European Master in Lexicography

Master's thesis

Lexicographic resources and machine translation. Challenges and perspectives

Student: Olha Novikova

Signature

Supervisor: Pablo Gamallo Otero

Signature GAMALLO OTERO PABLO - 36099559R Firmado digitalmente por GAMALLO OTERO PABLO - 36099559R Fecha: 2018.06.23 01:28:45 +02'00'

July 2018

Master's thesis presented at the Faculty of Philology of the University of Santiago de Compostela to obtain a Master's degree in Lexicography

Summary

In this thesis we describe and evaluate a tool for automatic generation of translations for multiword English terms into Spanish from a monolingual specialized Spanish corpus, compiled by means of web crawling. The resulting translations may be further used to expand or to revise the existing lexicographic resources, i.e. language for specific purposes dictionaries, terminological databases and translation memories. We evaluate the output with the precision and recall metrics and apply our approach to the small English-Spanish glossary in the legal domain that serves both as a reference dictionary and a source for input multiword terms. We analyze the obtained results and suggest possible solutions to the detected problems based on the theoretical part of this research devoted to the issues of machine translation, the use of dictionaries and corpora (non-parallel corpora in particular) for machine translation tasks, and the works on automatic terminology extraction.

Keywords: multiword expression, multiword term, corpus-based machine translation, rule-based machine translation, hybrid machine translation, lexicographic resource, parallel corpus, comparable corpus, terminology extraction, precision, recall, syntactic pattern, transfer rule.

Plan:

Introduction

In the current research we were motivated by the latest advances in the field of computational linguistics and terminology aimed at improving methods and tools for automatic compilation of lexicographic resources for human and machine translation tasks. We paid special attention to the research projects based on non-parallel corpora, since aligned parallel corpora, despite being very useful, are a scarce and an expensive resource. Moreover, we have tried to suggest an approach that would not be very sophisticated in terms of technology and could be used for different domains. The tool described in this paper has been designed to generate automatic translations of English multiword terms into Spanish by exploring a bilingual dictionary and a non-parallel (monolingual) corpus, which is practical for two reasons: 1) the majority of terms are multiword expressions, 2) non-parallel corpora are easier to get than parallel corpora.

It is also worth mentioning, that in this work we use the terms “multiword expression” and “multiword term” as mutually intelligible in most contexts. However, a multiword term (MWT) only applies to multiword expressions that are terms in the specific domain, while a multiword expression (MWE) is a more generic notion denoting all "idiosyncratic interpretations that cross word boundaries (or spaces)" (Sag et al., 2002) embracing both terms and non-terms.

Working on this paper we have set the following objectives:

- to investigate the past, present and future trends and challenges in the field of machine translation in the context of lexicography and terminography;
- to look into the ways the lexicographic resources (including dictionaries and corpora of different types) are employed in machine translation and to come up with an up-to-date definition of what constitutes a lexicographic resource of the XXI century;
- to explore the approaches to monolingual and bilingual terminology extraction from non-parallel corpora;
- to suggest a workable approach to build a bilingual lexicon (English-Spanish) from a monolingual (Spanish) corpus, and to experiment with web crawling for specialized corpus compilation;
- to identify and to analyze the limitations of our system in the process of evaluation;
- to suggest possible solutions and strategies for improvement of the method from the linguistic point of view.

To attain the abovementioned objectives, we started the paper with a brief overview of the history of machine translation and of the main approaches to machine translation

including rule-based (direct, interlingua, transfer systems), corpus-based (statistical, example-based, neural systems) and hybrid MT engines. We also explained the main principles of lexicographic resources compilation for machine translation systems, i.e. dictionaries for rule-based and corpora for corpus-based engines. The last section of the theoretical part deals with the issues of automatic terminology extraction without the use of parallel corpora. All theoretical data are supported with examples of related research works and projects we consider relevant and interesting for the purposes of this thesis. In the practical part we described the system architecture, the types of multiword terms (syntactic patterns) it is designed for, the evaluation methodology we use to assess its performance, and provided the quantitative table with the evaluation results. The limitations of different types were detected in the process of experiments we performed with the system and a small English-Spanish glossary of law terms. We summarize the practical part with the analysis of the weaknesses and search for remedies. We have included the list of literature sources as well as online and print reference dictionaries. The paper also contains two annexes: 1) the evaluation data set; 2) the list of abbreviations that appear throughout the paper.

