NLP Corpus Observatory – Looking for Constellations in Parallel Corpora to Improve Learners’ Collocational Skills

Gerold Schneider
English Department &
Institute of Computational Linguistics
University of Zurich
gschneid@ifi.uzh.ch

Johannes Graën
Institute of Computational Linguistics
University of Zurich
graen@cl.uzh.ch

Abstract

The use of corpora in language learning, both in classroom and self-study situations, has proven useful. Investigations into technology use show a benefit for learners that are able to work with corpus data using easily accessible technology. But relatively little work has been done on exploring the possibilities of parallel corpora for language learning applications.

Our work described in this paper explores the applicability of a parallel corpus enhanced with several layers generated by NLP techniques for extracting collocations that are non-compositional and thus indispensable to learn. We identify constellations, i.e. combinations of intra- and interlingual relations, calculate association scores on each relation and, based thereon, a joint score for each constellation. This way, we are able to find relevant collocations for different types of constellations.

We evaluate our approach and discuss scenarios in which language learners can playfully explore collocations. Our explorative web tool is freely accessible, generates collocation dictionaries on the fly, and links them to example sentences to ensure context embedding.

1 Introduction

Parallel corpora show a great potential for language learning, as they allow one to zoom into those areas where the linguistic differences between the native language and the target language are largest.

Data-driven Learning (DDL), although sometimes seen as either too complicated for learners (Hadley and Charles 2017), or furnishing texts of too high levels (Vyatkina and Boulton 2017), can benefit advanced learners, and even beginners, and also using very basic tools such as concordancers, as e.g. St. John (2001) describes for lexical tasks, Chujo et al. (2016) for grammatical tasks, and Vyatkina (2016) for collocations.

There are ample studies on creating corpus-informed teaching materials, for example dictionaries of collocations (Ackermann and Chen 2013; Durrant 2009; McGee 2012). The advantage of this approach is that students do not need to learn to use corpus interfaces. The disadvantage is that contextualisation is limited. Li (2017) shows that also direct corpus use improves learner competence in the area of collocations. They conclude that “[t]his exposure to attested language data raises learners’ awareness of using collocations in a more natural or near-native way … it would be beneficial for more researchers and teachers to investigate direct corpus applications in classroom settings.” (p. 165)

Ultimately, we need both corpus-derived teaching material and the direct corpus experience linked to it. Buyse and Verlinde (2013) show that using corpus-derived, contextualised resources (Linguee) led to better test performance and user satisfaction. They suggest that a further integration of tools would be desirable, allowing students to combine the immersion experience which Linguee offers and profit from abstracted customised resources such as collocation dictionaries.

The suggested integration involves using parallel corpora, like Linguee does, but deriving patterns that are particularly challenging for language learners from them, thus creating a registry of lexical-grammatical phenomena on which learners are likely to experience difficulties because literal translations do not suffice. The desired integration also requires linking the derived patterns back to the test, furnishing contextualised examples. We would like to contribute to this integration with our contribution.

Gerold Schneider and Johannes Graën 2018. NLP Corpus Observatory – Looking for constellations in parallel corpora to improve learners’ collocational skills. Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning at SLTC 2018 (NLP4CALL 2018). Linköping Electronic Conference Proceedings 152: 69–78.
In order not to start with preconceptions, we use as few initial constraints as possible, and let the data point out areas of linguistic contrast. We focus on English compared to Swedish, using four constructions: adjective-noun, verb-preposition, verb-object and verb-preposition-object. Namvar (2012) investigates nine constructions. The results show that verb-object collocations are most frequent in learner writing, followed by verb-preposition collocations. Källkvist (1998) observed that awkward collocations produced by advanced Swedish learners of English often involve an incorrect use of verbs. Verb-preposition constructions are particularly difficult to acquire for language learners (Gilquin and Granger 2011, pp. 59–60). Phrasal verbs represent “one of the most notoriously challenging aspects of English language instruction” (Gardner and Davies 2007, p. 339). Vyatkina (2016) shows particularly good results for learning German verb-particle structures with data-driven learning.

We go beyond purely collocation-based phrase search, in the following ways: first, while collocations do not entail non-compositionality, the fact that we need to reach collocational status in both languages leads to cleaner results, as in a double check. Secondly, by punishing literal translations, we also filter the majority of instances that are compositional cooccurrences.

