

## Compilation of the parallel corpus of international treaties

Andrei Nosov  
University of Tampere

Received: 01/05/2018

Accepted: 15/06/2018

### *Abstract*

This paper focuses on the description of the corpus «PEST-INTER»<sup>1</sup> in five languages and the process of its compilation and incorporation. The aim is to give step-by-step instruction on the corpus compilation. The further purpose is to show up the practical solutions for the problems raising in different stages of the corpus compilation. Describing the decisions taken and the strategies followed I discuss the corpus planning going into depth on web crawling, character and corpus encoding, automatic alignment and editing of the compiled texts.

### *Key Words*

Parallel Corpora, Encoding, Alignment, Incorporation.



### *Introduction*

In this article, I present the principles of collection of the parallel corpus PEST-INTER for n-gram based extraction of the lexicographically relevant data (LRD). By LRD, I mean a multilingual noun-oriented list of collocations extracted by statistical association measurements (AM) and normalized by specially developed syntactic restrictions. The basic unit of extraction of the lexicographically relevant data is the lexical *n-gram*, which is usually defined as a contiguous sequence of *n* words. Lists of n-grams obtained from mono- or multilingual corpora are a valuable source of LRD (terms, set expressions and

---

<sup>1</sup> Mikhailov, Mikhail (2016) 'PEST, Parallel Corpus of State Treaties. University of Tampere.' [Online] [Accessed on 4th March 2018] <https://puolukka.uta.fi/texthammer/>

clichés). Nowadays, there is a growing demand in translation and localization industry as well as in terminology and lexicography to quickly retrieve LRD from the text collections. Therefore, an effective model for this purpose could become a burning issue. The parallel corpus PEST-INTER will be specially developed to ensure the processing of this model. In order to be an appropriate tool the corpus should contain the texts in languages with different structure (Russian, Finnish, English, Swedish and French). In addition, the corpus should have a suitable structure, which will allow processing a large scale of research tasks and facilitating its incorporation in the larger database of state treaties PEST (COMS, University of Tampere).

This complex task defines the structure and strategy of the development of the specialized parallel corpus PEST-INTER in five languages (English, Finnish, Swedish, French and Russian). In order to provide the quality of the data collection and conformity with the previously mentioned task we created the automated system of collection of parallel texts and related information (i.e. metadata). I will present this system in detail below.

Nowadays, a number of huge international projects (i.e. European Parliament Proceedings Parallel Corpus<sup>2</sup>, Opus – The Open Parallel Corpus<sup>3</sup>, Hansard Corpus British Parliament<sup>4</sup>, etc.) are implementing automatic systems for collecting the texts from Internet. Carla Parra Escartín notes that there is no uniform approach to the description of the text collection for parallel corpora, because of the extensiveness and special features of this task [5]. At the same time Dan Tufiş and Jörg Tiedemann underline that one of driving factors in the field of corpus studies is the advantage of automated system's use over manual methods of processing [8] as they manage the data collection much better [7].

I find these assumptions instructive and develop the experimental guidelines for the corpus compilation by automated systems. The results of our case study may be applied in the further researches on the development of corpora of different types corresponding the broad spectrum of linguistic tasks (i.e. the automatic extraction of the multilingual lexicographic material for translation purposes,

---

<sup>2</sup> Koehn, Philipp (2005) 'A Parallel Corpus for Statistical Machine Translation.' *MT Summit* [Online] [Accessed on 8th February 2018] <http://statmt.org/europarl/>

<sup>3</sup> Tiedemann, Jörg (2012) 'Parallel Data, Tools and Interfaces in OPUS.' *In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)* [Online] [Accessed on 6th March 2018] <http://opus.nlpl.eu/>

<sup>4</sup> Alexander, Marc and Anderson, Jean (2012) 'The Hansard Corpus, 1803-2003.' *University of Glasgow, Glasgow, UK.* [Online] [Accessed on 9th March 2018] <https://www.hansard-corpus.org/>

modelling the reference dictionaries for improving the quality of machine translation, etc.).

For the collection of the texts for the corpus PEST-INTER, I specially developed an algorithm, which would be included in the larger system of the software products of the University of Tampere and would improve the existing corpus manager TextHammer<sup>5</sup>.

Bearing in mind the previously mentioned, my work may be considered in twofold perspective: (1) it has a practical impact for the purposes of collection of the parallel corpus PEST-INTER compiled for n-gram based extraction of the lexicographically relevant data (LRD); and (2) it provides the theoretical foundations for the automatic collection of parallel corpora as whole.

