Interconnecting lexical resources and word alignment: How do learners get on with particle verbs?

David Alfter[1] **& Johannes Graën**[1,2]

[1]Språkbanken, Department of Swedish, University of Gothenburg

[2]Grael, Department of Translation and Language Sciences, Pompeu Fabra University

Abstract

In this paper, we present a prototype for an online exercise aimed at learners of English and Swedish that serves multiple purposes. The exercise allows learners of these languages to train their knowledge of particle verbs receiving clues from the exercise application. At the same time, we collect information which will help us judge the accuracy of our graded word lists. As resources, we use lists with annotated levels from the proficiency scale defined by the Common European Framework of Reference (CEFR) and a multilingual corpus with syntactic dependency relations and word alignments for all language pairs. From the latter resource, we extract translation equivalents for particle verb constructions together with a list of parallel corpus examples that are used as clues in the exercise.

1 Introduction

Combinations of verbs and particles have been studied extensively in various aspects, e.g. particle placement with regard to cognitive processes (Gries, 2003), the relation between syntactical and semantic structure (Roßdeutscher, 2011) and their compositionality with respect to syntactic argument structure (Bott and Schulte im Walde, 2015). In the field of language learning, verb-particle combinations have been investigated in matters of their use of language learners of English (EFL) (Gilquin, 2015; Liao and Fukuya, 2004), also in comparison to native language speakers (Schneider and Gilquin, 2016) and with regard to pedagogical suggestions for language learning and teaching (Gardner and Davies, 2007).

The term 'phrasal verb' is used in most publications to refer to an English verb-particle combination that "behaves as a semantic unit" (Gilquin, 2015), while for other (mostly Germanic) languages term such as 'verb-particle constructions', 'verb-particle expressions' (Toivonen, 2002) or simply 'particle verbs' prevail (Zeller, 2001). Dehé (2015) compares particle verbs in Germanic languages and regards these terms as synonyms. We will thus refer to construction of verb and particle as particle verbs.

Particle verbs are especially difficult for learners since they present no discernible pattern in the selection of the particle. Gardner and Davies (2007) observe that “many nonnative English speakers actually avoid using phrasal verbs altogether, especially those learners at the beginning and intermediate levels of proficiency.” Not all verbs and particles are equally likely to take part in particle verbs. In English, “a number of lexical verbs such as take, get, come, put and go are particularly productive and frequent when they combine with adverbial particles” (Deshors, 2016). Gardner and Davies (2007) recommend learners to memorize those verbs and particles that occur frequently in verb-particle combinations.

Recently, so-called Games With A Purpose (GWAPs) (Lafourcade et al., 2015) have been used to collect information from players while offering a ludic interface that promotes participation. For example, JeuxDeMots (Lafourcade and Joubert, 2008; Lafourcade, 2007) has been used to find lexico-semantic relations between words, ZombiLingo (Fort et al., 2014) for the annotation of dependency syntax in French corpora, RigorMortis (Fort et al., 2018) for the identification of multiword expression by (untrained) learners, relying on their subjective opinion.

With the six reference levels of the Common European Framework of Reference (CEFR) (Council of Europe, 2001), henceforth CEFR levels, we can classify learners according to their level of proficiency. In Section 2.1, we introduce two resources that we build upon, which provide

Proceedings of the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa), pages 321–326
Turku, Finland, 30 September – 2 October, ©2019 Linköping University Electronic Press

lists of vocabulary units together with their estimated distribution over CEFR levels. In Section 2.2, we explain how we look up translation equivalents in several languages in a word-aligned multiparallel corpus, followed by a manual reassessment step described in Section 2.4.

In continuation, we present an application that implements a gamified exercise based on particle verbs in English and Swedish, their translation equivalents and corpus examples that demonstrate their use in authentic translations (Section 3). Learners playing the game try to not lose while the game automatically adapts to their current predicted knowledge level. The application keeps track of decisions taken by the user during the course of the game to provide them with feedback regarding their language skills, and points to potential weaknesses and (language-specific) factors for confusions. At the same time, we expect that a sufficiently large collection of decisions will help us assess the CEFR levels of our lexical resources and provide insights for future extensions.