Having classified the weaknesses of our approach according to their origin, we suggested possible solutions and improvements of the method, which could be taken as a basis for further academic research projects. Due to the system's high precision score, we argue that it has a practical value and could find its use for the tasks of automatic expansion and revision of LSP dictionaries and termbanks. In our opinion, the suggested improvements can also boost the system's performance significantly, especially increase its recall, making this method even more attractive for further research.

Chapter 1. Theoretical foundations.

1.1. Brief history of Machine Translation.

The processes of globalization and digital revolution marked a new period in the history of translation and lexicography. In the globalized world we can reach a different continent in less than a day, communicate with people from all over the world via social networks, make purchases without having to leave our house, and obtain a 3D live view of the places we will probably never visit. Notwithstanding, we still speak different languages and need to understand each other as never before. The Internet has created a demand for instant online translations, which human translators cannot possibly meet. The establishment of international organizations such as the United Nations, the Council of Europe, the European Court of Human Rights and the like gave growth to the market of translation and localization services and set a whole range of new challenges as well as opportunities for the ones employed in this sector. From now on the main task is translation of a huge amount of data into different languages as fast and as cheap as possible.

In this regard, we would like to underscore the interconnection of translation and lexicography, as the former is the direct user of lexicographic resources. On the other hand, the traditional perception of what constitutes a dictionary or a lexicographic resource has changed. With a shift to E-Lexicography we can no longer speak about a mere paper monolingual or bilingual dictionary, instead we should broaden the scope of lexicographic resources, including corpora (both parallel and comparable corpora, and web as corpus), online dictionaries and applications with multimedia items, translation memories and terminology databases, and a whole range of Natural Language Processing (NLP) tools which are based on the abovementioned lexicographic resources.

In rough terms, machine translation systems make use of lexicographic resources in a similar manner as human translators do. Hutchins (1995) defines "machine translation" (henceforth MT) as a set of computerized systems responsible for production of translations of natural languages with or without human assistance. These do not include computer-based translation tools which provide access to online dictionaries, terminology databases, transmission and reception of texts or the like.

The first attempts of MT may be traced back to the mid-20th century (Hutchins, 1995, 2006, 2014). Those begin in the 1930s with "translating machines" by a French-Armenian George Astrouni, and by a Russian Petr Smirnov-Troyanskii. The former invention constituted an automatic bilingual dictionary using paper tape; the latter was more significant

in terms of MT development and served to transform source language (SL) sentences analyzed by a human editor into the target language (TL) ones.

Taking into account the limited technical potential of the previous century, much effort was invested in improving hardware and in elaborating software for language processing. For political reasons MT research in the US was devoted to Russian-English translations of scientific and technical documents, and the Soviet experiments to English-Russian systems, correspondingly. For instance, the research at the University of Washington led by Erwin Reifler (in Hutchins, 1995) involved the compilation of large bilingual dictionaries, that included not only English translations of Russian lemmata, but also grammatical and syntactic rules for the output, phrases and clauses, and classification into sublanguages. In 1960s a research group at Harvard University worked on a Russian-English dictionary to produce word-for-word draft translations for the experts in the topic, which could be called a predecessor of modern computer-based dictionary aids for translators.

Despite the initial optimism the research in MT sphere was gradually coming to a standstill. MT itself was criticized by many scientists, among others Bar-Hillel and Victor Yngve (in Hutchins, 1995) as incapable of producing fully automatic translations of high quality due to "semantic barriers", i.e. the machine was not able to distinguish between different meanings and could not really understand any natural language the way humans did. To examine the current state of MT in the US the Automatic Language Processing Advisory Committee (ALPAC) was founded by the government sponsors of MT research. ALPAC issued its notorious report in 1966, referring to poor prospects of further funding of MT. On the contrary, it suggested investing into computational linguistic projects, particularly into the development of electronic dictionaries. The ALPAC report had a strong influence in the US, and virtually stopped the MT projects for almost ten years.