In the following, we present a method to explore constellations in parallel corpora. We then present our interactive and explorative web tool, which creates collocation dictionaries on the fly (indirect DDL) based on association scores, and links the dictionary entries to the parallel corpus examples (direct DDL). Users can explore and tailor the association metrics to their needs.

2 Related Work

The bilingual concordancers Glosbe,\(^1\) Linguee,\(^2\) Tradooit\(^3\) and our multilingual Multilingwis\(^4\) (Clematide, Graë\^n, and Volk 2016; Graë\^n and Clematide 2015; Graë\^n, Sandow, and Volk 2017) are web applications which allow translators and advanced learners to explore and compare translation variants (for an overview, see Volk, Graë\^n, and Callegaro 2014). No resources such as lists of phrases and collocations for the benefit of learners are automatically derived, however.

There is a long tradition of research in the area of phrasemes (Mel’čuk 1998; Wanner 1996). Collocations measures have been explored systematically (Evert 2004, 2008; Pecina 2009; Church and Hanks 1990) but it is unclear which measures are better suitable for the benefit of language learners. Huang et al. (2013) present a tool which allows learners to explore collocations using a variety of measures, but the results do not profit from parallel data, e.g. they are not weighted according to translation difficulty, as we intend to do.

To our knowledge, there has been no approach so far where data-driven NLP methods on parallel corpora are used for collocation retrieval for the benefit of language learning. Chujo et al. (2016) is partly similar to our approach. They compare a direct DDL tool in the form of a KWIC concordancer, and a separate indirect tool in the form of a word profiler. The word profiler delivers collocations once the user suggests a node. Our approach is more data-driven, as we assume no given nodes but generate results purely from the parallel corpora, and we fully integrate both into one tool, linking the lists of collocations to the examples in the parallel corpus.

In Graë\^n and Schneider (2017), we describe an approach where word lists are based on parallel corpora, but we restricted our research to the fixed frame of verb-preposition structures, and did not link the lists back to the corpora.

3 Data and Methods

The basis of our experiments is our FEP9 corpus (Graë\^n 2018), which comprises different layers of annotation (part-of-speech tags, lemmas, syntactic dependency relations) and alignment (sentence and word alignment) on top of the cleaned Europarl corpus (Graë\^n, Batinic, and Volk 2014). Europarl (Koehn 2005) consists of the transcribed and translated debates of the European Parliament over a period of 15 years.

From this corpus, we randomly sample a subset of 5% of parallel texts (contributions of individual speakers in Europarl) in English and Swedish. We filter word alignments for those, where three word aligners agree, namely GIZA++ (Och and Ney 2003), the Berkeley Aligner (Liang, Taskar, and Klein 2006) and efimaral (Östling and Tiede-
The fourth word aligner available in FEP9, fast_align (Dyer, Chahuneau, and Smith 2013) performs considerably inferior to the other aligners (see Graën 2018, Figure 4.21) and we therefore disregard its alignments. In total, our data set comprises 160 thousand sentence and 2.4 million word alignments.

We count cooccurrence frequencies on syntactic relationships (for each dependency label) and word alignments, both mapped to the respective lemmas in each language. Assuming the independence of two events (i.e. lemmas) observed together in either syntactical (interlingual) or word-correspondence relation (intralingual), we calculate the expected frequency of each lemma pair. Statistical association measures (see Evert 2004, 2008, for an overview) relate the observed frequency (O) to the expected frequency (E) and provide a ranking for a list of cooccurring events. Some association measure yield scores that have an information theoretic interpretation (Evert 2004, Section 3.1.7), but the scores of most measures need to be interpreted in comparison among themselves.

![Figure 1: A constellation consisting of two aligned verbs with corresponding aligned prepositions.](image)

Interlingual association measures, that is, the application of well-known association measures, which are frequently used to identify collocations in monolingual corpora, to parallel, word-aligned corpora are first described in (Graën n.d.). Our idea is to combine relations from syntactic analysis with word correspondence (i.e., the output of parsing and word alignment techniques) to find parallel patterns in two languages, which we call constellations. Figure 1 shows an example of parallel verb-preposition structures. Due to their complex structure (syntactic relations in both languages plus word alignment between the two), constellations are more error-prone than monolingual patterns (ibid., Section 4.2). However, the lowest possible threshold of two already suffices to filter out most errors, since systematic errors would need to coincide on the different levels, which is very rare.