2 Data Preparation

We extract particle verbs for CEFR levels from A1 to C1 from two lexical resources, one for English and one for Swedish.[1] For each particle verb that we find in these resources, we look up potential translation variants for several other languages, from a large multilingual word-aligned corpus. Since word alignment is less reliable when it comes to function words, we need to review the lists of translation variants and adjust word order and missing function words in multiword variants manually.

2.1 Lexical Resources

The CEFRLex project[2] offers lists of expressions extracted from graded textbook corpora for different languages. The languages currently available are French, Swedish and English. For this project, we use the Swedish list SVALex (François et al., 2016) and the English list EFLLex (Dürlich and François, 2018) from the CEFRLex project. Each resource lists single-word and multi-word expressions, as recognized by a syntactic parser, and their frequency in textbooks of different CEFR levels. Table 1 shows examples from the EFLLex list.

We extract particle verbs from both lists. For EFLLex, we use regular expressions to match all two-word expressions that are tagged as verbs. Manual inspection of the results shows that most expressions extracted this way are indeed particle verbs; we only had to exclude four expressions.[3]

For SVALex, we consider the subset of expressions tagged as verbal multi-word expressions. Since not all verbal multi-word expressions are particle verbs, we cross-check for the existence of each expression in the upcoming version of Saldo,[4] which includes particle verbs. Upon manual inspection of the resulting list we removed two reflexive particle verbs.[5] In total, we extracted 221 English and 362 Swedish particle verbs. As we are, among other things, interested in seeing how CEFR levels correlate with self-proclaimed proficiency, we assign each particle verb the CEFR level at which it first occurs in the respective resource, as has been previously done in various other experiments (Gala et al., 2013, 2014; Alfter et al., 2016; Alfter and Volodina, 2018).

2.2 Translation Equivalents from Parallel Corpus Data

The exercise is based on finding the correct particle for a particle verb in the target language based on translations in the source language. In other words, it means that, for example, learners of Swedish (target language) with knowledge of English (source language) will have to guess Swedish particle verbs based on English translations. For identifying translation equivalents in multiple languages, we use the Sparcling corpus (Graën, 2018; Graën et al., 2019), which, in addition to standard annotation such as part-of-speech tagging, features dependency relations from syntactic parsing in a number of languages (including English and Swedish) and bilingual word alignment for all language pairs. We use dependency relations to identify pairs of particles and their head verb matching the list that we extracted from EFLLex and SVALex.

For each occurrence of those pairs in the corpus, we look up aligned tokens in all other languages available to spot corresponding translation equivalents. We then filter the aligned tokens for content

[1] No particle verb has been classified as C2.

[2] http://cental.uclouvain.be/cefrlex/

[3] Those are ‘finger count’, ‘deep fry’, ‘go lame’ and ‘tap dance’, which use other part of speech than particles.

[4] https://spraakbanken.gu.se/eng/resource/saldo

[5] To wit ‘ge sig ut’ *‘go out’* and ‘klamra sig fast’ *‘cling to’*.

Expression	PoS	A1	A2	B1	B2	C1	C2
video	noun	65.19	0	67.87	81.76	111.06	90.93
write	verb	758.66	1421.51	1064.47	682.26	1104.72	1053.96
empty	adjective	0	28.83	28.65	102.29	37.84	61.88
shopping center	noun	0	45.12	9.80	0	15.50	11.45
dream up	verb	0	0	0	0	0.82	0.24

Table 1: Example entries from EFLLex.

words, that is, in terms of universal part-of-speech tags (Petrov et al., 2012), verbs, nouns, adjectives or adverbs. Functional parts of multi-word expressions are notoriously misaligned if the syntactic patterns of the corresponding expressions differ. For instance, English 'to cry out (for sth.)' can be expressed in Spanish with the fixed expression 'pedir (algo) a gritos'. In this case, we often see 'cry' aligned with 'pedir' and 'gritos', and the particle 'out' with the preposition 'a'. A similar expression is 'llevar (algo) a cabo' *'get through (sth.)'*, where 'carry' is aligned with 'llevar' and 'out' with 'cabo'; the preposition 'a' often remains unaligned in this case.

