18th Workshop on Building and Using Comparable Corpora (BUCC)

Co-located with COLING 2025,
Abu Dhabi, 20 January 2025
Invited Speaker:
Preslav Nakov,
Mohamed bin Zayed University of Artificial Intelligence,
Abu Dhabi

INVITED SPEAKER

Preslav Nakov

Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi

MOTIVATION

In the language engineering and linguistics communities, research in comparable corpora has been motivated by two main reasons. In language engineering, on the one hand, it is chiefly motivated by the need to use comparable corpora as training data for statistical NLP applications such as statistical and neural machine translation or cross-lingual retrieval. In linguistics, on the other hand, comparable corpora are of interest because they enable cross-language discoveries and comparisons. It is generally accepted in both communities that comparable corpora consist of documents that are comparable in content and form in various degrees and dimensions across several languages. Parallel corpora are on the one end of this spectrum, and unrelated corpora are on the other.

In recent years, the use of comparable corpora for pre-training Large Language Models (LLMs) has led to their impressive multilingual and cross-lingual abilities, which are relevant to a range of applications, including Information Retrieval, Machine Translation, Cross-lingual text classification, etc. The linguistic definitions and observations related to comparable corpora can improve methods to mine such corpora for applications of statistical NLP, for example, to extract parallel corpora from comparable corpora for neural MT or to improve cross-lingual transfer of LLMs. As such, it is of great interest to bring together builders and users of such corpora.

TOPICS

We solicit contributions on all topics related to comparable (and parallel) corpora, including but not limited to the following:

Building Comparable Corpora:

- Automatic and semi-automatic methods
- Methods to mine parallel and non-parallel corpora from the web
- Tools and criteria to evaluate the comparability of corpora
- Parallel vs non-parallel corpora, monolingual corpora

- Rare and minority languages, across language families
- Multi-media/multi-modal comparable corpora

Applications of comparable corpora:

- Human translation
- Language learning
- Cross-language information retrieval & document categorization
- Bilingual and multilingual projections
- (Unsupervised) Machine translation
- Writing assistance
- Machine learning techniques using comparable corpora

Mining from Comparable Corpora:

- Cross-language distributional semantics, word embeddings and pre-trained multilingual transformer models
- Extraction of parallel segments or paraphrases from comparable corpora
- Methods to derive parallel from non-parallel corpora (e.g. to provide for low-resource languages in neural machine translation)
- Extraction of bilingual and multilingual translations of single words, multi-word expressions, proper names, named entities, sentences, and paraphrases from comparable corpora, etc.
- Induction of morphological, grammatical, and translation rules from comparable corpora
- Induction of multilingual word classes from comparable corpora

Comparable Corpora in the Humanities:

- Comparing linguistic phenomena across languages in contrastive linguistics
- Analyzing properties of translated language in translation studies
- Studying language change over time in diachronic linguistics
- Assigning texts to authors via authors' corpora in forensic linguistics
- Comparing rhetorical features in discourse analysis
- Studying cultural differences in sociolinguistics
- Analyzing language universals in typologicbest student paper award at LREC-COLING-2024al research

This year we will run a shared task aimed at detecting translations of terms of comparable corpora for several language pairs. We have already prepared training and testing data, as well as the protocols for evaluation of submissions. This is a timely topic as evident from the best student paper award at LREC-COLING-2024.

PRACTICAL INFORMATION

The workshop is an in-person event. Workshop registration is via the main conference registration site.

The workshop proceedings will be published in the ACL Anthology.

IMPORTANT DATES

Deadlines are "anywhere on Earth."

30 Nov 2024	Paper submission deadline
8 Dec 2024	Notification of acceptance
12 Dec 2024	Camera-ready final papers
20 Jan 2025	Workshop date

For updates, please follow the present Web page.

SUBMISSION GUIDELINES

Please follow the style sheet and templates (for LaTeX, Overleaf, and MS-Word) provided for the main conference.

Papers should be submitted as a PDF file using the START conference manager.

Submissions must describe original and unpublished work and range from 4 to 8 pages plus unlimited references.

Reviewing will be double blind, so the papers should not reveal the authors' identity. Accepted papers will be published in the workshop proceedings, which will be included in the ACL Anthology.

Double submission policy: Parallel submission to other meetings or publications is possible but must be immediately (i.e. as soon as known to the authors) notified to the workshop organizers by e-mail.

For further information and updates see the present Web page.

PDF CFP : bucc2025-cfp.pdf
Last modified: 29 Jul 2025

BUCC 2025 SHARED TASK: Bilingual Term Alignment in Comparable Corpora in English-German, -Russian, and -French

The BUCC 2025 shared task addresses multilingual terminology alignment in comparable corpora. It expands on the BUCC 2022 and 2023 datasets by adding test data in English-German and English-Russian.

