PaCLE

Parallel Corpora for Language Learning Exercises

https://demo.spraakbanken.gu.se/johannes/PaCLE/

Short summary

This is an overview of useful features that can be used with regular expressions. A detailed explanation can be found on the following pages.

1. To find a single word such as Spanish “y” use spaces around that word « y » or the the code for an anchor between word and non-word characters “\y”: «\yy\y». This also works if a word is followed by a comma for example: “With y, as in yellow.”
2. Use \w+ for a sequence of letters, either as a replacement for a variable suffix (distribu\w+), prefix (\w+estimate) or as a placeholder for a word (from \w+ point of view)
3. Alternatives are specified in parentheses starting with “?:”, for example look (?:for|after) will match “look for” and “look after”
4. Alternatives between single letters are given in square brackets, [cz] means either c or z.
5. A question mark (?) means “optional”, an asterisk (*) means “optional, but can be repeated” and a plus sign means “at least one”.
6. An exclusion pattern given in parentheses and starting with “?!” does not allow the following sequence to occur at the given position.
   (?!distribución|distribuidora?)distribu\w+ will match any word that starts with “distribu” and comprises at least one more letter unless the word is “distribución”, “distribuidor” or “distribuidora”
7. Words enclosed in simple parentheses are shown in the header of each example. If no simple parentheses are used in the search expression, the whole match is used instead.

For more information about regular expressions please have a look at this page: regular-expressions.info

Introduction

PaCLE is a tool to facilitate the use of parallel corpora for language learning. It is both directed to teachers with regard to finding adequate authentic translation examples for exemplifying lexical choice or grammar and to language learners with regard to exploring unknown expressions and structures.

This manual explains PaCLE's basic functionality and provides real-world examples that can be found in the text collections. Currently, the only corpus in PaCLE is OpenSubtitles, which is based on movie subtitles translated by volunteers on the www.opensubtitles.org platform. Subtitles are to a large extent dialogs, which makes them more suitable for learners than other corpus material such as parliamentary debates or technical manuals.

Configuration

Before we can perform searches, we need to specify the language pair that we are interested in. One of the languages is the one that we want to learn or improve (target language), the other one is a language that we know better, e.g. the language that we speak natively or on a more advanced proficiency level.

A learner from Sweden, for example, who is learning Catalan, would use Swedish as source and Catalan as target language, but could also use Spanish as source language if her level allows her to comprehend the sentences that she reads.

Basic usage

For looking up single words, one only has to write that word in the search box of the language in question and hit enter to start the search. Here, we search for embarrassed in English with the other language being Spanish:

https://demo.spraakbanken.gu.se/johannes/PaCLE/#/query?pair=en-es&q1=embarrassed&order=4&limit=100

The results can be subdivided into different translation variants. In the first result…

Any little difference between me and the other children and I was embarrassed and humiliated in front of what felt like the world. Cualquier pequeña diferencia entre mí y los otros niños y yo estaba avergonzada y humillada frente a lo que parecía el mundo.

… we see that **embarrassed** translates to **avergonzada**. Further down the list, we find an example where **feel embarrassed** is translated with **avergonzarse**...

They ought to be alone on their honeymoon, and they might feel embarrassed. Deben de estar solos en su viaje de bodas y podrían avergonzarse.

… and one where the English adjective corresponds to a Spanish noun:

That's right, though I'm quite embarrassed about it. Así es, y no escondo mi vergüenza al admitirlo.

To filter out all words that have to do with **vergüenza** on the Spanish side, we can simply list all variants that we want to exclude separated by a pipe symbol or “vertical bar” and then change the search from inclusion (+) to exclusion (-):
vergüenza|avergonzado|avergonzada|avergonzar|avergonzarse

https://demo.spraakbanken.gu.se/johannes/PaCLE/#/query?pair=en-es&q1=embarrassed&q2=verg%C3%BCenza%7Cavergonzar%7Cavergonzado%7Cavergonzada&n2=1&order=4&limit=100

This will filter out any of the words given in that list, but will still let pass verb forms like **avergoncé**. However, now we also see many infrequent variants now like:

I'm embarrassed. Me pongo roja.

… or:

She's a little embarrassed. Ella se siente violenta.