By filtering out function words, we systematically miss any preposition, determiner or particle that forms part of the equivalent expression. Not filtering them out, on the other hand, leads to considerably noisier lists. The missing functional parts need to be added back later and the set of lemmas needs be put in the preferred lexical order (see Section 2.4). We retrieve lemmas of the aligned tokens as a set, disregarding their relative position in the text, and calculate frequencies for each translation equivalent. Translation equivalents are most frequently single verbs. The Swedish particle verb 'ha kvar' (literally *'have left'*), for instance, is aligned to the English verbs 'retain' 49 times, to 'maintain' 31 times and to 'remain' 26 times.

2.3 Example Sentence Selection

Alongside other options (see Section 3), we want to provide learners with authentic examples where the given particle verb is used as translation of a particular expression in another language. We typically find several example sentences per translation correspondence in the Sparcling corpus. The question now is how to select the most adequate one for the respective learner. In previous works, we have used the length of the candidate sentence pair as ranking criterion, downgrading those pairs that showed a substantial deviation in length (Schneider and Graën, 2018; Clematide et al., 2016).

While there is a substantial amount of previous work on finding good example sentences for use in dictionaries (e.g. GDEX (Kilgarriff et al., 2008)) or for language learners (e.g. HitEx (Pilán et al., 2017)), most of the features they use are language-specific, such as blacklists, 'difficult' vocabulary, or recognizing and excluding anaphoric expressions without referent in the same sentence.

For the purpose of this study, we have thus opted for a simple heuristics which works well across a number of different languages. We use sentence length and a weighted measure for lexical proficiency required to understand the target language sentence (since we do not have gradings for most of the source languages).

2.4 Manual Revision

Manual correction involves the removal of irrelevant translations, the re-ordering of words, in case a particle verb has been aligned to multiple other words, and the insertion of missing words into the translation variants (as in 'llevar *a* cabo'). In addition, we judge example sentences with regard to adequacy.

While the translation candidate extraction could be restricted to allow only verbal translations for particle verbs, this is a constraint that we do not want to impose. Indeed, certain languages tend towards more nominal ways of expression while other languages tend towards more verbal ways of expression (Azpiazu Torres, 2006). Thus, imposing such a constraint could possibly induce non-idiomatic or unnatural translation candidates.

Having multiple part-of-speech possibilities for translation variants also allows us to potentially control the difficulty of the exercise by only giving verbal translation variants to beginners while, as the learner progresses and improves, other part-of-speech variants could be included.

3 Crowdsourcing and Gamification

We use our gamified system to assess knowledge of language learners in their L2 (English or Swedish), and to judge the accuracy of the automatically assigned CEFR labels. The game presents one base verb each round, together with a list of particles to choose from and one initial clue in form of a translation variant for the particle verb that the player is supposed to guess. The player can gain points by choosing the right particle and loose points by choosing a wrong one. Additional clues can be traded off against points. These clues can also be example sentences in the target language or the elimination of several of the non-fitting particles.

The learner assessment is achieved by monitoring how players of certain self-proclaimed proficiency levels deal with expressions that they are supposed to master, according to the automatic CEFR level assignment method. If learners systematically struggle with expressions of their self-chosen proficiency level, we assume that they overvalued their level and provide feedback accordingly. If they show little or no difficulties in dealing with expressions deemed of their current self-proclaimed proficiency level, we assume that their actual proficiency is higher, and gradually increase the challenge by using particle verbs of higher levels and more difficult clues (e.g. less frequent translation variants).

The accuracy of the automatically assigned CEFR labels is measured by aggregating results over all players. We also take into account response times for individual exercises. Significantly large deviance from the average answering time or the average number of points used for 'trading' clues for particle verbs of the supposedly same proficiency level suggests that the particle verb in question could belong to a different level.