### *1. Basic principles of the corpus compilation*

In order to present some basic principles of compilation of a specialized corpus in compliance with Corpus Linguistics standards, I address to those, which seems to be the most appropriate. Firstly, I would mention that the development of the corpus PEST-INTER is based on the automated system of collection of texts, i.e. crawling. The fact that the corpus is available for a large scale of research tasks makes it universal tool and ensure its further use.

Another important point is about the structure of the corpus. The main challenge is to elaborate the corpus structure, which could support the methods of automatic extraction of LRD (i.e. collocations, keywords, n-grams, etc.). Here I issue from the basic terminological representations [2].

For the purpose to identify the corresponding segments in original and translated parts the texts, which are included in the corpus, should be aligned on the sentence level.

The corpus is an open source for multitask research. Therefore, we cared about inclusion in its structure the broad scope of the linguistic information with careful check of the results of automatic alignment, morphologic and syntactic markup. With the same purpose, we adopt the corpus PEST-INTER for the multi-level extraction of statistics (the expanded set of association measurements for extraction of n-grams of different length).

---

<sup>5</sup> Ibid.

## *2. Plan of the study*

In my research, I stayed on the following plan. Firstly, I will focus on the computer-assisted keyword retrieval of links for sources with international treaties of the labor organization (ILO) in French, English, Russian, Swedish and Finnish. The automatic compilation of the list of key-sources, which were used by web crawlers collecting texts. Secondly, I will develop the system of web crawlers compiling the corpus of international treaties in five languages following the selected links. The embedding of the extracted texts in the structure of the corpus PEST-INTER. The aim was to build the system of continuous updating of the corpus. This system would collect the related texts and incorporate them into the corpus. Thirdly, I will manage the manual 'blind' review of the collected materials for the corpus PEST-INTER according to the following requirements: easiness of parsing, minimum of grammatical errors and character mismatches, representativeness of texts (presence of different styles). Fourthly, I will make the tuning of statistical measurements in the obtained corpus of international treaties PEST-INTER (i.e. development and adaptation of the basic rules for the algorithm detecting the n-grams). Fifthly, I will focus on detection of errors, its classification and deleting.

## *3. Compiling the corpus PEST-INTER*

### *3.1. Principles of the selection of texts. The structural features of the texts of international treaties*

The corpus PEST-INTER is developed mainly for studies in the field of automated extraction of LRD from the languages with different structure and development of the technique of building the multilingual glossaries. Bearing in mind this task my decision was to collect the texts of international treaties in five languages (English, French, Russian, Finnish and Swedish) with an initial capacity of two million tokens (words), which should be extended further. Documents of the International Labor Organization (ILO) were the material for this purpose.

The first factor, which affected our choice, was the fact that the international documents of the ILO organization include a huge variety of contracts,

conventions, agreements, charters, declarations, memorandums, communique, governmental decrees and protocols of different international conferences translated into more than 30 languages.

The quality of the selected material predefined the second factor. In my case study, the main requirements to the quality of texts collected for extraction of LRD correspond to the rules developed for the translation of the international documents. The international treaties should have close translations. Working with the international documents the translator is responsible for the translation accuracy. The correctness of the important political decisions, the authenticity of the contracts and agreements depend on the quality of translations. These requirements are equally valid for not only the uniformity of contents of the documents, but also for the concise transfer of the details, which, at first glance, seem to be of limited scope. In addition, it is necessary to note that translations of international treaties are usually made by professionals with respect for all necessary procedures (i.e. clear and concise character, laconicism, logic representation and coherence) [6].

Hence, the main requirements to the quality of the extracted texts are as follows: *accuracy* – all provisions should be explained excluding double sense; *clarity* – laconicism of language should not affect the completeness of the transferred sense; *conciseness* – all provisions should be lapidary; *style compliance* – the text should satisfy the standards of the style of the source language (SL).

