Raising awareness of first-language interference using parallel corpora of subtitles

Elen Le Foll

This is a post-peer-review, pre-copyedit version of a paper published as:

Le Foll, Elen (in press/2022). **Raising awareness of first-language interference using parallel corpora of subtitles**. In Viana, V. (Ed.) *New Ways in Teaching with Corpora*. Abingdon, Oxon; New York, NY: Routledge. https://www.routledge.com/Teaching-English-with-Corpora-A-Resource-Book/Viana/p/book/9781032252971

Raising awareness of first-language interference using parallel corpora of subtitles

Elen Le Foll (Osnabrück University)

Abstract:
Negative first-language (L1) transfer is a common source of unidiomatic language use as learners often assume that words, phrases, collocations and other idiomatic expressions can be translated word for word. This chapter outlines a fun group activity that introduces learners to parallel corpora to tackle such unidiomatic language usage. The proposed paper-based data-driven learning (DDL) activity encourages learners to find more idiomatic ways to translate difficult words or phrases from their L1 into English by analysing concordance lines from TV and film subtitles. The chapter features a few examples, and the appendix includes detailed instructions for teachers to create their own worksheets on any language issue of relevance to their students. The activity draws on the subtitle subcorpus of the multilingual InterCorp corpus via the free corpus interface KonText. It can be adapted to the full range of proficiency levels and for more than 40 different learner L1s. Optionally, more advanced students can be introduced to the web interface to conduct their own parallel corpus searches online.

Levels: Elementary and above
Aims:

- Raise learners' awareness of issues of L1-L2 equivalence
- Practise conversation skills

Class time: 45 minutes
Preparation time: 30 minutes
Resources: Worksheet(s) designed by teacher (see online support materials)

Introduction

Negative first-language (L1) transfer is a common source of unidiomatic language use since learners often assume that words, phrases, collocations and other idiomatic expressions can be translated word for word. The fun group activity proposed in this plan is a stimulating way to introduce learners to parallel corpora offline and to show learners how to find more idiomatic ways to translate difficult words or phrases from their L1 into English. Drawing on the multilingual nature of the class, the lesson can be delivered in English, in the students' first language or using a combination of both.

Steps

1. Write a sentence featuring an L1-induced error that your students often make on the board. For instance, if you teach English to Spanish L1 speakers, a typical L1-induced error might be the use of the general noun *people* with a definite article and a verb in the singular form. Thus, Spanish L1 speakers may produce the following: **I don't think that a local public transport website is a good idea because the people does not have Internet access in the region.*
2. Ask students to discuss what is unidiomatic in this sentence in pairs. For example, in this context, *people* usually occurs without an article and is used with the plural verb form.
3. As a whole class, encourage students to think of reasons why this unidiomatic usage is common in the speech of Spanish L1 speakers. Here, this is likely to be because, in Spanish, *gente* (*people*) is typically used with the definite article *la* (*the*) and a verb in the singular form (e.g., *La gente no tiene acceso a Internet*). Therefore, this unidiomatic usage is likely due to L1 interference.
4. Group students in pairs or groups of three.
5. Hand out the subtitle worksheet(s) you have designed for the student groups (see online support materials for preparation procedure).
6. Explain that the sentences displayed come from film and TV subtitles in the students' L1 and in English.
7. Ensure that all students are aware of what subtitles are.
8. In their groups, ask students to identify different ways in which the focus word/phrase has been translated in the subtitles. Ensure that students pay particular attention to articles, verb forms, prepositions, and word order.
9. Tell students that there will likely be more than one idiomatic way to translate the focus word/phrase, so they should consider the context(s) in which each solution is most frequent. For instance, in the case of *la gente* (as suggested in the online support materials), students will see that it is usually translated by *people* without an article. However, some of the examples feature an adjective specifying a particular group of *people*, and, in these cases, an article is used (e.g., *And when the wrong people get shot?).*
10. Pool answers together as a class and summarise the information on the board.
11. Highlight the different ways to idiomatically translate the focus word/phrase. For instance, there are cases in which the phrase *la gente* is translated without using the word *people* at all (e.g., *Probaré que **la gente** del Sur ve la verdad , no el color de la piel* is translated as *I'll show **the men and women** of the South will look past color and see the truth* and *Es mejor hablar a **la gente*** as *It's better to talk to **the crowd***).
12. Let students in each group choose two or more examples from the worksheet that they particularly like and ask them to imagine a film scene in which the subtitles will feature in the dialogue.
13. Instruct students to write the dialogues of their film scenes.
14. Allow time for students to create and practise their dialogues. Every student in the group should have an active role.
15. Invite students to perform their film scenes in front of the rest of the class.

