Mining terminology from Eur-Lex corpora

Marek Pawelec
wasaty@wasaty.pl

This is slightly extended handout created for mini-workshop conducted during #TranslatingEurope Forum 2016, which was held in Brussels on October 27-28 2016.

The European Union legislative acts should be consulted whenever translating any content that is directly, and sometimes even only indirectly, affected by UE regulations. The acts, which are drafted in parallel in all UE languages establish common, multilingual terminology in broad range of subjects, from legal and financial frameworks, through chemicals, medical devices to particular, sometimes very specific subjects. Of course the terminology is not made up by the translators (there are exceptions...), but based on the terms used in the relevant fields, thus giving us excellent source of terminology when translating between any of the 24 official languages. What's more, in many cases, when dealing with translation of documents governed by regulations, the use of EC legislative terminology and phrases is mandatory.

Introduction to UE legislative acts

Let's start with a quick look of UE acts. There are two broad categories:

1. **Legislative acts**: within the meaning of the Treaty on the Functioning of the European Union (TFEU)): regulations[1], directives[2] and decisions which are adopted
2. **Non-legislative acts**: within the meaning of the Treaty on the Functioning of the European Union: regulations, directives and decisions which are not adopted by legislative procedure (delegated acts (Article 290) and implementing acts (Article 291) and acts based directly on the Treaties (acts relating to international agreements, CFSP decisions, etc.)), as well as other acts, such as recommendations and ECB guidelines.

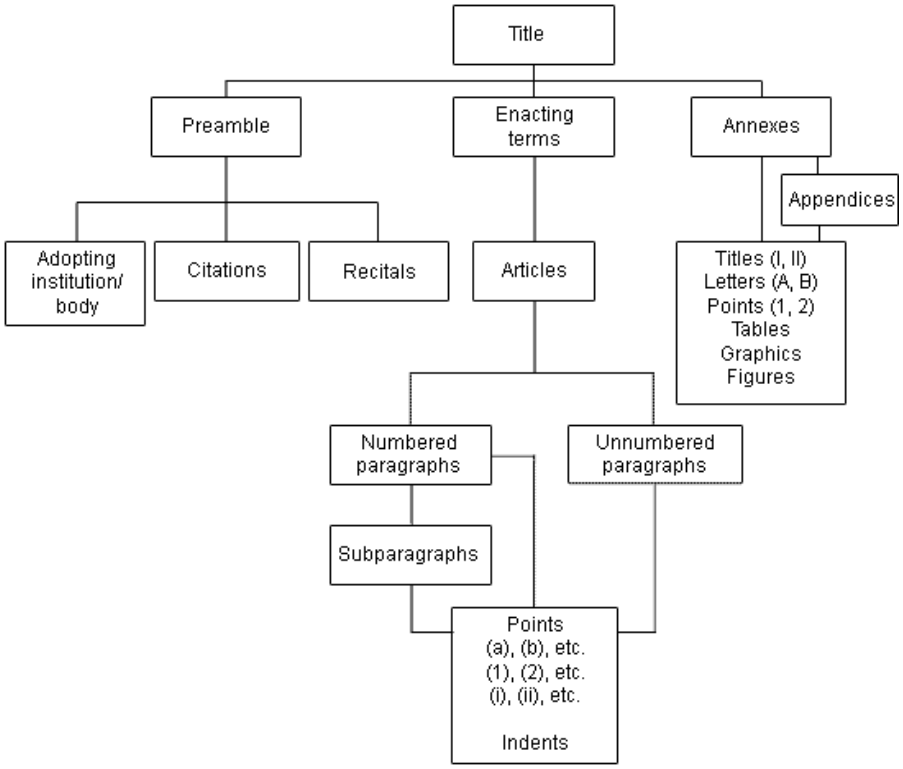
The legal act consist of:

- **Title**: Depicts type of act, number, date of adoption and subject matter.
- **Preamble**: Gives the legal basis for the act and reasons for the content of the enacting terms of an act (recitals).
- **Articles**: Enacting terms, which constitute the normative part of the act, divided into articles. Articles may be grouped in parts, titles, chapters and sections and subdivided into paragraphs, subparagraphs, points, indents and sentences. Where the enacting terms are simple, they may be set out in a 'Sole Article'. In directives and, where applicable, decisions, the addressees are specified in the last article.
- **Concluding formulas**: Place, date and signature.
- **Annexes**: The annex to an act generally contains rules or technical data which, for practical reasons, do not appear in the enacting terms, and which frequently take the form of a list or table.