### 3.2. Sources

The international treaties of the ILO organization are the material for the corpus PEST-INTER. These documents are publicly available in electronic form. The main sources from where they were taken are below:

- 1) *The databank NORMLEX*<sup>6</sup> (2018) – texts of treaties of the ILO organization in Russian, English and French;
- 2) *The databank FINLEX*<sup>7</sup> (2018) – texts of treaties of the ILO organization in Finnish;

---

<sup>6</sup> NORMLEX, Information System on International Labour Standards (2008-2018) [Online] [Accesssed on 4th January 2018] <http://www.ilo.org/dyn/normlex/en/?p=NORMLEXPUB:1>

<sup>7</sup> Data Bank FINLEX (1998-2018) [Online] [Accesssed on 5th January 2018] <https://www.finlex.fi/fi/>

- 3) *The website Svenska ILO kommittén*<sup>8</sup> (2018) – texts of treaties of the ILO organization in Swedish;
- 4) *Wikipedia Section «Luokka: Kansainväliset sopimukset»*<sup>9</sup> (2014) – separate articles of treaties of the ILO organization in Finnish;
- 5) *The databank EURLex*<sup>10</sup> (2018) – single treaties of the ILO organization in one or several of five stated languages.

### 3.3. Extraction methods

The main method of extraction was the development of automated systems for collecting and sorting the texts of international treaties from the Internet.

The programming language was Python, because of its laconicism, readability and high acceptance by the scientific community. The fact that there is a large set of libraries for linguistic purposes in this language further supported this choice.

I built five crawlers with the purpose of unifying them into one system, which allowed me to collect texts for the PEST-INTER corpus.

The first crawler retrieved all links to the ILO organization's treaties in one or several of the five stated languages. The retrieval was based on keywords.

From 52 retrieved links, the *five* most informative ones were selected. Due to differences in the HTML code structure, text formats and other particularities, each source required a different crawler.

The crawler processing worked as follows – my algorithm defined the strings in the sources' HTML code that specified the contents, metadata and language of each individual treaty. Then, it stored them in a given folder. After developing a basic algorithm on each single source, I obtained the texts that served my purpose and were to be included in my corpus. I developed the automatic extraction of the titles of treaties and the function that split the derived documents into paragraphs. In addition, I tuned the automatic sorting of the retrieved documents according to the source (target) language and its sequence

---

<sup>8</sup> Swedish Information System on International Labour Standards (2000-2018) [Online] [Accessed on 6th January 2018] <http://svenskailo-kommitten.se/>

<sup>9</sup> Wikipedia «Luokka: Kansainväliset sopimukset» (2014) [Online] [Accessed on 7th January 2018] [https://fi.wikipedia.org/wiki/Luokka:Kansainv%C3%A4liset\\_sopimukset](https://fi.wikipedia.org/wiki/Luokka:Kansainv%C3%A4liset_sopimukset)

<sup>10</sup> EUR-Lex data for commercial or non-commercial purposes (1998-2018) [Online] [Accessed on 9th January 2018] <http://eur-lex.europa.eu/>

number. Hence, I obtained 200 folders, each containing five documents in TXT or PDF (not edited) format. The documents in PDF (not edited) format were recognized using the ABBYY FineReader OCR program<sup>11</sup> and saved in TXT format. Any non-recognized characters were added manually. Then, the texts were added to the main folder.

Collecting the documents with their translations in all five languages, creating acceptable TMX files and incorporating them in the PEST-INTER corpus database turned out to be a tough challenge. In order to minimize the need to cut extracted documents for which at least one of the five languages was missing, my algorithm used Google Translate<sup>12</sup> to translate the titles of the extracted documents. In this way, based on the translated titles, we were able to locate the missing translations on the Internet and add them to the main massive. This method allowed to reduce the list of the incomplete documents considerably. Another set of translations of the treaties (~ 2%) was found, recognized and added manually. The last method was applied in the cases, where it was inappropriate or impossible to use the methods of automated extraction.

#### *4. Editing and loading the texts in the corpus PEST-INTER*

In order to facilitate the processing of the collected texts and the extracting from them a necessary linguistic information a number of structural problems (i.e. a character encoding, sentence alignment, insertion of the metadata, POS markup, etc.) were solved.

##### *4.1. Character encoding*

Using the separate encoding type introduces several restrictions in the extraction of the single characters and their interoperability with other encoding types [1]. Therefore, it is more appropriate to operate with a multilingual array saved in XML format [8]. Hence, all automatically retrieved documents were saved in UTF-8 before creating TMX<sup>13</sup> files for the subsequent processing.