We also present an interactive interface that facilitates the exploration of different association measures on different relations (Graën and Bless 2017). Based on a list of verbs and their direct objects, the user chooses one of five “simple association measures” presented in (Evert 2008, Chapter 4) or the absolute frequency for ranking verb-object pairs. On the source language side (English, German or Italian), the association score is either calculated on the syntactic relation between verb and object or one of their alignment relations. This limitation to the original idea of combining association scores on all relations to a single constellation score sketched in (Graën n.d.) is what we address in this work. In addition to support verb constructions with direct objects, we also define constellations for support verb constructions with prepositional objects (see, for instance, Figure 2), adjectival modifiers of noun and verb-preposition combinations.

![Figure 2: A constellation consisting of two aligned verbs with corresponding aligned prepositions and aligned prepositional objects.](image)

In this work, we implement the idea of free combinations of association scores on different relations. Our objective is to identify non-compositional expressions, such as support verb construction, that a language learner is required to learn by heart. Translation difficulties arise particularly frequently wherever non-compositionality is involved, that is, wherever literal translations
lead to incorrect or non-nativelike expressions. Non-compositional features include any form of idiom and collocation, as for example phrasal verbs, support verb constructions and technical terms. We hence combine the parallel search for phrasemes in both languages with word correspondences in the form of alignments.

Retrieval of the constellations from our corpus is similar to the retrieval performed in (Graën n.d.), but we expect that our data holds more reliable word alignments, as they are obtained by agreement of three different word aligners instead of a single one. From the observed (O) and expected (E) cooccurrence frequencies, we calculate the respective association score for each relation. To make different association measures for syntactic dependency relations and word alignment comparable, we normalize all association scores to values between 0 and 1.

A straightforward way to do so is to linearly project all positive association scores to the range from 0 to 1: \( \text{score}_\text{norm} = \frac{\text{score}}{\max(\text{score})} \). If the maximum association score is attained by an outlier (some association measures favour rare combinations (see Graën 2018, Figure 5.4)), all association scores of the relation in question are penalised.

Another way to normalize values is to use the tangens hyperbolicus: \( \text{score}_\text{norm} = 1 - \frac{2}{1 + e^{\text{score}}} \). Some association measures yield high values that, after being normalized with the tangens hyperbolicus, are indistinguishably close to 1. We therefore propose to apply two subsequent normalisations: first to divide by the average score to obtain a distribution around 1 (\( \text{score}_\emptyset = \text{score}/\text{score} \)), and second to apply the tangens hyperbolicus to the resulting normalized scores:

\[
\text{score}_\text{norm} = 1 - \frac{2}{1 + e^{\text{score}_\emptyset}}
\]

Our application allows for experimenting with these three normalizations, as well as different association measures for syntactical and word-correspondence relations. The formula for the final score of a particular constellation example can be any mathematical operation on the respective association scores and the raw frequency. As we expect an element of surprise in the correspondence of expressions in both languages, we use the association score on one of the word correspondence relations to downgrade the final score. In the case of support verb constructions, we prefer verb pairs that are not used frequently as translations. The combinations of association measures that worked best for the respective constellations are explained in Section 4.

We facilitate the memorisation of those expressions by providing authentic parallel corpus examples. The example list comprises all examples from our small corpus subset ordered by number of tokens in both sentences of the respective example (longer sentences are supposed to be more difficult to capture) and the difference in number of tokens between both languages. We expect the latter number to differ since English sentences comprises relatively more tokens than Swedish sentences, but an overly large number typically originates from a non 1-to-1 sentence alignment or untranslated parts in one of the languages. We have considered adding other measures, such as syntactic complexity or variation in alignment, but a length-based sorting already yields satisfactory results. Short sentences allow the user to concentrate on the constellation in context, while long sentences offer so much context that users easily get distracted.

4 Results

Best-scoring results for three different constellations consisting of four tokens are shown on page 7 ff. On page 8, we list the best results for a constellation of six tokens (verbs with prepositional objects). Users can interactively change the collocation formula that are used in our experimental application.