Many research groups are working on this problem using a wide variety of approaches. However, as there is no standard way to measure the performance of the systems, the published results are not comparable and the pros and cons of the various approaches are not clear. The shared task aims at solving these problems by organizing a fair comparison of systems. This is accomplished by providing corpora and evaluation datasets for a number of language pairs and domains.

Moreover, the importance of dealing with multi-word expressions in Natural Language Processing applications has been recognized for a long time. In particular, multi-word expressions pose serious challenges for machine translation systems because of their syntactic and semantic properties. Furthermore, multi-word expressions tend to be more frequent in domain-specific text, hence the need to handle them in tasks with specialized-domain corpora.

Through the 2025 BUCC shared task, we seek to evaluate methods that detect pairs of terms that are translations of each other in two comparable corpora, with an emphasis on multi-word terms in specialized domains.

Provided resources

The BUCC shared task provides several datasets of the following form:

- A pair of comparable corpora C_1 and C_2 in languages L_1 and L_2.
- A list of terms D_1 that occur in C_1 and a list of terms D_2 that occur in C_2. Term lists may include both single-word and multi-word terms.

- For training only, a gold standard dictionary $D_{1,2}$ in the form of a list of pairs of terms (t_1, t_2) that are translations of each other, with t_1 in D_1 and t_2 in D_2.

The task participants may additionally use any external resources, except the CCAligned corpora, from which the task datasets have been extracted. When reporting their results, participants are required to specify which resources they used. They are also encouraged to test conditions in which they only use the provided resources.

Task

Given a test dataset with comparable corpora C_1 and C_2, and lists of terms D_1 and D_2, participant systems are expected to produce an ordered list of term pairs in (D_1, D_2) that are translations of each other, in descending order of confidence.

Note that D_1 and D_2 may have different sizes, that not every term in D_1 may have a translation in D_2, that some terms in D_1 might have multiple translations, and conversely. For practical reasons, we limit the length of a submitted term pair list to a ceiling of 10 times the average length of D_1 and D_2. (This can be seen as meaning that, on average, a system may submit up to 10 alignment hypotheses for each term in D_1 or in D_2.)

The test datasets will include both the same language pairs as those provided for training and also other language pairs.

Participants can submit up to 5 system runs for each test dataset.

Evaluation

The evaluation metric will be the Average Precision of the predicted bilingual term pair list, where the relevance of a term pair is determined by its presence in the (hidden) gold standard dictionary $D_{1,2}$. This models the task as an information retrieval task: retrieve all relevant term pairs (t_1, t_2) (documents) from the cross-product $D_1 \times D_2$ (virtual pool of documents), presenting them in descending order of confidence. Average Precision is the area under the recall $\times$ precision curve. It is computed as the average over all m relevant term pairs (t_i, t_j) (i.e., all term pairs in the gold standard) of the precision value obtained for the set of top n_k term pairs existing after each relevant term pair (t_{i_k}, t_{j_k}) is retrieved, from the first to the last relevant term pair. Relevant term pairs that are not retrieved receive a precision of zero, hence decrease Average Precision. Average Precision (AP) is defined as:

$$AP = \frac{1}{m} \sum_{k=1}^{m} P(R_k)$$

where R_k is the set of ranked predicted term pairs from the top to the position at which k relevant term pairs have been retrieved. Given the gold standard dictionary $D_{1,2}$, the precision of a set of predicted term pairs R is defined as $P(R) = \frac{|R \cap D_{1,2}|}{|R|}$. Evaluation code is provided on github.

Helper note: To optimize Average Precision, a system must find all relevant term pairs and put them at the top of the list. Average Precision increases when true predictions (relevant term pairs) are added anywhere in the prediction list. Average Precision also increases when false predictions, if any, are pushed towards the bottom of the list. Note that Average Precision cannot decrease when more predictions, whether true or false, are added to the bottom of the list. Also note that Average Precision is equivalent to Mean Average Precision (MAP) with exactly one query:

Q: find all term pairs in $D_1 \times D_2$ that are translations of each other

The present evaluation therefore does not model the task with a query per source term, but with one global query (Q above) that considers all source terms together. The systems are thus expected to rank all their chosen (t_1, t_2) candidate term pairs in descending order of confidence. To state it another way, a system should *not* rank term pairs first according to the source term t_1, then in descending order of confidence.

File format

All files use UTF-8 encoding, with LF end-of-line markers.