---

<sup>11</sup> ABBYY FineReader OCR (1989-2018) [Online] [Accessed on 15th February 2018] <https://www.abbyy.com/en-eu/finereader/>

<sup>12</sup> Google LLC (2018) [Online] [Accessed on 15th February 2018] <https://translate.google.fi/?hl=ru&tab=TT>

<sup>13</sup> TMX (Translation Memory eXchange) is an XML specification for exchange of translation memory between various CAT (computer-assisted translation) and localization tools.

#### 4.2. Sentence alignment

At this stage, I loaded, integrated and aligned the TXT files in five languages with UTF-8 encoding on sentence level by means of the LFAAligner program<sup>14</sup>. The most part of this work was realized automatically thanks to high-quality of the documents extracted by the automated systems at the previous stage. Nevertheless, some sequences had to be corrected manually by means of Alignedit<sup>15</sup> application.

Main corrections:

1. The coercion of the titles of conventions in five languages to the uniform format (Number-Title-Year);
2. The correction of the errors of the alignment;
3. The replacement of not translated sentences with the sign "XXX";
4. The correction of the errors in words, which were not recognized;
5. The deleting of the character mismatches mostly in the Swedish treaties (most of these treaties were initially in PDF format);
6. The completing of the Finnish treaties. I did not succeed to find the treaties containing all articles. Practically each treaty contained initially a number of references to the articles, which were translated in the other sources. Collection of the missed information was carried out separately, and the additions were made manually at the stage of editing;
7. The adding of the uniform format of the language tags containing the information about the language of the document;
8. The deleting of the blank lines.

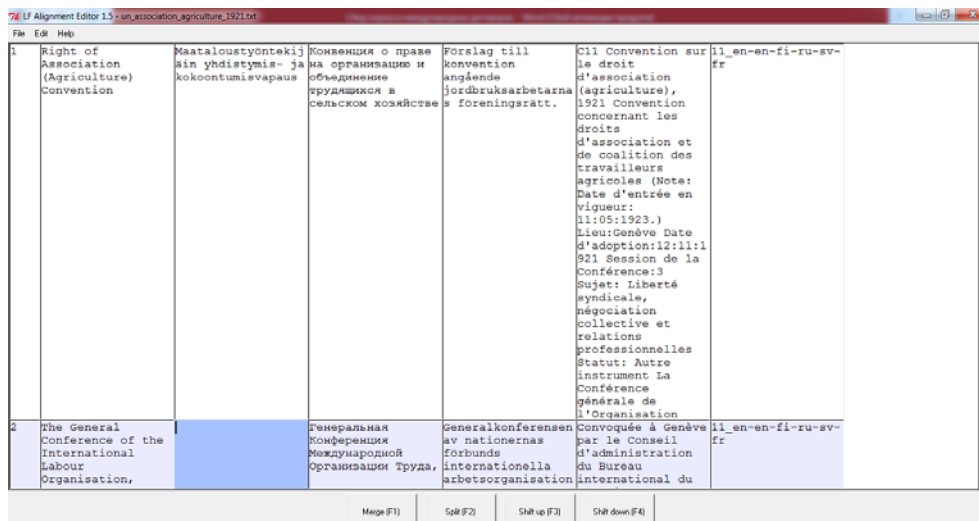
---

<sup>14</sup> LFAAligner software (2018) [Online] [Accesssed on 15th January 2018] <https://sourceforge.net/projects/aligner/>

<sup>15</sup> software is included in the package LFAAligner




## Compilation of the parallel corpus of international treaties



1	Right of Association (Agriculture) Convention	Maataloustyöntekijain yhdistymis- ja kokoontumisvapaus	Конвенция о праве на организацию и объединение трудящихся в сельском хозяйстве	Förslag till konvention angående jordbruksarbetarnas föreningsrätt.	C11 Convention sur le droit d'association (agriculture), 1921 Convention concernant les droits d'association et de coalition des travailleurs agricoles (Note: Date d'entrée en vigueur: 11/05/1923.) Lieu:Genève Date d'adoption:12/11/1921 Session de la Conférence:3 Sujet: Liberté syndicale, négociation collective et relations professionnelles Statut: Autre Instrument La Conférence générale de l'Organisation	11_en-en-fi-ru-sv-fr
2	The General Conference of the International Labour Organisation,		Генеральная Конференция Международной Организации Труда,	Generalkonferensen av nationernas förbunds internationella arbetsorganisation	Convoquée à Genève par le Conseil d'administration du Bureau International du	11_en-en-fi-ru-sv-fr