Before the actual game starts, learners have to choose the language that they want to train. They are also asked to indicate their mother tongue and any other languages they know, including a self-assessment of their proficiency in the respective languages (beginner, intermediate, advanced). This rough scale is translated to the levels A1 and A2, B1 and B2 and C1 respectively.

Having finished the self assessment, the learner gets a predefined amount of points, as a virtual currency. More points can be gained each round by finding the right particle for the given verb with as few clues as possible. A wrong answer is worth an equally negative amount of points that could have been gained by choosing the right answer. We employ a function to calculate the reward based on hints used and difficulty of the hints in terms of language knowledge, i.e. a clue in a lower-rated language will cost the learner less points than, for instance, in his mother tongue. The game ends when the player is out of points or the game is out of particle verbs. The final score is used to create an entry on a leaderboard.

4 Discussion and Future Work

With the development of new CEFR graded multi-word expression lists, including a wider range of expressions, the exercise can be extended to other types of expressions. With the advent of CEFR graded multi-word lists in other languages, the exercise can also be extended to encompass a more diverse set of languages.

One aspect that is not specifically addressed in this study is the issue of polysemy. Indeed, a particle verb can have multiple meanings, and thus multiple different translations. This aspect will prove problematic when the particle verbs are shown in context, as one has to ensure that both the original as well as the translation pertain to the same sense of the expression.

Another question concerns the accuracy of the automatic assignment of CEFR levels based on the method used. While we surmise that we can gain insights about the accuracy of the assigned levels through the proposed prototype, a separate investigation should be carried out. One could possibly compare the automatically assigned levels from EFLLex to the levels given in English Vocabulary Profile.[6]

Acknowledgements

The authors would like to express their gratitude towards Lars Borin for making available the upcoming version of the Saldo lexical resource. This research is partly supported by the Swiss National Science Foundation under grant P2ZHP1_184212 through the project "From parallel corpora to multilingual exercises – Making use of large text collections and crowdsourcing techniques for innovative autonomous language learning applications".

[6] http://www.englishprofile.org

References

David Alfter, Yuri Bizzoni, Anders Agebjörn, Elena Volodina, and Ildikó Pilán. 2016. From Distributions to Labels: A Lexical Proficiency Analysis using Learner Corpora. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, 130, pages 1–7. Linköping University Electronic Press.

David Alfter and Elena Volodina. 2018. Towards Single Word Lexical Complexity Prediction. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*.

Susana Azpiazu Torres. 2006. Stylistic-contrastive analysis of nominality and verbality in languages. In *Studies in Contrastive Linguistics: Proceedings of the 4th International Contrastive Linguistics Conference*, 170, pages 69–77. Universidade de Santiago de Compostela.

Stefan Bott and Sabine Schulte im Walde. 2015. Exploiting Fine-grained Syntactic Transfer Features to Predict the Compositionality of German Particle Verbs. In *Proceedings of the 11th International Conference on Computational Semantics (IWCS)*, pages 34–39.

Simon Clematide, Johannes Graën, and Martin Volk. 2016. Multilingwis – A Multilingual Search Tool for Multi-Word Units in Multiparallel Corpora. In Gloria Corpas Pastor, editor, *Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives – Fraseologia computacional y basada en corpus: perspectivas monolingües y multilingües*, pages 447–455. Tradulex.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Press Syndicate of the University of Cambridge.

Nicole Dehé. 2015. Particle verbs in Germanic. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen, and Franz Rainer, editors, *Word-formation: an international handbook of the languages of Europe*, 40 edition, volume 1 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, chapter 35, pages 611–626. De Gruyter Mouton.

Sandra C. Deshors. 2016. Inside phrasal verb constructions: A co-varying collexeme analysis of verb-particle combinations in EFL and their semantic associations. *International Journal of Learner Corpus Research*, 2(1):1–30.

Luise Dürlich and Thomas François. 2018. EFLLex: A Graded Lexical Resource for Learners of English as a Foreign Language. In *11th International Conference on Language Resources and Evaluation (LREC)*.