[1] A regulation is a legal act of the European Union that becomes immediately enforceable as law in all member states simultaneously.

[2] A directive is a legal act of the European Union, which requires member states to achieve a particular result without dictating the means of achieving that result.

The structure can be summarized with the image below.



Source: http://publications.europa.eu/

The legal acts of European Union are publicly available through Internet Eur-Lex, the official website of European Union law and other public documents of the European Union (EU), published in 24 official languages of the EU. Each document can be downloaded in several forms (PDF, HTML) or displayed in mono- bi- and trilingual version. The acts are also available as bilingual Translation Memory Exchange (TMX) files, which can be used directly in all Computer Aided Translation (CAT) tools.

If we want to familiarize ourselves with the legal basis and terminology of the given subject field, the ideal solution would be to read relevant acts. However, this is usually not feasible – for example Regulation (EC) No 1272/2008 on classification, labeling and packaging of substances and mixtures, which is fundamental when it comes to chemistry-related translations (SDS), is over 111,000 words long and written in complex language that is not easy to read. So, how can we efficiently use the available legal acts as a source of relevant terminology and phraseology? My suggested solution is outlined below.

Terminology mining

First of all, why “mining”, and not “extracting”? Because what I’m suggesting here includes not only proper terms, as recognized by terminologists[3], but also “bastardized”, yet extremely pragmatic definition of terms: “whatever you store in your term base or glossary” (because it’s useful), especially recurring phrases. As translators working within the scope of a given subject field and specific regulations, for maximum credibility we need to deliver our translation using the same phrases used in the referenced acts. And additional bonus is that such term bases or glossaries, when

[3] Terms are words and compound words or multi-word expressions that in specific contexts are given specific meanings [Wikipedia]

used in modern CAT tools, can markedly improve our translation throughput. In general, terminologists distinguish two types of terminology work: systematic and ad hoc.

The former is kept in higher regards, as purposeful activity conducted in order to establish or extend terminology resources of a given domain. The most relevant – from our point of view – element of this kind of activity is identification of concepts, associated terms and their equivalents in target language. This requires either going over the source and/or bilingual corpora texts to in order to manually identify terms ("gold standard" of terminology work), or use of some kind of automated solution for term identification. The simplest, but quite useful approach is to use statistical analysis for term identification, optimally with the use of stop words lists to avoid prepositions, common words etc. The result of such activity is a list of "term candidates" or "potential terms", which is a starting point for identification of real terms and their target language equivalents.

Ad hoc terminology work means that terms are identified and (ideally) recorded as an element of translation process. This is usually perceived as less time-consuming, since there is no dedicated activity of term identification. However, to gain any profits from ad hoc work one must record identified terms in a term base or glossary (term base is organized according to concepts, while glossary is simply a list of terms and their equivalents).

When does it make sense to perform a dedicated, systematic terminology extraction? It's most beneficial when entering new domain, e.g. when you receive for translation some documents on the subject you have no deeper background knowledge. Spending some time on mining terminology and often used phrases can give you a solid head start and pay off quickly. Ad hoc work makes more sense when working in a field you generally know – simply record new terms you deem important or useful to achieve better consistency, ensure correct terminology and speed-up your work.

It's also worth noting that translated EC terminology may be inconsistent between or within legal acts and is subject to change over time, so it is always a good idea to record not only term translation, but also the source where the translation was found (e.g. act number and publication year).

So, how to approach the term extraction process in systematic terminology work?

1. **Identify relevant legislative acts**.
   When translating documents in regulated field, you need to be aware it's regulated and need to know relevant regulations, otherwise you won't have a chance to deliver translation which is not only conveys the right meaning, but also uses the right regulatory terms and phrases, and in some areas (e.g. drug registration), use of non-approved wording will disqualify the document. If you enter new subject field, simply use search engines to find as much about it, as you can, adding to your search criteria keywords like "directive" or "regulation". If it's regulated, you won't have problem finding guides to regulations governing that field. You can also use online corpora (e.g. Linguee) to look for key phrases – if the terms or phrases turn up in UE acts, display them and check their content.
2. **Familiarize yourself with the acts**.
   If you are not going to specialize in this particular area, you don't have to read all the documents thoroughly, but go over them using the bilingual view in Eur-Lex. Note especially purpose and scope of the act and presence of any definitions, placed at the beginning of enacting terms. These are the "real terms" you have to translate correctly. These clearly defined terms can be added to your term base, but you can also use IATE term base[4], which contain such official terminology, or some domain-specific term bases, e.g. ECHA term[5] for