Figure 1. Document in the program Allignedit before the correction was made



1	Right of Association (Agriculture) Convention	Maataloustyöntekijain yhdistymis- ja kokoontumisvapaus	Конвенция о праве на организацию и объединение трудящихся в сельском хозяйстве	Förslag till konvention angående jordbruksarbetarnas föreningsrätt.	Convention sur le droit d'association (agriculture)	11_en-en-fi-ru-sv-fr
2	The General Conference of the International Labour Organisation,	Kansainvälisen työjärjestön yleiskokous,	Генеральная Конференция Международной Организации Труда,	Generalkonferensen av nationernas förbunds internationella arbetsorganisation,	La Conférence générale de l'Organisation internationale du Travail,	11_en-en-fi-ru-sv-fr
3	Having been convened at Geneva by the Governing Body of the International Labour Office, and having met in its Third Session on 25 October 1921, and	Kansainvälisen työtoimiston johtokunta on koonnut Genevessä ja tapasi kolmannessa istunnoissaan 25 lokakuuta 1921, ja	созванная в Женеве Административным Советом Международного Еуро Труда и собравшаяся 25 октября 1921 года на свою третью сессию,	vilken av styrelsen för internationella arbetabyrån sammankallats till Genève och där samlats den 25 oktober 1921 till sitt tredje sammanträde	Convoquée à Genève par le Conseil d'administration du Bureau international du Travail, et s'y étant réunie le 25 octobre 1921, en sa troisième session,	11_en-en-fi-ru-sv-fr
4	Having decided upon the adoption of certain proposals with regard to the rights of association and combination of agricultural workers, which is included in the fourth item of the agenda of the Session, and	Päättaessään tietyjen ehdotusten hyväksymisestä yhdistymisvapauden ja maataloustyöntekijöiden yhdistämisen osalta, joka sisältyy istunnon esityslistan neljanteen kohtaan ja	постановив принять ряд предложений о праве на организацию и объединение трудящихся в сельском хозяйстве, что является четвертым пунктом повестки дня сессии,	samt beslutit - antaga vissa förslag angående jordbruksarbetarnas förenings- och församlingsrätt, vilken fråga innefattas i fjärde punkten på dagordningen för sammanträdet,	Après avoir décidé d'adopter diverses propositions relatives aux droits d'association et de coalition des travailleurs agricoles, question comprise dans le quatrième point de l'ordre du jour de la session, et	11_en-en-fi-ru-sv-fr
5	Having determined that these proposals	Päättaessään, että nämä ehdotukset ovat	решии придать этим предложениям статус	ävensom beslutit, att dessa förslag skola	Après avoir décidé que ces propositions	11_en-en-fi-ru-sv-fr

Figure 2. Document in the program Allignedit after the correction was made

When the alignment was made, the obtained files were saved in the TMX format by means of TMX maker 3.0<sup>16</sup>. In addition, I manually added the metadata considered in structure of the corpus PEST-INTER.

<sup>16</sup> Ibid.

```

<?xml version="1.0" encoding="utf-8" ?>
<!DOCTYPE tmx SYSTEM "tmx14.dtd">
<tmx version="1.4">
  <header
    creationtool="LF_TMX_maker"
    creationtoolversion="3.0"
    datatype="unknown"
    segtype="sentence"
    adminlang="EN"
    srclang="EN"
    o-tmf="Tw4win 2.0 Format"
  >
  </header>
  <body>
<tu creationdate="20171228T170829Z" creationid="LF_TMX_maker_3.0"><prop type="Ttxt::Note">11_en-en-fi-ru
<tuv xml:lang="EN"><seg>Right of Association (Agriculture) Convention</seg></tuv>
<tuv xml:lang="FI"><seg>Maataloustyöntekijäin yhdistymis- ja kokoontumisvapaus</seg></tuv>
<tuv xml:lang="RU"><seg>Конвенция о праве на организацию и объединение трудящихся в сельском хозяйстве<
<tuv xml:lang="SV"><seg>Förslag till konvention angående jordbruksarbetarnas föreningsrätt.</seg></tuv>
<tuv xml:lang="FR"><seg>Convention sur le droit d&apost;association (agriculture)</seg></tuv> </tu>

<tu creationdate="20171228T170829Z" creationid="LF_TMX_maker_3.0"><prop type="Ttxt::Note">11_en-en-fi-ru
<tuv xml:lang="EN"><seg>The General Conference of the International Labour Organisation,</seg></tuv>
<tuv xml:lang="FI"><seg>Kansainvälisen työjärjestön yleiskokous,</seg></tuv>
<tuv xml:lang="RU"><seg>Генеральная конференция Международной Организации Труда,</seg></tuv>
<tuv xml:lang="SV"><seg>Generalkonferensen av nationernas förbunds internationella arbetsorganisation,<
<tuv xml:lang="FR"><seg>La conférence générale de l&apost;organisation internationale du Travail,</seg><

<tu creationdate="20171228T170829Z" creationid="LF_TMX_maker_3.0"><prop type="Ttxt::Note">11_en-en-fi-ru