4.1 Adjective-Noun Collocations

For adjective-noun collocations, we show the following formula:

\[
\text{score} = a_{s1}^2 \cdot a_{s4}^4 \cdot \frac{a_{s3}^3}{(a_{s2}^4)^2}
\]

The score consists of the linear combination of the association score between adjective and noun in English (\( a_{s1}^2 \)) and Swedish (\( a_{s1}^4 \)), and the association score of the alignment between the nouns (\( a_{s3}^3 \)), divided by the squared association score of the alignment of the adjectives (\( a_{s2}^4 \)). This formula has the effect that associations from both languages are reported, particularly those in which the noun is a literal translation, but the adjective is non-literal: the fact that adjective alignment association scores are used in the denominator assures that generally unlikely translations are preferred.
When does the Council intend to reach a decision on the establishment of this future observatory? När kommer rådet att fattat beslut om att inrätta detta framtida organ?

It has attempted to reallocate budgetary resources from the Progress programme to the microfinance facility before the European Parliament has reached a decision. Den har försökt omfördela budgetresurser från Progressprogrammet till instrumentet för mikrokreditern innan Europaparlamentet har fattat ett beslut.

Furthermore, the decision-making process itself can be unclear, as the convention submits proposals and the Intergovernmental Conference has to reach decisions. Dessutom kan det bli oklart kring själva beslutsfattandet, eftersom konventet lägger fram förslag och regeringskonferensen måste fatta beslut.

When the matter comes before Parliament, therefore, we often have to reach our decisions very quickly if we want to make the internal market a reality for the citizens of Europe. Kommer ärendet sedan till parlamentet, måste vi ofta fatta mycket snabba beslut, eftersom vi vill öppna den gemensamma marknaden för medborgarna.

With regard to the forestry strategy of the Community in general, and in particular the question whether forestry activities should be governed by Community legislation, the Commission will also shortly reach a decision on such a forestry strategy, which will likewise be communicated to Parliament. Beträffande gemenskapens beskogningsstrategi, i synnerhet frågan om gemenskapsrättsliga bestämmelser för skogsbruket, kommer kommissionen snart att fatta beslut om en beskogningsstrategi och informera parlamentet om detta.

In reaching its decision it concluded after prolonged debate, in the presence of Mr Le Pen and colleagues of his who were there to support him, that the legitimate procedure had been complied with in every respect and that no breach of the basic rule establishing parliamentary immunity had taken place, so that the Member was free to carry out his duties while at the same time the institution of Parliament was not being undermined. För att fatta sitt beslut drog det, efter långvarig diskussion där även Le Pen och kolleger som stöder honom var närvarande, slutsatsen att det juridiska förfarandet var absolut korrekt, så att inget brott begås mot grundregeln som fastställer parlamentarisk immunitet, för att ledamoten skall kunna utöva sina plikter oberoende utan att den parlamentariska grundregeln samtidigt undermineras.

Table 1: Examples for the verb-direct object constellation “reach decision”/”fatta beslut” ordered by increasing length and minimal length difference. Example 2 shows a direct translation, sentence 4 shows adjective to adverb variation, sentence 6 an English continuous form.

The 80-best list illustrates that, for example, stor uppmärksamhet corresponds to English great attention, where the noun is a direct translation, but the adjective is non-literal. Swedish native speakers learning English can thus see that close attention is a more native-like translation than great attention or even big attention. In the opposite direction, English speakers learning Swedish can equally see that stor uppmärksamhet is a more native-like translation than nära uppmärksamhet.

Clicking on the results displays example sentences sorted by estimated complexity, which helps learners to contextualise idiomatic and collocational expression. We show the example of “reach decision” corresponding to “fatta beslut” (row 4) in Table 1.

4.2 Verb-Object Collocations

For verb-object collocations, we show the following formula:

\[ score = a_s^2 \cdot a_s^4 \cdot \frac{a_s^4}{(a_s^1)^2} \cdot \text{freq} \]

The formula is similar to the one used for adjective-nouns, this time punishing direct translation of verbs, with the difference that frequency is also used. Frequency is an important factor for the identification of light verb constructions (Ronan and Schneider 2015). Swedish learners of English can see in the table of results that have de-
bate is a more native-like translation than lead debate, the literal translation, while English learners of Swedish can e.g. see that nömla exempel or ta exempel is often preferable to the direct translation of ge exempel. A further small difference is that we have used the t-score association metric here, while z-score was used for adjective-noun collocations.