[4] http://iate.europa.eu/
[5] https://echa-term.echa.europa.eu/

chemicals. Don't forget to go over annexes – they often contain translator's "treasure trove" of terms and phrases directly usable in translations, especially if they include tables, document templates and/or examples. In such cases statistical term extraction won't give satisfactory results – if a regulatory phrase occurs only once or twice, it may be omitted despite being extremely important for translated text.
It is the main reason I prefer to use alignment over ready-made TMX files (see below) – corpora files or "bilingual" Excel files can be used to see parallel documents in context and manually "extract" relevant, important terminology and phrases.

3. **Align monolingual act versions into bilingual file or download and extract TMX file**.
Bilingual resources are something all translators love, especially if they are reliable and normative. In most cases the UE acts can be aligned quickly and reliably using free or commercial software, resulting in bilingual corpora, aligned Excel files or TMX files – one of the many available solutions is described below.
EU legislation is also available as TMX collections, which can be downloaded freely and extracted[6]. This has the benefit of skipping the alignment step, however the resulting TMX will contain a batch of regulations, so the use of additional software (e.g. Olifant[7]) would be required to extract TM for specific regulation, which is simple procedure, but falls outside the scope of this document.
4. **Extract terminology**.
Depending on the available software, there are several options for automatic or semiautomatic terminology extraction. In general, they can be divided into two approaches:
    - Monolingual terminology extraction with manual identification of target-language equivalents: supported by variety of free and commercial software.
    - Bilingual terminology extraction: in practice limited to costly, specialized commercial software. Although there are free tools available, they are very unreliable (e.g. Anchovy) or limited to small volumes of text.

Given the scope of this presentation, I will present the alignment, extraction and equivalent finding workflow with three free tools: LF Aligner, Okapi Rainbow and ApSIC Xbench (alternatively AntPConc). Please note, that you'll get better extraction results when using bigger corpus – either longer single act, or set of acts on the same subject: the bigger the sample, the better statistical significance can be obtained.