```

Figure 3. Document in the format TMX before the metadata was added

```

<?xml version="1.0" encoding="utf-8" ?>
<!DOCTYPE tmx SYSTEM "tmx14.dtd">
<tmx version="1.4">
  <header
    creationtool="LF_TMX_maker"
    creationtoolversion="3.0"
    datatype="unknown"
    segtype="sentence"
    adminlang="EN"
    srclang="EN"
    o-tmf="Tw4win 2.0 Format"
  >
  </header>
  <body>
<textdef code="un_association_agriculture_1921_en" title="Right of Association (Agriculture) Convention
<textdef code="un_association_agriculture_1921_fi" title="Maataloustyöntekijäin yhdistymis- ja kokoontu
<textdef code="un_association_agriculture_1921_ru" title="конвенция о праве на организацию и объединени
<textdef code="un_association_agriculture_1921_sv" title="Förslag till konvention angående jordbruksarb
<textdef code="un_association_agriculture_1921_fr" title="Convention concernant les droits d'associatio

<tu creationdate="20171228T170829Z" creationid="LF_TMX_maker_3.0"><prop type="Ttxt::Note">11_en-en-fi-ru
<tuv xml:lang="EN"><seg>Right of Association (Agriculture) Convention</seg></tuv>
<tuv xml:lang="FI"><seg>Maataloustyöntekijäin yhdistymis- ja kokoontumisvapaus</seg></tuv>
<tuv xml:lang="RU"><seg>Конвенция о праве на организацию и объединение трудящихся в сельском хозяйстве<
<tuv xml:lang="SV"><seg>Förslag till konvention angående jordbruksarbetarnas föreningsrätt.</seg></tuv>
<tuv xml:lang="FR"><seg>Convention sur le droit d&apost;association (agriculture)</seg></tuv> </tu>

<tu creationdate="20171228T170829Z" creationid="LF_TMX_maker_3.0"><prop type="Ttxt::Note">11_en-en-fi-ru
<tuv xml:lang="EN"><seg>The General Conference of the International Labour Organisation,</seg></tuv>

```

Figure 4. Document in the format TMX after the metadata was added

The metadata contained the name of the document in the source and target languages, the subject, the year of issue of the document and the year of its translation as well as the tag of the target language.

```
<textdef code="un_association_agriculture_1921_en" title="Right of Association (Agriculture) Convention" subject="labor_policy"
yearorig="1921" yeartr="1921" lang="en" />
<textdef code="un_association_agriculture_1921_fi" title="Maataloustyöntekijäin yhdistymis- ja kokoontumisvapaus"
subject="labor_policy" yearorig="1921" yeartr="1921" lang="fi" />
<textdef code="un_association_agriculture_1921_ru" title="Конвенция о праве на организацию и объединение трудящихся в сельском
хозяйстве" subject="labor_policy" yearorig="1921" yeartr="1921" lang="ru" />
<textdef code="un_association_agriculture_1921_sv" title="Förslag till konvention angående jordbruksarbetarnas föreningsrätt"
subject="labor_policy" yearorig="1921" yeartr="1921" lang="sv" />
<textdef code="un_association_agriculture_1921_fr" title="Convention concernant les droits d'association et de coalition des travailleurs
agricoles" subject="labor_policy" yearorig="1923" yeartr="1923" lang="fr" />
```

Figure 5. Extract of the metadata added manually to the files in TMX format

### 4.3. Corpus markup and loading in the server

When all TMX documents were edited, I loaded them on the puolukka<sup>17</sup> server by means of the client SSH Tectia – Secure File Transfer.<sup>18</sup> Further, by means of the publicly available client PuTTY<sup>19</sup> I obtained the access to the server. Before the automated syntactic markup of my array was made by means of the parsers installed in the server, I should solve one more problem connected to the predefined requirements for all files, which I would load into the corpus PEST-INTER. With the purpose to increase the corpus processing and to exclude possible system mistakes, I restricted length of metadata up to 50 characters. Hence, I needed to rename all titles which length exceeded the predefined. This work was performed by means of the editor VIM<sup>20</sup> that allowed to rename automatically all metadata. By means of the command *///*, I changed the names of the files edited in accordance with the correct metadata.

