University of Zurich^UZH

Using Parallel Corpora for Generating Language Learning Exercises

Challenges and Pitfalls

Johannes Graën

Friday 8th July, 2022

Abstract

We want to generate language learning exercises automatically from (parallel) corpora (CALL).

In order to do so, we need to identify good examples from suitable corpora.

We need to provide reliable feedback to learners. This is why we need to pay special attention to exercise types and NLP techniques employed when generating exercises without teacher interaction.

Outline

- Data-driven learning
    - Using corpora for language learning
    - ICT literacy of teachers & students
    - Benefits of parallel corpora
- Example selection
    - GDEX for language learners
    - Proficiency levels (CEFR)
    - Additional criteria for parallel corpora
- Exercise generation
    - Exercise types
    - Applicability of ICALL methods
    - Challenges and pitfalls (and solutions)

Data-driven learning

Concept

Data-driven learning (DDL)

- Learners explore real-world data (corpora),
- … develop hypotheses,
- … and ‘prove’ them true or wrong.

Discovery learning and *discover & reconstruct* are similar concepts.

Reception of data-driven learning

- Data-driven learning has proven to be more effective and more efficient than ‘traditional’ learning methods
  - “lack of empirical studies exploring the actual impact of corpus methods on the learning outcomes” (Meunier 2011)
  - “corpus-based learning is more efficient than traditional treatments” (Boulton and Cobb 2017; Cobb and Boulton 2015)

Corpora for language learning

- Exposure to authentic language vs. constructed textbook examples
- But: existing tools are either too lightweight (Google-like) or too complex (corpus query languages) for typical learners
- Incidental learning of lexical items or grammatical structures possible

Corpora

- Very diverse landscape
- Best suited for language learners:
  - mode: here only text corpora
  - ⇒ text corpora of transcribed speech
  - text types: stories, dialogues, …
  - domains/genres: adjuvant if of interest to learner
- Suitability depends to a large part on learning goal
- Can be analyzed by means of NLP methods (part-of-speech tagging, lemmatization, morphological analysis, syntactical parsing, coreference resolution, named entity recognition, sentiment analysis, topic modeling, …) ⇒ ICALL

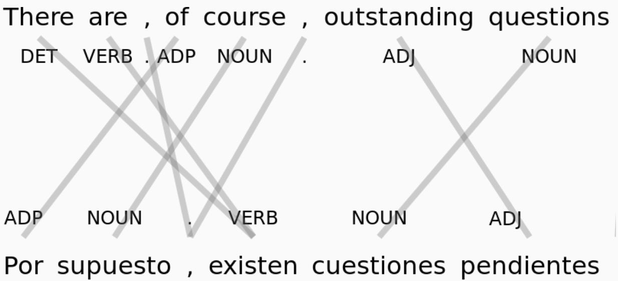
ICT literacy

- Learners need to be sufficiently proficient with the tools they use
- Effective use of ICT requires a particular learner level (Cruz Piñol 2015)
- Different types of tools for different learner proficiency levels (Buyse 2014)
- The use of technology is best taught in classrooms (Buyse 2014; Buyse and Verlinde 2013; Cassany 2016; Cobb and Boulton 2015; Vázquez-Calvo 2016)

Parallel corpora

- Typically translations (also from a third language or indirect)
- Many resources freely available:
    - The OPUS collection
    - The Zurich Parallel Corpus Collection
- One corpus to highlight: OpenSubtitles
    - Subtitles are usually short (unlike parliamentary debates, patents, or legal texts)
    - Domain and text type vary depending on the respective movie (usually ‘standard language’)
    - The corpus is huge in terms of tokens and languages
    - Translation made and reviewed by volunteers (no professional translations)

Word alignment


There are , of course , outstanding questions
DET VERB . ADP NOUN . ADJ NOUN
ADP NOUN . VERB NOUN ADJ
Por supuesto , existen cuestiones pendientes

- Links between corresponding tokens
- Automatically derived
- They are not necessarily word-by-word translations (e.g. functional parts of expressions)
- Many-to-many alignments that humans would expect are often not found by the algorithms

Benefits

- Translation to L1 (or a strong L2) can help disassembling structures
- ⇒ Let learners access annotations
- Different senses can be distinguished with the aid of the respective translations
- Aggregated alignment frequencies provide insight into different uses (as part of expressions or in terms of senses) (Graën and Schneider 2020)
- The combination of syntactic relations, alignments and alignment frequencies can be used to identify corresponding (idiomatic) constructions (Graën and Schneider 2017; Schneider and Graën 2018)