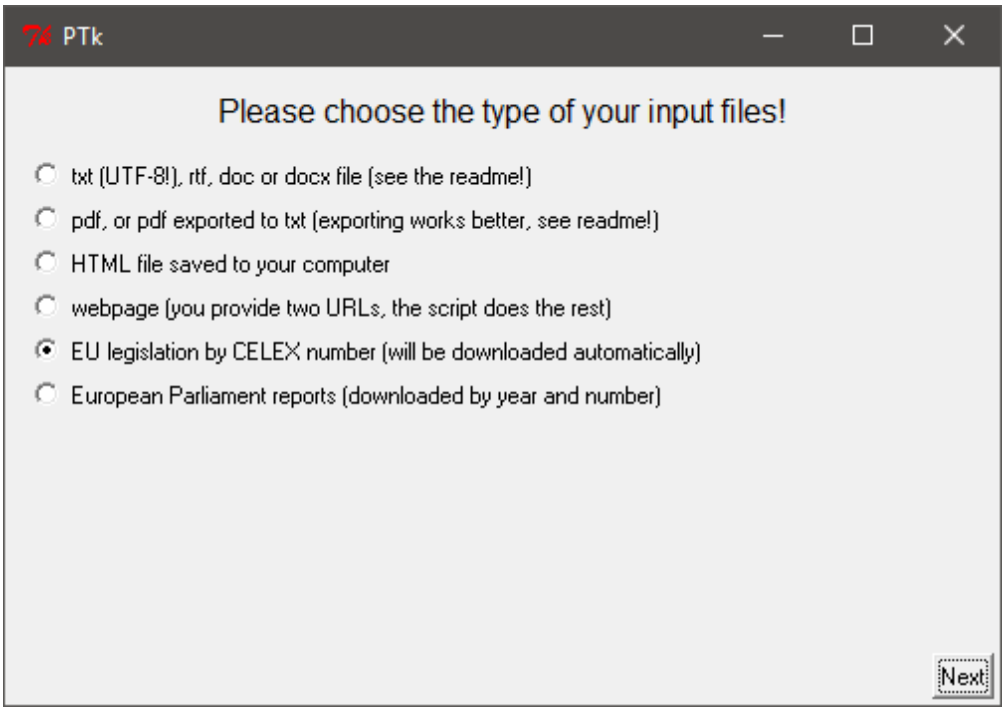
Terminology mining procedure

I. Legal act alignment

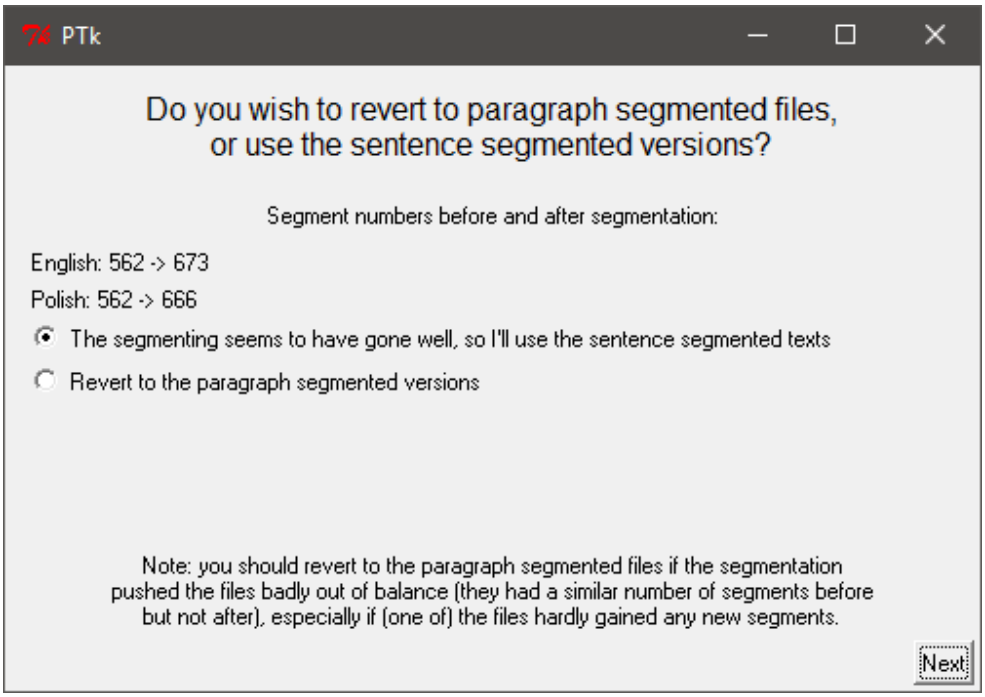
In this step we'll use free software called **LF Aligner**[8], intended for alignment of monolingual documents. After downloading and unpacking the software:

1. Run LF Aligner by double clicking **LF_aligner_4.1.exe** program icon
2. In the **Please choose type of your input files!** dialog select **EU legislation by CELEX number**. Click **Next**.

[6] https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory
[7] http://okapi.sourceforge.net/downloads.html
[8] https://sourceforge.net/projects/aligner/



3. In the **Please choose the folder where your files will be saved!** dialog click **Browse** and select relevant folder. Click **Next**.
4. In the **Specify the languages of your texts:** dialog select source and target languages. Click **Next**.
5. In the **Enter the CELEX number!** dialog enter Celex (e.g. 32015R0830). Click **Next**. Program will download documents from Eur-Lex service.
6. In the following dialog check the number of source and target segments for sentence segmentation. If the numbers are identical or very similar, click **Next**. If there is a big difference in number of sentences, select **Revert to the paragraph segmented versions**. Paragraph alignment is generally more reliable, but less useful than sentence-based bilingual document. Click **Next**.



7. In the **Review the aligned file to correct any incorrectly paired segments** select **Generate an xls and open it for reviewing**. Click **Next**.

Note*: you can also use native graphical editor or skip the review step, e.g. if the number of segments for source and target was identical.*

8. The alignment will be displayed in Excel. Review alignment and correct any problems. When finished, save the file and close Excel.