---

<sup>17</sup> Mikhailov, Mikhail (2016) 'PEST, Parallel Corpus of State Treaties. University of Tampere.' [Online] [Accessed on 4th March 2018] <https://www.puolukka.uta.fi>

<sup>18</sup> Tectia® Client 6.4. User Manual (1995–2017) [Online] [Accessed on 15th January 2018] <https://www.ssh.com/manuals/client-user/64/index.html>

<sup>19</sup> PuTTY: latest release (0.70) (2018) [Online] [Accessed on 15th January 2018] <https://www.putty.org/>

<sup>20</sup> Vi/Vim Editor. User Manual (2018) [Online] [Accessed on 15th January 2018] <http://help.ubuntu.ru/wiki/vim>

At the final stage I created the load list of all files by means of the ls command `ls *.xml > load.bat` and the texts were loaded in the corpus PEST-INTER on the puolukka server.

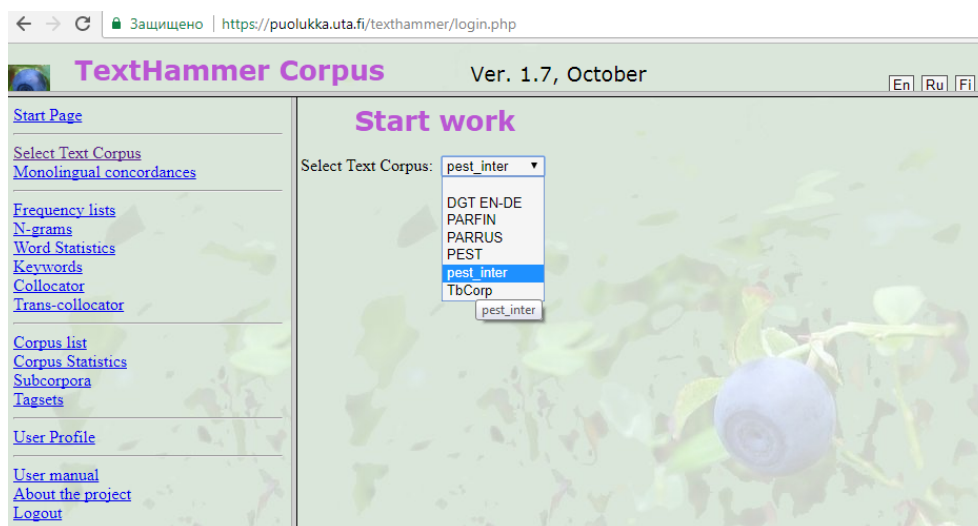


Figure 6. PEST-INTER corpus in the database PEST

The screenshot shows the 'Corpus list' section of the TextHammer Corpus interface. It features a table with columns for 'Code', 'Author', 'Translator', 'Text', 'year', 'Publisher', and 'Language'. The table lists several ILO conventions in five languages (en, fi, ru, sv, fr). The background of the interface is a decorative image of blueberries.

Code	Author	Translator	Text	year	Publisher	Language
mini_age_sea_1936_en			Minimum Age (Sea) Convention (Revised 1936)	1936		en
mini_age_sea_1936_fi			Meriyökön käytettävien lasten minimi-ikä vahvistaminen (muutettu, 1936)	1936		fi
mini_age_sea_1936_ru			Компенсия о минимальном возрасте допуска детей на работу в море (пересмотренная в 1936 году)	1936		ru
mini_age_sea_1936_sv			Förslag angående fastställande av minimialder för barns användande i arbete till sjöss (reviderad 1936)	1936		sv
mini_age_sea_1936_fr			Convention (révisée) sur l'âge minimum (travail maritime)	1936		fr
un_association_agriculture_1921_en			Right of Association (Agriculture) Convention	1921		en
un_association_agriculture_1921_fi			Maaataloustyöntekijän yhdistymis- ja kokoomusvapaus	1921		fi
un_association_agriculture_1921_ru			Компенсия о праве на организацию и объединение трудящихся в сельском хозяйстве	1921		ru
un_association_agriculture_1921_sv			Förslag till konvention angående jordbruksarbetarnas föreningsrätt	1921		sv
un_association_agriculture_1921_fr			Convention concernant les droits	1921		fr

Figure 7. Extract of the list of ILO conventions in five languages in PEST-INTER corpus

Now the corpus PEST-INTER contains over two million tokens and available for carrying out the experiments with n-gram models. The corpus is constantly growing. The next update will be done on September 2018.