The squared association between the verbs has the effect of slightly exaggerating the urge to find verbal differences: the list gives both have responsibility as translation of bära ansvar (rank 2), as well as bear responsibility as translation of ha ansvar (rank 175, off the short top of the list). Users can thus experiment with less strong punishment for verb-verb alignment and again inspect the examples, and equally explore a range of association metrics. The interface allows users to interactively and playfully explore native speaker associations.

4.3 Verb-Preposition Collocations

Next, we focus on verb-preposition and phrasal verb constructions. The formula shown here is identical to the one for verb-object, this time punishing direct translations of prepositions (as \( s_2 \) in the denominator is the score calculated on the alignment of the two prepositions):

\[
\text{score} = \frac{as_1^2 \cdot as_3^4 \cdot \frac{as_3^3}{(as_2^3)^2} \cdot freq}
\]

We can see e.g. that congratulate on is a more native-like translation of Swedish gratulera till than the direct translation congratulate to.

4.4 Verb-Preposition-Object Constellations

Finally, we give an example of a construction involving more than two words: verb-PP constructions where the noun in the PP is also idiomatic.

\[
\text{score} = \frac{as_1^2 \cdot as_2^3 \cdot as_3^4 \cdot as_5^6 \cdot \frac{as_5^6}{(as_1^3)^2} \cdot freq}
\]

The formula that we illustrate here combines positive association between all the elements except for the verb alignment (as \( s_1^3 \)), where negative association, i.e. non-direct translation is sought for. English learners of Swedish can detect that the idiomatic translation of come into force is träda i kraft. We also notice that the Swedish lemmatizer is producing a systematic error by lemmatizing the supine form trätt to träta ‘to quarrel’ instead of träda ‘to step’.

5 Evaluation

While the lists presented in Section 4 may look intuitively convincing, the question arises up to which point learners fail to produce the collocations suggested in the lists, and instead produce direct translations, influenced by L1 transfer. We thus address the question if learners actually produce the awkward collocations that the list suggests. As test case, we assume a situation in which a native speaker of Swedish is producing English collocations. The question is whether his or her collocations are less native-like than those in a reference corpus of native speakers. We use the ICLE corpus (Granger et al. 2009) as learner corpus to assess if the level of these awkward collocations is higher in than in a native speaker corpus, for which we use the BNC (Aston and Burnard 1998).

The picture is complicated by several facts. First, the awkward collocations are all correct, and also found in the BNC, but typically with a slight meaning shift, and not as the major variant. We are thus addressing the question if the suggested English collocation is more dominant in the native than in the learner texts. Second, due to sparse data reasons, we had to include all learners, irrespective of their native language. Third, ICLE contains data of University students, advanced learners who chose native-like collocations in the majority of cases.

We evaluate adjective-noun structures, in the two following ways. First, for all cases where

- the Swedish adjective has a direct translation,
- one that is different from the one suggested in the collocation under observation,
- but semantically similar to the English one in the list,
- the translation of the noun is direct,
- whenever we have at least 3 hits in ICLE in total (maximally one zero count in any cell is replaced by a smoothing count of 0.1)

then we compare the numbers.

For example, stor uppmärksamhet \((t_4, t_5)\) could be directly translated to English great attention \((t'_2, t'_3)\), but the suggested English collocation is close attention \((t_2, t_1)\). close attention occurs 106 times in the BNC, great attention only 47 times, the ratio \(r_{BNC} = t_2/t'_2 = 2.25\). In ICLE, great attention occurs 9 times, while close attention occurs twice,
Verb-Preposition-Noun Constellations (4.4)