Consequently, the attention to MT in the 1970s was higher in Canada and Europe. On the one hand, the bicultural policy of the Canadian government stimulated a great demand for English-French translations. On the other hand, there existed an acute need within the European Community for rapid translations of administrative, legal and scientific documentation from and into the Community languages. The significant projects tackling these issues were:

- the TAUM (Traduction Automatique de l'Université de Montréal) project that involved the elaboration of METEO translation system, designed for the restricted vocabulary of weather forecasts (Hutchins, 2014).
- Systran operational system developed by Peter Toma in 1968. Its earliest version served

for translating Russian texts into English for the US Air Force during the Cold War. Systran was installed for other language pairs (French-English, English-Italian etc.) at a number of intergovernmental organizations such as the Commission of the European Communities, NATO and the International Atomic Energy Authority (Hutchins, 2006);

- the main rival of SYSTRAN was the commercial MT Logos system. Originally it was developed for German-English language pair and further expanded for other languages. Even though the dictionaries of the mainframe systems like SYSTRAN and Logos were adapted for specific subject domains such as aircraft engineering, the systems were designed for general application (Hutchins, 2006).

The availability of microcomputers and improved text-processing software during the 1980s reduced the price of MT. During this period the dominant MT strategy was the translation via intermediary representations with semantic, syntactic and morphological analysis and non-linguistic "knowledge bases". The DLT project, stands for Distributed Language Translation (Witkam, 2006), operated over computer networks where each terminal served for translation exclusively from and into one language. The output texts were transmitted between terminals in an intermediary language – a modified form of Esperanto. Combining linguistic (syntactic and morphological analysis) and extralinguistic (Esperanto database) information, the system computed probability scores for dependency-linked words. The project was a vital advancement in the creation of large lexical databases from a corpus of texts translated by a human, thus anticipating the corpus-based MT developments coming in the next decade.

Researchers at IBM suggested a method of MT based on the corpus of English and French texts of reports of Canadian parliamentary debates. Their method consisted in aligning phrases and words of the parallel texts and calculating probabilities that a word in a SL sentence corresponds to an aligned TL sentence. The result was then revised according to TL word-for-word frequencies obtained from the bilingual corpus. This method turned out to be a success, about a half of the phrases offered adequate translations. Since this time, statistical machine translation (SMT) based on IBM method became the focus of many MT research groups. The next corpus-based approach is known as "example-based machine translation" (further EBMT). It was first proposed by Makoto Nagao in 1981 (in Hutchins, 2006). The main idea of EBMT is that translation may be performed through examples based on previous translations. Thus, equivalent phrases were to be extracted from aligned parallel texts. The innovativeness of this approach was the use of real translations produced by professionals and, therefore, it could boast high-quality idiomatic translations.

At the same time, many professional translators recognized the importance of machine assistance in their daily work. This was not obligatory a fully-automated translation, but rather computer aids called “translation workstations” that aimed at making translation process less time-consuming and more precise. Those included dictionary creation, terminology management, concordancing tools, translation memories, and MT systems for translating texts or text segments. Among the earliest products were Trados (Translator’s Workbench) and STAR AG (Transit). During the 1990s and 2000s many more appeared: SDL, Xerox (XMS), Terminotix , MultiCorpora, MetaTexis, and ProMemori (Hutchins, 1995).

On the other hand, the Internet created a need of immediate translations of e-mails, web pages etc., thus turning MT into a mass-market service. The use of MT by non-professional users permitted sacrificing quality for the sake of speed and free of charge access. In the XXI century the main applications of MT are tools for domain-restricted terminology with large organizations as target users and the systems for non-translators (i.e. for users of numerous online services).