37	Done at Brussels, 28 May 2015.	Sporządzono w Brukseli dnia 28 maja 2015 r.
38	For the Commission	W imieniu Komisji
39	The President	
40	Jean-Claude JUNCKER	Jean-Claude JUNCKER
41	(1) OJ L 396, 30.12.2006, p.	Przewodniczący
42	1 . (2) Regulation (EC) No 1272/2008 of the European Parliament and of the Council of 16 December 2008 on classification, labelling and packaging of substances and mixtures, amending and repealing Directives 67/548/EEC and 1999/45/EC, and amending Regulation (EC) No 1907/2006 (OJ L 353, 31.12.2008, p.	
43	1).	(1) Dz.U. L 396 z 30.12.2006, s. 1 .
44		(2) Rozporządzenie Parlamentu Europejskiego i Rady (WE) nr 1272/2008 z dnia 16 grudnia 2008 r. w sprawie klasyfikacji, oznakowania i pakowania substancji i mieszanin, zmieniające i uchylające dyrektywy 67/548/EWG i 1999/45/WE oraz zmieniające rozporządzenie (WE) nr 1907/2006) (Dz.U. L 353 z 31.12.2008, s. 1).
45	(3) Commission Regulation (EU) No 453/2010 of 20 May 2010 amending Regulation (EC) No 1907/2006 of the European Parliament and of the Council on the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) (OJ L 133, 31.5.2010, p.	(3) Rozporządzenie Komisji (UE) nr 453/2010 z dnia 20 maja 2010 r. zmieniające rozporządzenie (WE) nr 1907/2006 Parlamentu Europejskiego i Rady w sprawie rejestracji, oceny, udzielania zezwoleń i stosowanych ograniczeń w zakresie chemikaliów (REACH) (Dz.U. L 133 z 31.5.2010, s. 1).

9. In the **Do you want to generate a TMX file?** dialog select **Yes**. Click **Next**.
10. If necessary, modify language codes and add note. Click **Next**. Target TMX file will be generated in folder selected in step 3.

II. Terminology extraction

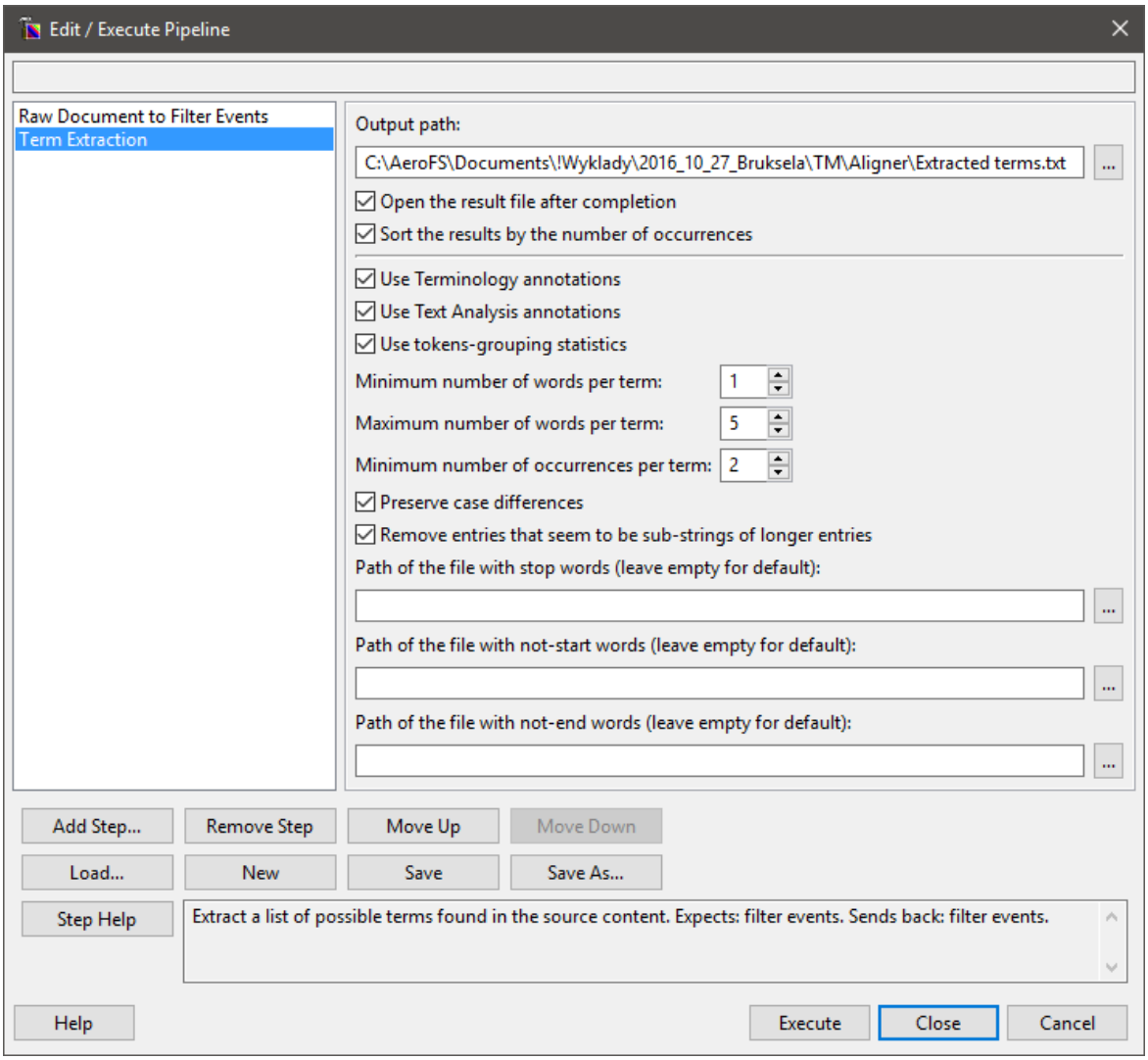
In this step we'll use **Okapi Rainbow**[9]. It's an element of Okapi localization tools which offers wide variety of features, including terminology extraction. Okapi requires Java to run. Also note, that while we will be using bilingual TMX file, only source language term candidates will be extracted. Monolingual file can be used for that purpose too. And there are plenty of other free monolingual terminology extraction tools, some offering better results (for example, Rainbow treats numbers as words) – however, the tool was chosen because of its versatility and ability to work with bilingual files.