Example selection

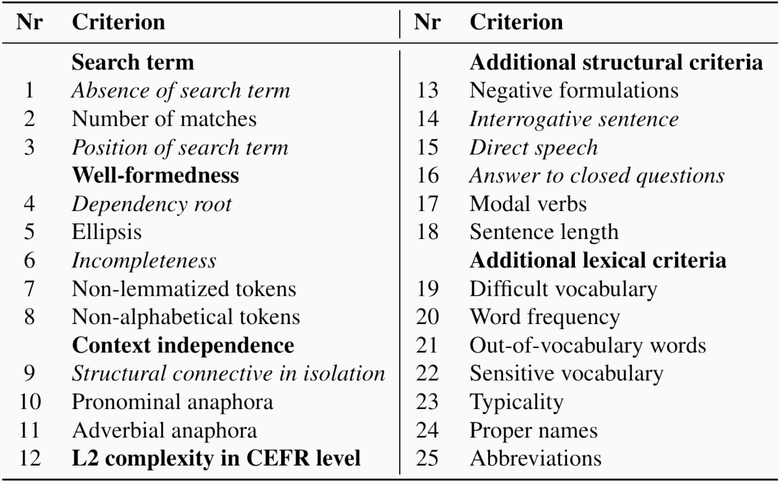
Identification of suitable examples

- *Good Dictionary Examples* (GDEX) in Lexicography (Kilgarriff et al. 2008)

  > *Sentences are evaluated with respect to their length, use of complicated vocabulary, presence of controversial topics (politics, religion...), sufficient context, references pointing outside of the sentence (e.g. pronouns), brand names and other criteria.* (`https://www.sketchengine.eu/guide/gdex/`)

- ⇒ Avoid *PARSNIP* (Politics, Alcohol, Religion, Sex, Narcotics, Isms, Pork) at all costs?

Criteria for automatic example selection



Nr	Criterion	Nr	Criterion
	Search term		**Additional structural criteria**
1	*Absence of search term*	13	Negative formulations
2	Number of matches	14	*Interrogative sentence*
3	*Position of search term*	15	*Direct speech*
	Well-formedness	16	*Answer to closed questions*
4	*Dependency root*	17	Modal verbs
5	Ellipsis	18	Sentence length
6	*Incompleteness*		**Additional lexical criteria**
7	Non-lemmatized tokens	19	Difficult vocabulary
8	Non-alphabetical tokens	20	Word frequency
	Context independence	21	Out-of-vocabulary words
9	*Structural connective in isolation*	22	Sensitive vocabulary
10	Pronominal anaphora	23	Typicality
11	Adverbial anaphora	24	Proper names
12	**L2 complexity in CEFR level**	25	Abbreviations

Criteria used in HitEx framework (Pilán, Volodina, and Borin 2016)

Resources

In monolingual corpora:

- CEFRLex framework for single lexical items
- Several readability measures for estimating the required proficiency level
- Complexity of derived syntactical structure (e.g. nestedness)

Additionally, in parallel corpora:

- Degree of idiomaticity by comparison
- Word alignment frequency (conditional probability)
- Derive word senses using alignment

Exercise generation

Exercise types (text only)

- Identify parts of speech, lemmas, morphological, …
- Reordering shuffled sentences (reconstruct storyline)
- Reordering shuffled words in a sentence
- Gap-filling/cloze exercises (with distractors, with or without given options)
- $\Rightarrow$ Subtype: bundled gap-filling
- Odd-One-Out (lexical)
- Adjust tense
- Reading comprehension questions
- …

Example 1: Bundled gaps

- Type: Fill-in-the-gap; the learner is presented a sentence with a gap (one word missing) and is asked to enter the missing word.
- Challenge: There could be several options to fill the gap that we created automatically.
- Approach: Use several (four) sentences with the same gap.
- Implementation: Available online
- Publication: (Wojatzki, Melamud, and Zesch 2016)

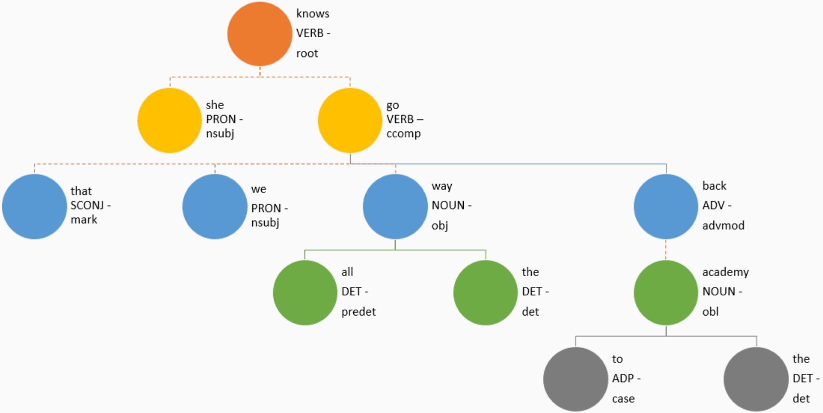
Example 2: Particle verbs and their prepositions