| no. | t₁ (verb on) | t₂ (prep) | t₃ (noun) | t₄ (verb) | t₅ (prep) | t₆ (noun) | freq. | α₁² | α₂² | α₃² | score |
|-----|--------------|-----------|-----------|-----------|-----------|-----------|-------|-----|-----|-----|-------|
| 1   | vote         | for       | report    | rösta     | för       | betänkande| 54    | 1.0000| 1.0000| 1.0000| 54.000|
| 2   | enter        | into      | force     | träda     | i         | kraft     | 31    | 0.9958| 1.0000| 1.0000| 31.258|
| 3   | thank        | for       | work      | tacka     | för       | arbete    | 31    | 1.0000| 1.0000| 1.0000| 31.000|
| 4   | be           | in        | interest  | ligga     | i         | intresse  | 29    | 1.0000| 1.0000| 1.0000| 28.999|
| 5   | thank        | for       | report    | tacka     | för       | betänkande| 25    | 1.0000| 1.0000| 1.0000| 25.000|
| 6   | be           | of        | importance| vara      | av        | betydelse | 25    | 1.0000| 1.0000| 1.0000| 25.000|
| 7   | congratulate  | on        | report    | gratulera | till      | betänkande| 18    | 1.0000| 1.0000| 1.0000| 18.000|
| 8   | vote         | against   | report    | rösta     | mot       | betänkande| 18    | 1.0000| 1.0000| 1.0000| 17.971|
| 9   | speak        | with      | voice     | tal    | med       | röst      | 18    | 1.0000| 1.0000| 0.9998| 17.825|
| 10  | come         | from      | country   | komma    | från      | land      | 18    | 1.0000| 1.0000| 1.0000| 16.000|
| 11  | vote         | for       | resolution| rösta    | för       | resolution| 16    | 1.0000| 1.0000| 1.0000| 15.987|
| 12  | thank        | for       | cooperation| tacka    | för       | samarbete | 15    | 1.0000| 1.0000| 1.0000| 15.000|
| 13  | be           | of        | importance| vara      | av        | vikt      | 15    | 1.0000| 1.0000| 1.0000| 15.000|
| 14  | be           | at        | stake     | stå      | på        | spel      | 13    | 1.0000| 1.0000| 0.9866| 12.824|
| 15  | come         | into      | force     | träda    | i         | kraft     | 12    | 0.9865| 1.0000| 1.0000| 12.329|
| 16  | participate  | in        | debate    | delta    | i         | debatt    | 12    | 1.0000| 1.0000| 0.9996| 12.000|
| 17  | take         | on        | Thursday  | äga      | på        | torsdag   | 12    | 1.0000| 1.0000| 0.9996| 11.856|
| 18  | go           | in        | hand      | gå       | i         | hand      | 11    | 1.0000| 1.0000| 0.9999| 10.999|
| 19  | thank        | for       | support   | tacka    | för       | stöd      | 11    | 1.0000| 1.0000| 1.0000| 10.997|
| 20  | come         | into      | force     | träda    | i         | kraft     | 9     | 0.9300| 1.0000| 1.0000| 10.280|
| 21  | propose       | by        | Commission| föreslå   | av        | kommission| 10    | 1.0000| 1.0000| 1.0000| 9.971|
| 22  | be           | in        | situation | befina   | i         | situation | 9     | 1.0000| 1.0000| 1.0000| 9.001|
| 23  | adopt         | by        | Committee | anta     | av        | utskott   | 9     | 1.0000| 1.0000| 1.0000| 8.997|
| 24  | contribute    | to        | development| bidra    | till      | utveckling| 9     | 1.0000| 1.0000| 1.0000| 8.996|
| 25  | be           | in        | line      | ligga    | i         | linje     | 9     | 1.0000| 1.0000| 0.9998| 8.995|

Table 2: Evaluation of adjective-noun constellations

\[
r_{BNC} = \frac{t_2}{t_3} = 0.22. \quad r_{ICL}^{ICL} \quad (r, \text{last column}) \text{ is then } 10.15, \text{which can be interpreted as relative dominance, expressing that the suggested collocation is 10.15 times more dominant in the BNC than in ICLE.} \]

Second, we measure the absolute dominance of the English collocation, as follows: the frequency of the collocation, divided by the frequency of the noun modified by any adjective. For close attention in the BNC, this is \( \text{dom}(\text{BNC}) = \frac{106}{4805} = 0.022 \), in ICLE it is \( \text{dom}(\text{ICLE}) = \frac{76}{286} = 0.007 \). \( \text{dom}(\text{BNC})/\text{dom}(\text{ICLE}) \) is thus 3.15. The mean of the absolute dominance is 5.3, which means that the suggested collocation is found 5.3 times more often in BNC than in ICLE.

The evaluation has shown that in the majority of cases, our method yields good results, and allows learners to explore various constellations.