Most of the functions in Rainbow can be invoked through so-called pipelines, which can involve multiple steps.

After downloading and unpacking the software:

1. Run Rainbow by double clicking **rainbow.exe** program icon.
2. Click **Input** > **Add documents** and select TMX file generated in step I.10. Alternatively simply drag the TMX file and drop in the Rainbow window. If you aligned more acts on a given subject, repeat procedure to add all of them into the list.
3. Select **Language and Encodings** tab, choose correct languages and encoding (UTF-8) for source and target.
4. Click **Utilities** > **Edit/execute pipeline**.
5. Click **Add step** > **Raw documents to filter events** > **OK**.
   *Each pipeline step requires either "raw documents" or "filter events" as input. The type required is stated clearly in the step description in the **Edit/execute pipeline** dialog ("Expects" and "Sends back").*
6. Click **Add step** > **Term Extraction** > **OK**.
7. Click Browse button [**...**] to select folder and file name in the **Output path** field.
8. Check **Open the result file after completion** and **Sort the results by number of occurrences** check boxes.
9. Optionally change extraction settings e.g. **Minimum/maximum number of words per term**.
10. Click **Execute**. Extraction on the source language will be run and the result file will be displayed.

[9] http://okapiframework.org/



11. OPTIONALY: create an empty .txt file and provide paths to it for stop words – this will allow you to extract phrases including words like “and”, “or”, “with” etc., and not only potential terms. You can save these results to separate extraction file. Generally you can experiment with available settings to find optimal results.