Points for consideration and alternative steps

- You can easily create bespoke worksheets for each group (see online support materials), thus adapting the activity to various levels of proficiency or L1 backgrounds within one class.

- This activity does require the teacher to spend some time preparing the worksheet(s); however, from the second worksheet onwards, the procedure will be relatively quick and is well worth the time investment since the KonText interface (Machálek 2020) allows for the creation of bespoke worksheet(s) created on the basis of authentic spoken-like language.
- If you teach older and more advanced students and have access to at least one computer/tablet with Internet access per pair of students, students can also work with KonText directly. They should follow the procedure described in the online support materials. The advantage is that they will learn to search parallel corpora and have access to many more examples than you can provide on a single worksheet. In future, they may also be able to use this method to perform other queries and solve other L1-L2 equivalence issues. Note, however, that this option is probably not suitable for young learners due to the strong language and inappropriate content of some of the films featured in the subtitle datasets of the InterCorp corpora.
- With a monolingual class, it is also possible to extend the activity by requiring students to write subtitles to their film scene in their L1 to accompany the live performances of their scenario. This provides further opportunities for students to gain greater awareness of L1-L2 equivalence issues and can help foster interlingual mediation skills.
    - Use Microsoft PowerPoint or LibreOffice Impress to create a plain slide presentation with just one long text box at the bottom of each slide.
    - Ask students to write one subtitle per slide.
    - At the same time as they perform their film scene, display the subtitles using a projector and allow a student to click through the slides at the appropriate moment.

References and suggested reading

Čermák, F., & Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics, 17*(3), 411–427.

Gambier, Y., Caimi, A., & Mariotti, C. (Eds.) (2014). *Subtitles and language learning: principles, strategies and practical experiences*. Bern/New York: Peter Lang.

Machálek, T. (2020): KonText: Advanced and Flexible Corpus Query Interface. In: Proceedings of LREC 2020, pp. 7005–7010. http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.865.pdf

Tiedemann, J. (2009). News from OPUS: A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, & R. Mitkov (Eds.), *Recent advances in natural language processing* (vol V; pp. 237–248), Amsterdam/Philadelphia: John Benjamins.

Yamashita, J., & Jiang, N. (2010). L1 influence on the acquisition of L2 collocations: Japanese ESL users and EFL learners acquiring English collocations. *Tesol Quarterly*, *44*(4), 647–668.

Online support materials

Procedure for teacher preparation

1. Open a web browser and connect to KonText (https://kontext.korpus.cz). It is not necessary to register at KonText to prepare this lesson but registering is free and will give you access to some useful additional options such as being able to save your searches.
2. If the interface is not automatically displayed in English, click on "English" in the top-right corner.
3. Click on "syn2020", which is the default corpus at the time of writing, then click on "All corpora" and on "parallel" (see Figure 1).