Karën Fort, Bruno Guillaume, and Hadrien Chastant. 2014. Creating Zombilingo, a Game With A Purpose for dependency syntax annotation. In *Gamification for Information Retrieval Workshop (GamifIR)*.

Karën Fort, Bruno Guillaume, Mathieu Constant, Nicolas Lefebvre, and Yann-Alan Pilatte. 2018. "Fingers in the Nose": Evaluating Speakers' Identification of Multi-Word Expressions Using a Slightly Gamified Crowdsourcing Platform. In *Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions*, pages 207–213.

Thomas François, Elena Volodina, Ildikó Pilán, and Anaïs Tack. 2016. SVALex: a CEFR-graded Lexical Resource for Swedish Foreign and Second Language Learners. In *10th International Conference on Language Resources and Evaluation (LREC)*.

Núria Gala, Thomas François, Delphine Bernhard, and Cédrick Fairon. 2014. Un modèle pour prédire la complexité lexicale et graduer les mots. In *21ème Traitement Automatique des Langues Naturelles*, pages 91–102.

Núria Gala, Thomas François, and Cédrick Fairon. 2013. Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. *E-lexicography in the 21st century: thinking outside the paper*.

Dee Gardner and Mark Davies. 2007. Pointing Out Frequent Phrasal Verbs: A Corpus-Based Analysis. *TESOL quarterly*, 41(2):339–359.

Gaëtanelle Gilquin. 2015. The use of phrasal verbs by French-speaking EFL learners. A constructional and collostructional corpus-based approach. *Corpus Linguistics and Linguistic Theory*, 11(1):51–88.

Johannes Graën. 2018. *Exploiting Alignment in Multiparallel Corpora for Applications in Linguistics and Language Learning*. Ph.D. thesis, University of Zurich.

Johannes Graën, Tannon Kew, Anastassia Shaitarova, and Martin Volk. 2019. Modelling large parallel corpora:the zurich parallel corpus collection. In *Proceedings of the 7th Workshop on Challenges in the Management of Large Corpora (CMLC)*.

Stefan Thomas Gries. 2003. *Multifactorial analysis in corpus linguistics: A study of particle placement*. Open Linguistics. A&C Black.

Adam Kilgarriff, Milos Husák, Katy McAdam, Michael Rundell, and Pavel Rychlỳ. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX International Congress*.

Mathieu Lafourcade. 2007. Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *7th International Symposium on Natural Language Processing (snlp)*, page 7.

Mathieu Lafourcade and Alain Joubert. 2008. JeuxDe-Mots: un prototype ludique pour l'émergence de relations entre termes. In *Journées internationales d'Analyse statistiques des Données Textuelles (JADT)*, pages 657–666.

Mathieu Lafourcade, Alain Joubert, and Nathalie Le Brun. 2015. *Games with a Purpose (GWAPS)*. John Wiley & Sons.

Yan Liao and Yoshinori J. Fukuya. 2004. Avoidance of phrasal verbs: The case of Chinese learners of english. *Language learning*, 54(2):193–226.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA).

Ildikó Pilán, Elena Volodina, and Lars Borin. 2017. Candidate sentence selection for language learning exercises: from a comprehensive framework to an empirical evaluation. *TAL*, 57(3/2016):67–91.

Antje Roßdeutscher. 2011. Particle Verbs and Prefix Verbs in German: Linking Theory versus Word-syntax. *Leuvense Bijdragen*, 97:1–53.

Gerold Schneider and Gaëtanelle Gilquin. 2016. Detecting innovations in a parsed corpus of learner English. *International Journal of Learner Corpus Research*, 2(2):177–204.

Gerold Schneider and Johannes Graën. 2018. NLP Corpus Observatory – Looking for Constellations in Parallel Corpora to Improve Learners' Collocational Skills. In *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning (NLP4CALL)*, pages 69–78. Linköping Electronic Conference Proceedings.

Ida Toivonen. 2002. Verbal particles and results in swedish and english. In *Proceedings of the West Coast Conference in Formal Linguistics*, volume 21, pages 457–470.

Jochen Zeller. 2001. *Particle verbs and local domains*, volume 41. John Benjamins Publishing.