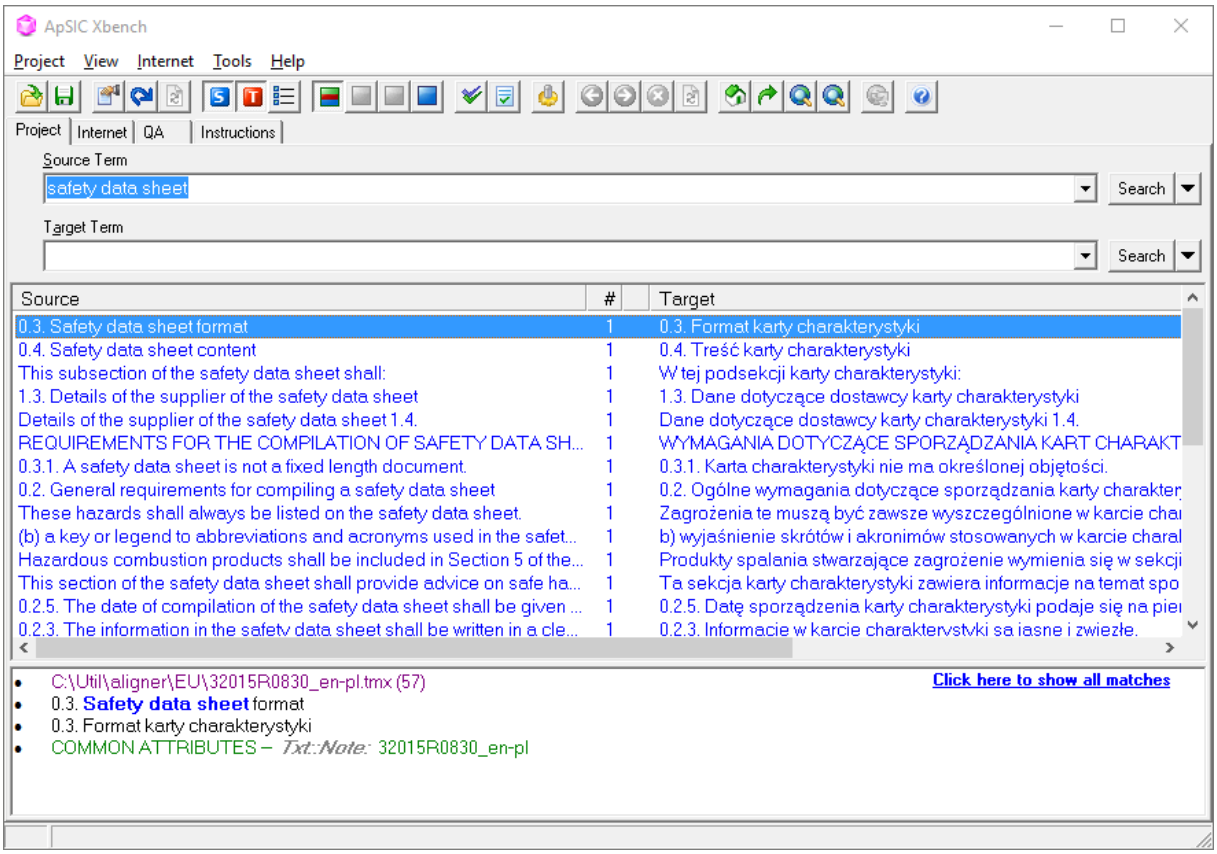
Note: *When using TMX files extracted from EU TMX collection, please note different encoding (UTF 16 LE) and full language designations, eg. “EN-GB” instead of “EN”.*

III. Target language equivalent finding with ApSIC Xbench

ApSIC Xbench[10] is a powerful dictionary/QA software available both in free and commercial version. Main limitation of the free version is lack of UTF support. After downloading and unpacking the software:

1. Run the program by double clicking Xbench.exe icon.
2. Click **Project** > **Properties** (F2).
3. Click **Add**.
4. Select **TMX memory**.
5. Click **Next**.
6. Click **Add file** and browse to the TMX file created in step I.13. **Click OK**, then **Next**.
7. Click **OK** and again **OK**, to **close Project properties** dialog.
8. Use **Source term** field to display terms or phrases extracted in stage II.