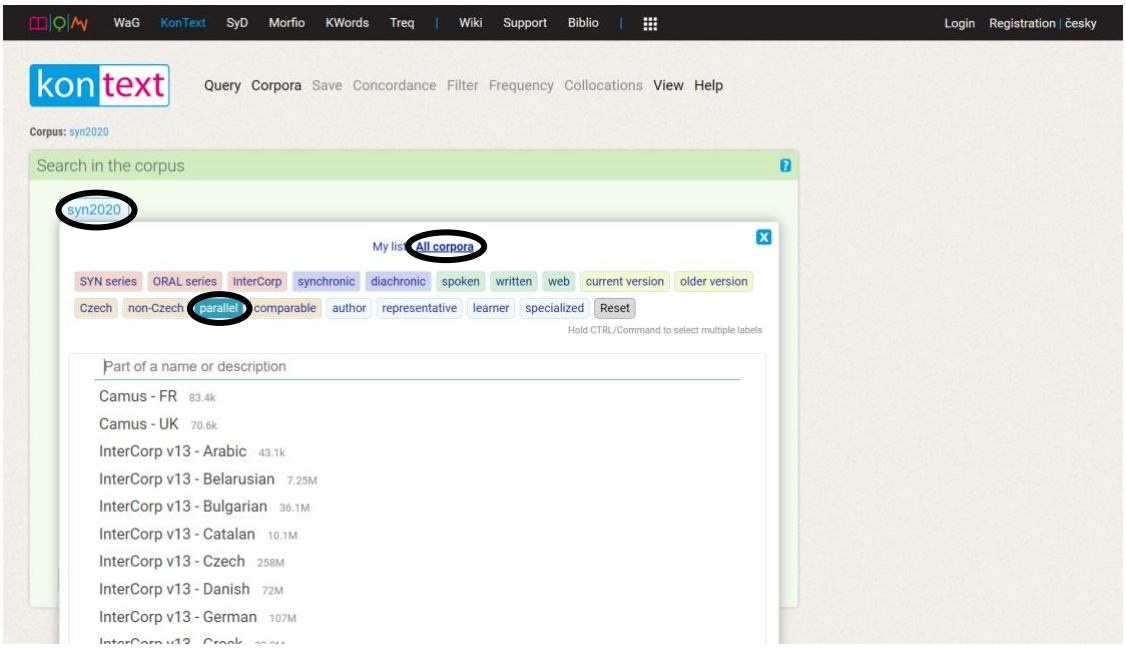

Figure 1: Corpus selection

4. Click on the latest version of the InterCorp corpus (Čermák & Rosen 2012) which corresponds to your students' L1 (e.g., if you teach English to German students, click on "InterCorp v13 - German"). If your students have different L1s, you will need to repeat the procedure for each L1 (or see Options and caveats section). The InterCorp corpora are parallel corpora containing a wide range of text registers. This lesson plan only draws from the TV and film subtitle subcorpus (source: http://www.opensubtitles.org). Sequences of up to 100 words are aligned to the corresponding translation in a number of languages. The number of languages and the sizes of the InterCorp corpora are constantly growing. At the time of writing, InterCorp parallel corpora were currently available on the KonText web interface in Albanian, Arabic, Belarusian, Bulgarian, Catalan, Chinese, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hebrew, Hindi, Hungarian, Icelandic, Italian, Japanese, Latvian, Lithuanian, Macedonian, Malay, Maltese, Norwegian, Polish, Portuguese, Romani, Romanian, Russian, Slovak, Slovenian, Serbian, Spanish, Swedish, Turkish, Ukrainian, and Vietnamese.
5. Choose a language issue caused by L1 transfer that your students regularly face and that you would like them to improve on. This could be negative L1 transfer derived from a single word, collocation, phrase, or construction.
    - For example, German native speakers often struggle to translate the ubiquitous word *selbst,* which can mean *automatic(ally), even, in person, self-, (by)*

myself/yourself/herself etc. Negative L1 transfer can lead to the production of utterances such as: **This is a self-made cake.*

- In French, many constructions use the verb *faire* and, as a result, L1 French speakers tend to overuse the English verb *make* and produce unidiomatic collocations such as: **I made my homework*, or **I always used to make nightmares*.
- Spanish uses the phrase *la gente* to refer to people in general. Consequently, Spanish L1 speakers will often use *people* with an article and with a verb in singular form, e.g., **The people here is very friendly.*

6. Type the language item(s) you have selected in the search box.
   - If the issue is a single invariable word (e.g., *selbst* in German), type *selbst* in the search box.
   - If the issue stems from a construction containing more than one word in the learners' L1, such as *la gente* in Spanish, type the full phrase (e.g., *la gente*) in the search box.
7. In the search box, click on the option "Aligned corpora" and, in the drop-down menu, select the latest version of the English InterCorp (e.g., "InterCorp v13 - English").
8. Click on the option "Restrict search". In the section "text.txtype", tick "subtitles". This restricts the search to texts drawn from aligned film and TV series subtitle data.
9. Scroll down to the bottom of the page and click on the "Search" button.
10. Take a moment to explore the results page (see Figure 2) which shows a list of subtitles containing the L1 word or phrase queried on the left, and the English versions of the subtitles on the right. The number of hits (A) corresponds to the number of times the queried word, lemma or phrase was found in subtitles in InterCorp. The results are displayed in a random order.