- Type: Fill-in-the-gap with multiple options; the learner is presented a sentence with a gap and asked to pick a preposition from a list.
- Challenge: There are often numerous prepositions that form a particle verb together with an often semantically light verb. We cannot be certain that one clue provides sufficient information.
- Approach: More clues and other means of help can be traded for a virtual currency.
- Prototype: Available online
- Publication: (Alfter and Graën 2019)

Example 3: Sentence reconstruction on bilingual sentence pairs

- Type: Novel type of exercise; the learner is asked to identify matching tokens between source and target language.
- Challenge: There might be multiple options and function words can often not be assigned in a meaningful way.
- Approach: Group tokens in the source language, so that the assignment is between tokens and chunks.
- Implementation: Experimental study online
- Publication: (Zanetti, Volodina, and Graën 2021)

Grouping through clustering



Other aspects

- Gamification, GWAP
- Supervised approaches
    - indirect corpus consultation by teachers
    - Crowdsourcing of language learning materials
    - Optional feedback on each exercise by learners (?)
- Learners’ attitude towards technology is key; half-baked solutions might have a negative impact.

Questions/comments?

References i

Alfter, David and Johannes Graën (2019). “Interconnecting lexical resources and word alignment: How do learners get on with particle verbs?” In: *Proceedings of the 22nd Nordic Conference of Computational Linguistics (NODALIDA)*. Turku, Finland: Linköping University Electronic Press, pp. 321–326.

Boulton, Alex and Tom Cobb (2017). “Corpus Use in Language Learning: A Meta-Analysis”. In: *Language Learning* 67.2, pp. 348–393.

Buyse, Kris (2014). “Una hoja de ruta para integrar las TIC en el desarrollo de la expresión escrita: recursos y resultados”. In: *Journal of Spanish Language Teaching* 1.1, pp. 101–115.

Buyse, Kris and Serge 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.

Cassany, Daniel (2016). “Recursos lingüísticos en línea: Contextos, prácticas y retos”. In: *Revista signos* 49, pp. 7–29.

Cobb, Tom and Alex Boulton (2015). “Classroom applications of corpus analysis”. In: *The Cambridge Handbook of English Corpus Linguistics*. Ed. by Douglas Biber and Randi Reppen. Cambridge University Press, pp. 478–497.

Cruz Piñol, Mar (2015). “Léxico y ELE: enseñanza/aprendizaje con tecnologías”. In: *Journal of Spanish Language Teaching* 2.2, pp. 165–179.

References ii

Graën, Johannes and Gerold 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.

— (2020). “Exploiting Multiparallel Corpora as a Measure for Semantic Relatedness to Support Language Learners”. In: *Strategies and Analyses of Language and Communication in Multilingual and International Contexts*. Ed. by David Levey. Cambridge Scholars Publishing, pp. 153–167.

Kilgarriff, Adam, Miloš Husák, Katy McAdam, Michael Rundell, and Pavel Rychlỳ (2008). “GDEX: Automatically Finding Good Dictionary Examples in a Corpus”. In: *Proceedings of the 13th EURALEX International Congress*. Ed. by Janet DeCesaris Elisenda Bernal. Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra.

Meunier, Fanny (2011). “Corpus linguistics and second/foreign language learning: exploring multiple paths”. In: *Revista Brasileira de Linguística Aplicada* 11.2, pp. 459–477.

Pilán, Ildikó, Elena Volodina, and Lars Borin (2016). “Candidate sentence selection for language learning exercises: from a comprehensive framework to an empirical evaluation”. In: *Traitement Automatique des Langues* 57.3, pp. 67–91.

References iii

Schneider, Gerold and Johannes Graën (Nov. 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 at SLTC 2018 (NLP4CALL)*. Linköping Electronic Conference Proceedings, pp. 69–78.

Vázquez-Calvo, Boris (2016). “Digital language learning from a multilingual perspective: the use of online language resources in the one-to-one classroom”. PhD thesis. Universitat Pompeu Fabra. Departament de Traducció i Ciències del llenguatge.

Wojatzki, Michael, Oren Melamud, and Torsten Zesch (June 2016). “Bundled Gap Filling: A New Paradigm for Unambiguous Cloze Exercises”. In: *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*. San Diego, CA: Association for Computational Linguistics, pp. 172–181.

Zanetti, Arianna, Elena Volodina, and Johannes Graën (2021). “Automatic Generation of Exercises for Second Language Learning from Parallel Corpus Data”. In: *International Journal of TESOL Studies* 3.2: *Technology in Applied Linguistics*. Ed. by Marina Dodigovic and Stephen Jeaco, pp. 55–71.