- Single-term lists D_1 and D_2 contain one term per line.
- Corpora C_1 and C_2 contain one sentence per line.
- The gold standard dictionary $D_{1,2}$ contains two terms per line, separated by a tabulation: t_1<TAB>t_2
- The system output submitted by a participant contains two terms per line, separated by a tabulation <TAB>. Its lines are ordered in decreasing order of confidence.

Sample data

A small sample dataset is provided in bucc2023_sample.zip for the English-French language pair. It contains:

- A pair of comparable corpora C_1 =*src_corpus_sample.txt* and C_2 =*tgt_corpus_sample.txt* in languages L_1 =*en* and L_2 =*fr*.
- A list of terms D_1 =*src_term_list_sample.txt* that occur in C_1 and a list of terms D_2 =*tgt_term_list_sample.txt* that occur in C_2. Term lists may include both single-word and multi-word terms.
- For training only, a gold standard dictionary $D_{1,2}$ =*gold_dictionary_sample.txt* in *en*×*fr*.

Training data

A training dataset is provided in bucc2023_training.zip for the English-French language pair. It contains:

- A pair of comparable corpora C_1 =*corpus-en.txt* and C_2 =*corpus-fr.txt* in languages L_1 =*en* and L_2 =*fr*.
- A list of terms D_1 =*terms-en.txt* that occur in C_1 and a list of terms D_2 =*terms-fr.txt* that occur in C_2. Term lists may include both single-word and multi-word terms.
- For training only, a gold standard dictionary $D_{1,2}$ =*terms-en-fr.txt* in *en*×*fr*.

Note that the sizes of D_1, D_2 and $D_{1,2}$as well as the proportions of terms in D_1or D_2 that have a translation in $D_{1,2}$ are likely to be different in the test datasets.

Test data

Test datasets in three language pairs (en-fr, en-de, en-ru). A test dataset contains:

- A pair of comparable corpora C_1 =*corpus-en.txt* and C_2 =*corpus-fr.txt* in languages L_1 =*en* and L_2 =*fr*.
- A list of terms D_1 that occur in C_1 and a list of terms D_2 that occur in C_2. Term lists may include both single-word and multi-word terms.

Time schedule

Any time	Expression of interest to the shared task contact point (see below). This will allow us to register you o
20 July 2024	Sample dataset and training data are available for the English-French language pair (en-fr)
tbd	Test data release
tbd	Submission of system runs by participants (up to 5 per dataset) by e-mail to the shared task contact po
30 Nov 2024	Submission of shared task papers to the BUCC workshop

Shared task organizers and contact

Pierre Zweigenbaum (Université Paris-Saclay, CNRS, LISN, Orsay, France)

Serge Sharoff (University of Leeds, United Kingdom)

Ayla Rigouts Terryn (Université de Montréal, Mila, Canada)

Reinhard Rapp (University of Mainz, Germany)

Shared task contact: please send expressions of interest to:

- pz (at) lisn (dot) fr

Last modified: 29 Jul 2024

ORGANIZERS AND CONTACT

Serge Sharoff (University of Leeds, United Kingdom)

Ayla Rigouts Terryn (Université de Montréal, Mila, Canada)

Pierre Zweigenbaum (Université Paris-Saclay, CNRS, LISN, Orsay, France)

Reinhard Rapp (University of Mainz, Germany)

Contact for workshop: s (dot) sharoff (at) leeds (dot) ac (dot) uk

PROGRAMME COMMITTEE

- Ebrahim Ansari (Institute for Advanced Studies in Basic Sciences, Iran)
- Eleftherios Avramidis (DFKI, Germany)
- Gabriel Bernier-Colborne (National Research Council, Canada)
- Thierry Etchegoyhen (Vicomtech, Spain)
- Alex Fraser (University of Munich, Germany)
- Natalia Grabar (University of Lille, France)
- Amal Haddad Haddad (Universidad de Granada, Spain)
- Amir Hazem (University of Tokyo, Japan)
- Kyo Kageura (University of Tokyo, Japan)
- Natalie Kübler (Université Paris Cité, France)
- Philippe Langlais (Université de Montréal, Canada)
- Yves Lepage (Waseda University, Japan)
- Shervin Malmasi (Amazon, USA)
- Michael Mohler (Language Computer Corporation, USA)
- Emmanuel Morin (Nantes Université, France)

- Dragos Stefan Munteanu (RWS, USA)
- Ted Pedersen (University of Minnesota, Duluth, USA)
- Nasredine Semmar (CEA LIST, Paris, France)
- Silvia Severini (Leonardo Labs, Italy)
- Richard Sproat (Google, USA)
- Marko Tadić (University of Zagreb, Croatia)
- François Yvon (Sorbonne Université, France)

Last modified: 17 Jul 2024