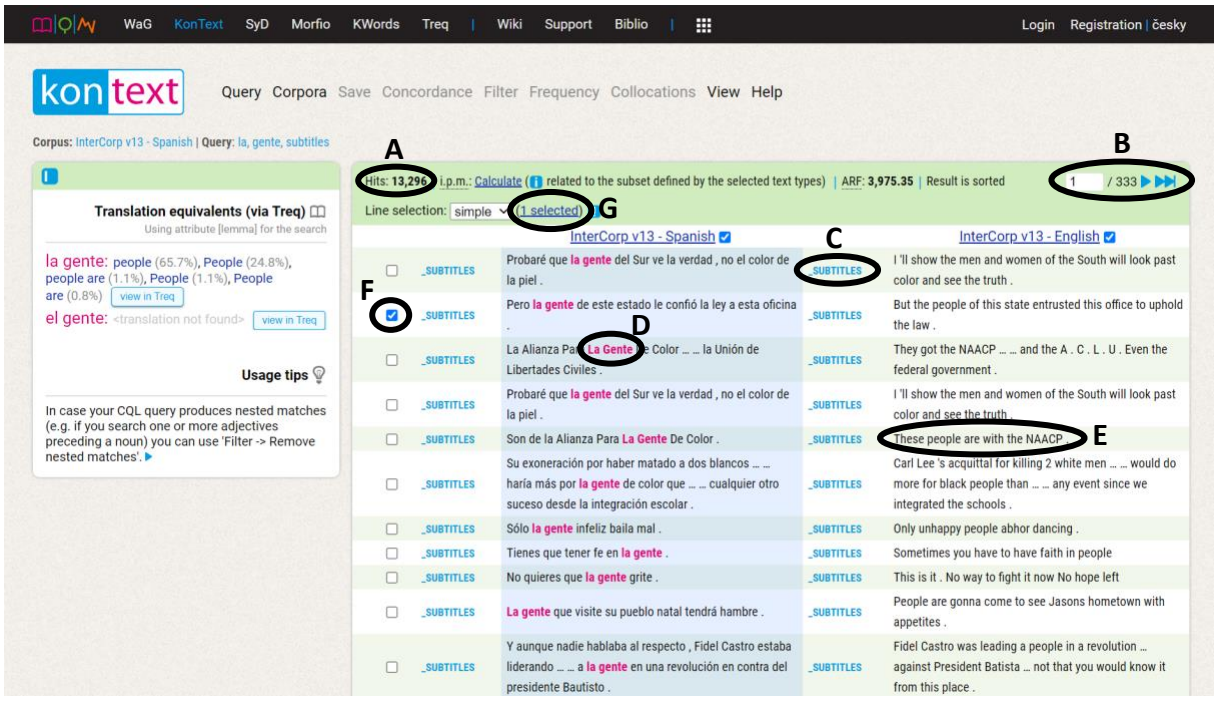

Figure 2: Results page for *la gente* in InterCorp Spanish and English

11. Click on the forward and back buttons (B) to display more results. If you click on "_SUBTITLES" (C), an additional window will open, containing, among other information, the title of the film (under "text.title"). By clicking on the pink focus words/phrases within

the extracts of the L1 subtitles (D), you can read more of the co-text from the L1 film transcripts. Clicking on a subtitle on the right-hand side (E) will open an additional window displaying the co-text in which the subtitle is found in English.

12. By ticking the corresponding boxes (F), select subtitles which:
    - have been accurately and idiomatically translated (bearing in mind that, even though many of the subtitles come from DVDs, some of the translations fail to meet professional standards);
    - contain lexical and grammatical features suitable for your students' level;
    - are appropriate for your students in terms of content (especially since strong language is not uncommon in films and TV series).
13. Once you have made your selection, click on the blue link "[number] selected" (see G in Figure 2) and choose "Keep only selected lines" from the "Action" drop-down menu. This opens a new window containing only the subtitles that you have selected.
14. Print your selection or save as a PDF (landscape is recommended). This is your subtitle worksheet.

Resources

Čermák, F. & Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics* 17(3). 411–427. https://doi.org/10.1075/ijcl.17.3.05cer.

Machálek, T. (2020). KonText: Advanced and Flexible Corpus Query Interface. In: Proceedings of LREC 2020, pp. 7005–7010. http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.865.pdf