6 Conclusions and Future Work

We have implemented and evaluated an interactive tool for data-driven learning of constellations (i.e., parallel collocation structures) in which language learners experience particular difficulties. Our system features full integration of direct and indirect data-driven learning. Collocation dictionaries are generated on the fly, and linked to the parallel examples in the aligned corpus to ensure contextualisation. Our approach is based on the use of association measures of collocations and of alignments. Advanced users can also customise the association scores.

As future steps, we plan to test the tool with learners, to train on the entire Europarl corpus, and to add more languages to our approach.

https://pub.cl.uzh.ch/purl/constellations
References

Ackermann, K. and Y. H. Chen (2013). “Developing the Academic Collocation List (ACL): A corpus-driven and expert-judged approach.” In: Journal of English for Academic Purposes 12.4, pp. 235–247.

Aston, G. and L. Burnard (1998). The BNC Handbook. Exploring the British National Corpus with SARA. Edinburgh: Edinburgh University Press.

Buyse, K. and S. Verlinde (2013). “Possible effects of free on line data driven lexicographic instruments on foreign language learning: The case of Linguee and the interactive language toolbox”. In: Procedia-Social and Behavioral Sciences 95, pp. 507–512.

Chujo, K., Y. Kobayashi, A. Mizumoto, and K. Oghigian (2016). “Exploring the Effectiveness of Combined Web-based Corpus Tools for Beginner”. In: Linguistics and Literature Studies 4.4, pp. 262–274.

Church, K. W. and P. Hanks (1990). “Word Association Norms, Mutual Information, and Lexicography”. In: Computational Linguistics 16.1, pp. 22–29.

Clematide, S., J. Graën, and M. Volk (2016). “Multilingwis – A Multilingual Search Tool for Multi-Word Units in Multiparallel Corpora”. In: Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives – Fraseología computacional y basada en corpus: perspectivas monolingües y multilingües. Ed. by G. C. Pastor. Geneva: Traduxel, pp. 447–455.

Durrant, P. (2009). “Investigating the viability of a collocation list for students of English for academic purposes”. In: English for Specific Purposes 28.3, pp. 157–169.

Dyer, C., V. Chahuneau, and N. A. Smith (2013). “A Simple, Fast, and Effective Reparameterization of IBM Model 2”. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Association for Computational Linguistics (ACL), pp. 644–649.

Evert, S. (2004). “The Statistics of Word Cooccurrence: Word Pairs and Collocations”. PhD thesis. University of Stuttgart.

Graën, J. (n.d.). “Identifying Phrasemes via Interlingual Association Measures”. In: Lexemkombinationen und typisierte Rede im mehrsprachigen Kontext. Ed. by C. Konecny, E. Autelli, A. Abel, and L. Zanasi. Tübingen: Stauffenburg Linguistik. In press.

– (2018). “Exploiting Alignment in Multiparallel Corpora for Applications in Linguistics and Language Learning”. PhD thesis. University of Zurich. In press.

Graën, J., D. Batinic, and M. Volk (Oct. 2014). “Cleaning the Europarl Corpus for Linguistic Applications”. In: Proceedings of the Conference on Natural Language Processing (KONVENS) (Hildesheim). Stiftung Universität Hildesheim, pp. 222–227.

Graën, J. and C. Bless (2017). “Exploring Properties of Intra- and Interlingual Association Measures Visually”. In: Proceedings of the 21st Nordic Conference of Computational Linguistics (NODALIDA). Linköping Electronic Conference Proceedings 131. Linköping University Electronic Press, Linköpings universitet, pp. 314–317.

Graën, J. and S. Clematide (2015). “Challenges in the Alignment, Management and Exploitation of Large and Richly Annotated Multi-Parallel Corpora”. In: Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC) (Lancaster). Ed. by P. Bański, H. Biber, et al., pp. 15–20.

Graën, J., D. Sandoz, and M. Volk (2017). “Multilingwis² – Explore Your Parallel Corpus”. In: Proceedings of the 21st Nordic Conference of Computational Linguistics (NODALIDA). Linköping Electronic Conference Pro-

A. Lüdeling and M. Kytö. Vol. 2. Berlin: Walter de Gruyter, pp. 1212–1248.

Gardner, D. and M. Davies (2007). “Pointing out frequent phrasal verbs: A corpus-based analysis”. In: TESOL Quarterly: A Journal for Teachers of English to Speakers of Other Languages and of Standard English as a Second Dialect 41.2, pp. 339–359.