… or:

If I embarrassed you, forgive me, but I wanted you to see that they are not a threat to national security. Perdóneme si le ha resultado embarazoso, pero quería que viera que mis padres no son un peligro para los EE.UU.

Using regular expressions

Instead of a list of words or word forms to exclude, we can also define a general pattern by means of regular expressions. Regular expressions are a very powerful way to express letter sequences and they can become quite complex. But for describing inflection, we merely need a small subset of their functionality.

Instead of vergüenza|avergonzado|avergonzada|avergonzar|avergonzarse, we can write a?verg(?:üe|o)n[cz]\w+ instead, which we will disassemble and discuss now in detail.

https://demo.spraakbanken.gu.se/johannes/PaCLE/#/query?pair=en-es&q1=embarrassed&q2=a%3Fverg%28%3F%3A%C3%BCe%7Co%29n%5Bcz%5D%5Cw%2B&n2=1&order=4&limit=100

1. a? – The question mark after a letter marks that letter as optional, so we allow an “a” or nothing for this part.
2. verg – This is just a literal sequence of letters v, e, r, and g.
3. (?:üe|o) – For this part, we allow either the sequence “üe” or just an “o”
4. n – Again a literal “n”
5. [cz] – At this point, we either allow “c” or “z”.
6. \w+ – A backslash introduces a special letter class, in this case any letter that typically belongs to “word” as opposed to “digits” (\d) or “space characters” (\s) for example. The plus sign marks a variable number of occurrences starting with 1 of what precedes the sign. In this case, we expect at least one “word” letter, but potentially many of them.

If we look at words that are supposed to be matched, we can identify the 6 parts that we see above:

- vergüenza: 1: (empty), 2: verg, 3: üe, 4: n, 5: z, 6: a
- avergonzado: 1: a, 2: verg, 3: o, 4: n, 5: z, 6: ado
- avergonzarse: 1: a, 2: verg, 3: o, 4: n, 5: z, 6: arse
- avergüences: 1: a, 2: verg, 3: üe, 4: n, 5: c, 6: es
- avergoncé: 1: a, 2: verg, 3: o, 4: n, 5: c, 6: é
- avergonzásemos: 1: a, 2: verg, 3: o, 4: n, 5: z, 6: ásemos

We can also use the same pattern for matching word instead of excluding them:

https://demo.spraakbanken.gu.se/johannes/PaCLE/#/query?pair=en-es&q1=embarrassed&q2=a%3Fverg%28%3F%3A%C3%BCe%7Co%29n%5Bcz%5D%5Cw%2B&order=4&limit=100

The part that matches the search expression is shown in the header of each example unless we use the expression to filter out sentences (exclusion mode).

In Romance languages, verbs are heavily inflected. The most common way of inflecting verbs is to add a suffix to the stem of that verb. Other inflection processes involve vowel changes (e.g. Spanish probar > pruebo; French gérer > je gère; Italian venire > vieni).

Using the pattern for variable endings from above, we can try to catch all verb forms of the Spanish verb **distribuir** by using the stem and adding the pattern “any number of any word letter” (\w+) to it: distribu\w+

https://demo.spraakbanken.gu.se/johannes/PaCLE/#/query?pair=en-es&q2=distribu%5Cw%2B&order=4&limit=100

In the result set, we find different forms, for example **distribuye, distribuirán, distribuido, distribuyeron**, but also the nouns **distribuidor/distribuidora** and **distribución**.

We can exclude those from matching words that we don't want to see in the result set by using an exclusion pattern that regular expressions provide. It is defined inside parentheses and starts with a question mark followed by an exclamation mark: (?!...)

That way, we can exclude **distribuidor, distribuidora** and **distribución:**

https://demo.spraakbanken.gu.se/johannes/PaCLE/#/query?pair=en-es&q2=%28%3F%21distribuci%C3%B3n%7Cdistribuidora%3F%29distribu%5Cw%2B&order=4&limit=100

Multi-word units

So far, we have only looked at finding single-words examples in the corpus. However, multi-word units are equally important when learning a language. Since PaCLE matches complete sentences against the search patterns given by its users, we can similarly perform a lookup for a multi-word unit by simply writing it into the search box and hitting enter.