[10] http://www.xbench.net/



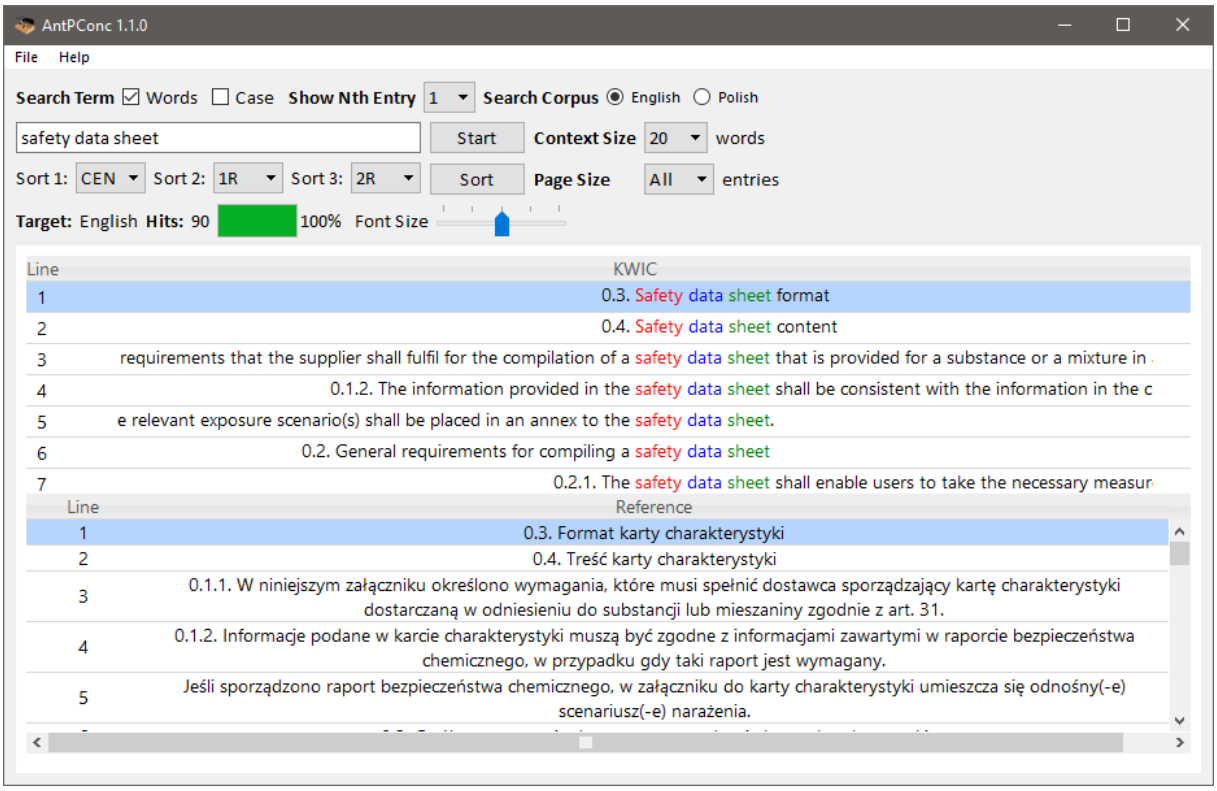
IV. Target language equivalent finding with AntPConc

AntPConc[11] is a freeware parallel corpus tool for concordancing and text analysis. After downloading the program (no extraction or installation is necessary):

1. Run program by double clicking **AntPConc.exe** icon.
2. Click **File** > **Build/edit corpus**.
3. Enter "Source" or source language name (e.g. "English") in the **Display Name** field.
4. Click **Choose files** in the **Corpus 1** tab.
5. Browse to LF aligner folder created in step I.3 and select source-language .txt file created by LF aligner and click **OK**.
6. Select **Corpus 2** tab and enter "Target" or target language name (e.g. "Polish") in the **Display Name** field.
7. Click **Choose files** in the **Corpus 2** tab.
8. Browse to LF aligner folder created in step I.3 and select target-language .txt file created by LF aligner and click **OK**.
9. Click **Update corpus**.

You can now use editing field of the main AntPConc window to display search terms in context. Click match to display target language equivalent. Use list of extracted terms or phrases to find their equivalents and record them e.g. in Excel sheet.

[11] http://www.laurenceanthony.net/software/antpconc/



Please note that parallel files created by LF Aligner may contain differences corrected during alignment review step. This can be resolved by using Rainbow:

1. Execute steps 1-5 from stage II using TMX as input file.
2. Click **Add step** > **Format Conversion** > **Parallel Corpus Files**.
3. Browse to folder location in the **Output path** field and enter a name for resulting files – they will be suffixed with language designations.
4. Click **Execute**.

On a final note, commercial tools often offer more features and greater ease of use plus overall integration of resources in the translation environment – both specialized terminology extraction tools and translation environments. In my preferred tool – **memoQ**[12] – alignment can be run quickly with great results, generating corpus which can be used directly for translation matches, as a reference file and source for monolingual terminology extraction with instant access to target language parallel translation for quick equivalents finding to facilitate bilingual term base creation.

[12] https://www.memoq.com/en/, 45-day free trial available