Gilquin, G. and S. Granger (2011). “From EFL to ESL: Evidence from the International Corpus of Learner English”. In: Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap. Ed. by J. Mukherjee and M. Hundt. Amsterdam: John Benjamins, pp. 55–78.
ceedings 131. Linköping University Electronic Press, Linköpings universitet, pp. 247–250.

Graën, J. and G. Schneider (2017). “Crossing the Border Twice: Reimporting Prepositions to Alleviate L1-Specific Transfer Errors”. In: Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning & 2nd Workshop on NLP for Research on Language Acquisition. Linköping Electronic Conference Proceedings 134. Linköpings universitet Electronic Press, pp. 18–26.

Granger, S., E. Dagneaux, F. Meunier, and M. Paquot (2009). International Corpus of Learner English v2 (Handbook + CD-Rom). Presses universitaires de Louvain, Louvain-la-Neuve.

Hadley, G. and M. Charles (2017). “Enhancing extensive reading with data-driven learning”. In: Language Learning & Technology 21.3, pp. 131–152.

Huang, P.-Y., C.-M. Chen, N.-L. Tsao, and D. Wible (2013). “A Corpus-Based Tool for Exploring Domain-Specific Collocations in English”. In: 27th Pacific Asia Conference on Language, Information, and Computation (PAACLIC), pp. 542–549.

Källkvist, M. (1998). “Lexical infelicity in English: the case of nouns and verbs”. English. In: Perspectives on Lexical Acquisition in a Second Language. Ed. by K. Haastrup and Å. Viberg. Lund University Press.

Koehn, P. (2005). “Europarl: A parallel corpus for statistical machine translation”. In: Machine Translation Summit (Phuket). Vol. 5. Asia-Pacific Association for Machine Translation, pp. 79–86.

Li, S. (2017). “Using corpora to develop learners’ collocational competence”. In: Language Learning & Technology 21.3, pp. 153–171.

Liang, P., B. Taskar, and D. Klein (2006). “Alignment by Agreement”. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Association for Computational Linguistics (ACL), pp. 104–111.

McGee, I. (2012). “Collocation dictionaries as inductive learning resources in data-driven learning: An analysis and evaluation”. In: International Journal of Lexicography 25.3, pp. 319–361.

Mel’čuk, I. (1998). “Collocations and Lexical Functions”. In: Phraseology. Theory, Analysis, and Applications. Ed. by A. P. Cowie, pp. 23–53.

Namvar, F. (Jan. 2012). “The relationship between language proficiency and use of collocation by Iranian EFL students”. In: The Southeast Asian Journal of English Language Studies 18 (3), pp. 41–52.

Och, F. J. and H. Ney (2003). “A Systematic Comparison of Various Statistical Alignment Models”. In: Computational Linguistics 29.1, pp. 19–51.

Östling, R. and J. Tiedemann (2016). “Efficient word alignment with Markov Chain Monte Carlo”. In: Prague Bulletin of Mathematical Linguistics 106, pp. 125–146.

Pecina, P. (2009). Lexical Association Measures: Collocation Extraction. Vol. 4. Studies in Computational and Theoretical Linguistics. Praha, Czech Republic: Institute of Formal and Applied Linguistics, Charles University in Prague.

Ronan, P. and G. Schneider (2015). “Determining Light Verb Constructions in Contemporary British and Irish English”. In: International Journal of Corpus Linguistics 20.3, pp. 326–354.

St. John, E. (2001). “A case for using a parallel corpus and concordancer for beginners of a foreign language”. In: Language Learning & Technology 5.3, pp. 185–203.

Volk, M., J. Graën, and E. Callegaro (2014). “Innovations in Parallel Corpus Search Tools”. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC) (Reykjavik). Ed. by N. Calzolari et al. European Language Resources Association (ELRA), pp. 3172–3178.

Vyatkina, N. (2016). “Data-driven learning for beginners: The case of German verb-preposition collocations Data-driven learning for beginners: The case of German verb-preposition collocations”. In: ReCALL 28.2, pp. 207–226.

Vyatkina, N. and A. Boulton (2017). “Corpora in language learning and teaching: Commentary”. In: Language Learning & Technology 21.3, pp. 1–8.

Wanner, L. (1996). Lexical Functions in Lexicography and Natural Language Processing. Vol. 31. John Benjamins Publishing.