If we want to explore translation variants of the English support verb construction **(to) pay attention**, we can use pay attention as a search expression:

https://demo.spraakbanken.gu.se/johannes/PaCLE/#/query?pair=en-es&q1=pay%20attention&order=4&limit=100

To not miss any inflected form of the verb **pay**, we can improve the search by adding alternatives with expression (?:...) that we used above for the alternative between "üe" and "o" in (?:üe|o). The resulting search expression could look like this: (?:pay|pays|paid|paying) attention

https://demo.spraakbanken.gu.se/johannes/PaCLE/#/query?pair=en-es&q1=%28%3F%3Apay%7Cpays%7Cpaid%7Cpaying%29%20attention&order=4&limit=100

As translation variants, we see **prestar atención**, **poner atención,** or **hacer caso** in the example list.

The expression above will match occurrences of **pay attention** only if both words appear side by side in the text**.** We will still miss those occurrences where an adjective is used to modify the noun **attention**. If we expect a single word in between **pay** and **attention**, we can use the pattern \w+ to match any sequence of "word" letters as described above. The resulting search expression thus looks like this: (?:pay|pays|paid|paying) \w+ attention

https://demo.spraakbanken.gu.se/johannes/PaCLE/#/query?pair=en-es&q1=%28%3F%3Apay%7Cpays%7Cpaid%7Cpaying%29%20%5Cw%2B%20attention&order=4&limit=100

The result set includes **more:**

I pay more attention to my performances than I do to my makeup.
Le doy más importancia a mis performances que a mi maquillaje.

… and **less**:

Now you're dead, perhaps some of your old friends will pay less attention to you. Give you more elbow room.
Ahora que está muerto, quizás sus viejos amigos le dejen en paz, con más espacio para operar.

We also find **any**:

He refuses to pay any attention to a servant girl's story.
Se niega a prestar atención a la historia de una criada.

… and **no**:

If she greets you, don't pay any attention to her.
Si ella te saluda, ignórala.

Among the actual adjectives, we see **close**:

Now... pay close attention.
Mira... fíjate bien.

… **particular**:

In watching this story... I want you to pay particular attention... to the three undraped ladies... who dance in the final scene.
Al ver este episodio... quiero que preste especial atención... a las tres mujeres desnudas... que bailan en la escena final.

… **little**:

My parents had paid little attention to me.
Mis padres se ocuparon poco de mí.

… and several more. To highlight the part that we are interested in, we simply need to put parentheses around it: (?:pay|pays|paid|paying) (\w+) attention

https://demo.spraakbanken.gu.se/johannes/PaCLE/#/query?pair=en-es&q1=%28%3F%3Apay%7Cpays%7Cpaid%7Cpaying%29%20%28%5Cw%2B%29%20attention&order=4&limit=100

Partial matches

If we don't specify pattern sequences at the beginning or - more likely - at the end of a word, what we match might actually be only part of a word. If we are searching for view on, we also match **review on:**

Make the Duke of Willenstein commander-in-chief, assemble your troops, pass the review on horseback, surprise them! Haga al Duque de Willenstein comandante en jefe, prepare las tropas, pase revista a caballo, ¡sorpréndalos!

… **preview on**:

You have to make arrangements for the preview on Friday. Si tiene que organizar el preestreno del viernes, él irá conmigo.

… and **interview on**:

I've got this 8:30 interview on this toxic waste story. Tengo una entrevista a las 8:30 en una planta de desechos tóxicos.

To remedy the search expression, we can either add a space in front of **view** or we use another feature of regular expressions, which is an anchor point, which matches the beginning and end of a letter sequence. It is expressed by \y

The resulting search expression is thus \yview on

https://demo.spraakbanken.gu.se/johannes/PaCLE/#/query?pair=en-es&q1=%5Cyview%20on&order=4&limit=100

Other search ideas

- Which preposition is required by the verb **depend**? depend (\w+)
- Does **separated** use **by** or **with** as preposition? separated (by|with)
- Which verbs go with **… to death**? (\w+d) to death
- Which ways are there to say the Spanish expression **a fin de cuentas** in English? a fin de cuentas