleaning the Europarl Corpus for Linguistic Applications. In Konvens 2014. Stiftung Universität Hildesheim.

Koehn, P., 2005. Europarl: A parallel corpus for statistical machine translation. In Machine Translation Summit. pp. 79–86.

Och, F.J. & Ney, H., 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational linguistics*, 29(1), pp.19–51.

Petrov, S., Das, D. & McDonald, R., 2012. A Universal Part-of-Speech Tagset. In *Proc LREC 2012*. pp. 2089–2096.

Volk, M., Graën, J. & Callegaro, E., 2014. Innovations in Parallel Corpus Search Tools. In Proc LREC 2014. pp. 3172–3178.