Thanks to the functions of the specially developed corpus manager TopicWords several tools are available to researchers, including the system of search, the function of addition and extension of the research material, the functions of extraction of different statistical information. Access to them is provided on demand.

### *Concluding remarks*

In this article, I provided a number of special cases within the development of the corpus PEST-INTER. These problems become the burning issue nowadays. In addition, I described the difficulties, which most of young researchers encounter at the stage of the data collection and its design in the electronic corpus. I gave the answers to such questions as:

- What features are considerable in writing algorithms for data collection from the Internet?
  - What format is the most convenient for the purposes of extraction of LRD (i.e. UTF-8 for character encoding and TMX for corpora as whole)?
  - What are the particularities of alignment and editing of the documents to store in the database?
  - How to build the corpus relevant to the large scale of the research tasks?
- In the initial stage of the data collection, I aimed to compile the corpus, which can reconcile not only the specific requirements of my research interests, but also can be used by other researchers for their purposes. In this regard, the systematic analysis of the structure and the features of the corpus mentioned above has not only theoretical, but also a practical impact.

In addition, I underline the fact that the project has an open character. Writing this article, I continue to develop the corpus PEST-INTER. When my research on n-gram based extraction of LRD will be done, the corpus development will last on the principles of openness and accessibility of the materials thanks to the corpus manager TopicWords providing the function of continuous data collection.

Results of my case study are as follows:

- Development of the system of crawlers, collecting the data for the corpus of international treaties PEST-INTER;
- Development of the open corpus of international treaties PEST-INTER in five languages;
- Integration of the functions of the corpus manager TopicWords into the corpus PEST-INTER;

Further work:

- To develop fully automated system of corpus compilation;
- To add the treaties in 5 languages from other fields of the Law in the corpus PEST-INTER;
- To tune an algorithm for a ratio of statistics and lengths of n-gram in the corpus PEST-INTER.

In conclusion, I would note that this article could be a source for further researches, as well as it could be appreciated as the instruction for creation of other corpora.

#### *Acknowledgments*

I am grateful to professor Mikhail Mikhailov I have had the pleasure to work during this project. He has provided me extensive personal and professional guidance and taught me a great deal about scientific research. I would especially like to thank Juho Härme, PhD student of COMS. As my colleague and friend, he has contributed a lot in all the stages of my case study. The research is supported by the grant 10.04.2017/TM-17-10530/EDUFI Fellowship/WS21.

#### *Bibliography*

- Fantinuoli C., Zanettin F. (2015) *New directions in corpus-based translation studies*. Berlin: Language Science Press.
- Grinev-Grinevich S.V. (2008) *Terminology*. Moscow: Academy.
- McEnery T., Xiao R., Tono Yu. (2006) *Corpus-based language studies: An advanced resource book*. Taylor & Francis.
- Mikhailov M., Cooper R. (2016) *Corpus Linguistics for Translation and Contrastive Studies: A Guide for Research*. London & New York: Routledge.

- Parra Escartín C. (2013) 'Encoding a parallel corpus: The TRIS corpus experience'. *The many facets of corpus linguistics in Bergen*, 3, № 1, Bergen: Bergen Language and Linguistics Studies (BeLLS), pp. 61-80.
- Schweitzer A.D. (1988) *Translation theory. Status, problems, aspects*. Moscow: Nauka.
- Tiedemann J. (2007) 'Building a Multilingual Parallel Subtitle Corpus'. In Dirix, P., Schuurman, I., Vandeghinste, V. & Van Eynde, F. (eds.). *Proceedings of CLIN 17*, pp. 147-162.
- Tufis D., Steinberger R., Pouliquen Bruno, Widiger A., Ignat C., Erjavec T., Varga D. (2006) 'The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages'. In *The 5<sup>th</sup> International Conference on Language Resources and Evaluation – Proceedings*, pp. 2142-2147