Modern applications of MT split into three large groups (Armentano Oller et al., 2007): assimilation, communication and dissemination. In the first case the MT output is used to convey a general idea of the original, and the translation quality is not that important as soon as this demand is met. A similar application is the communicative one, i.e. when speakers of different languages want to understand each other (e.g. via chat or e-mail) with the help of an MT system. Finally, dissemination requires translations of superior quality, for the output targets at general public. High precision and adequacy of translations for dissemination may be achieved by different strategies: use of controlled language, restricting the linguistic domain to one sublanguage (e.g. human rights), controlled vocabulary and syntax (avoiding polysemous words). Moreover, MT translations for dissemination generally require human post-editing. Nowadays, despite all the recent advancements in the field, high-quality MT translation is still an ambitious objective, which has, however, been attained at least for assimilation tasks. Thus, millions of Internet users are profiting from MT on a daily basis to translate web pages or to make purchases online. To conclude, in the globalized society with increased information production and consumption, the demand for MT is more evident than ever before, calling for new multidisciplinary tasks and professions in the fields of computer linguistics, translation and lexicography.

1.2. Main approaches to MT.

In terms of their basic characteristics, Hutchins (1995) distinguished bilingual (designed for two particular languages) and multilingual (more than one language pair) MT systems. The former may be unidirectional, i.e. a TL cannot be a SL and v.v. or bidirectional (any language may be a SL and TL). The majority of bilingual systems are unidirectional. There are two principal strategies for generation of MT systems: rule-based and corpus-based systems. Lately, the difference between MT approaches has become less significant, especially with the appearance of hybrid MT strategies, combining features of both rule- and corpus-based engines. Innovative and promising neural machine translation (NMT) approach gained recognition after 2014. In further sections we will discuss the given MT strategies in more detail.

1.2.1. Rule-based MT approaches.

Rule-based MT systems are classical MT systems based on linguistic data retrieved from lexicographic resources: monolingual and bilingual dictionaries, grammars covering main morphological, syntactic and semantic rules of the languages they treat. Such MT systems generate translations via morphological syntactic and semantic analysis of the SL and TL. Rule-based systems embrace the following approaches:

Direct translation is the oldest and the simplest type in MT history. SYSTRAN[1] was originally designed as a direct MT system. It is a bilingual unidirectional system designed specifically for one language pair that performs translation at the word level. The syntactic and semantic SL analysis is limited and tailored to the particular TL. The first generation MT systems were based on traditional lexicographic resources, i.e. large machine-readable and electronic dictionaries, and employed a direct translation approach. These dictionaries contained lists of words that served to express word senses, represent its meaning, or specify the syntactic frames in which a word can appear. This approach was very limited in its essence. It lacked an analysis of the internal structure of the original text and did not take into account the grammatical relationship between the principal sentence parts. The lack of computational knowledge at the beginning of MT research was also an issue.

The second generation MT systems employed more ambitious **interlingua** and **transfer approaches**. The former involves an intermediary semantico-syntactic representation of the SL text, i.e. interlingua, from which translations may be produced for different TL. Translation is a two-step process: from the SL to interlingua and from interlingua to the TL.

1. http://www.systransoft.com/systran/

Interlinguas may be based on artificial or auxiliary languages such as Esperanto or on natural languages (a set of semantic primitives common to all languages). The rules for SL analysis are oriented towards a particular SL, and not TL-specific. On the contrary, the procedures for TL generation are TL-specific and independent of the SL. In this respect, the interlingua approach is especially attractive for multilingual systems such as DLT (Witkam, 2006). This technique is not devoid of drawbacks either. Interlingua research has been devoted mostly to defining the universal and language-independent interlingua. It constitutes the main difficulty even for related languages (e.g. Spanish, Portuguese, Galician). Furthermore, it is complicated to extract meanings from the SL text to generate an interlingua representation.

Because of the complexity of the interlingua approach, less ambitious **transfer-based approach** was proposed. Unlike the previous one, it is designed for a specific language pair. Therefore, it doesn't require the resolution of all the SL ambiguities to enable translation in any TL in a way interligua MT does, but rather focuses on the resolution of the difficulties inherent to the languages dealt with. Transfer MT involves three steps: 1) analysis: SL texts are converted into abstract SL-oriented representations with the help of the SL parser; 2) transfer: these are converted into equivalent TL-representations; 3) generation: final TL output is generated with the TL morphological analyzer. Transfer systems usually consist of three types of dictionaries: monolingual SL dictionaries with morphological, grammar and syntactic data, similar TL dictionaries, bilingual dictionaries to relate SL and TL forms. Apertium[2] may be mentioned as an example of such system. One of the main problems of this approach is that rules must be applied for all three steps of translation.

1.2.2. Corpus-based MT approaches.

As we have already mentioned in the first section, since 1989 corpus-based MT becomes the dominant strategy in the field. It is an alternative strategy, aimed at overcoming the drawbacks of RBMT systems. As its name, suggests, corpus-based technique makes use of bilingual parallel corpora as a basis for translation. The main benefit of corpus-based MT systems is that they can learn translations of terminology and even stylistic phrasing from texts translated by human translators. Corpus-based systems split into Statistical Machine Translation (SMT) and Example-based Machine Translation (EBMT).

The idea of **Statistical machine translation (SMT)** is that any sentence in one language is a possible translation of any sentence in the other language. The best translation is picked based on the probability scores assigned by the system. There are various methods to train

2. https://www.apertium.org/index.eng.html?dir=cym-eng#translation

this, several start with automatically obtained word alignment and then collect phrase pairs of any length that are consistent with the word alignment. The SMT mechanism operates in the following way: first word sequences (referred to as "phrases" or "clumps" in SMT literature) of SL and TL texts are aligned, i.e. brought to correspondence. On the basis of these alignments a "translation model" of SL-TL frequencies and a "language model" of TL sequences are deduced. The core of the language model is the probabilistic phrase translation table that is learned from parallel corpora. Translation is done via selection of the most probable TL output for each SL word or phrase, and defining the most probable sequence of TL words. SMT depends on a language model, a translation model and a decoding algorithm. The translation model ensures that the MT system produces target hypothesis corresponding to the source sentence, in other words it provides for the faithfulness of translation. The language model ensures the grammatically correct results and the fluency of translation. The decoder uses the foreign string, heuristics and other methods to perform the search through all SL strings, limit the search space and at the same time to keep acceptable quality (Hutchins, 2005). Most systems are language-independent, and building a SMT system for a new language pair is mostly a matter of availability of parallel texts.

The issues of SMT to be tackled:

- parallel text alignment is needed to identify corresponding sentences in SL and TL. It is a challenging task, because in the process of translation sentences may be split, merged, omitted or reordered by a translator;
- statistical anomalies. Okpor (2014) provides a good example: the sentence "*I took a train to Berlin*" is often mistranslated as "*I took a train to Paris*", because "*train to Paris*" occurs much more often than "*train to Berlin*" in the training set;
- data dilution: taking into account a limited quantity of brand-specific corpora in the field of terminology, training sets are often used from alternative sources for a specific brand or domain. Those may confuse the terminology, text format and style;
- different word orders: SMT does not work well for the languages with significantly different word orders (e.g. European languages and Arabic);
- high costs of the parallel corpus creation. For this reason, many SMT researchers lately have focused on the application of non-parallel corpora (for more details see section 1.3.3. of the current paper).

Google translate[3] was launched as a SMT system. It makes use of the

3. https://translate.google.com/?hl=es

United Nations and European Parliament transcripts to calculate the probability of translations. It first translates the SL into English and then into any other TL.

The foundation of **example-based MT (EBMT)** is the idea of translation by analogy, i.e. examples of previously translated sentences are used to translate similar sentence types with the similar language pair. EBMT systems are trained on bilingual parallel corpora. The basic premise is that, if a previously translated sentence occurs again, the same translation is likely to be correct again (Carbonell et al., 2004). Alignment, matching and retrieval in EBMT are identical in its core to comparable methods in SMT. Some researchers, for example Somers (2003), even broaden the scope of the EBMT approach, embracing a range of approaches to corpus-based MT as its variants. The main tasks of EBMT: example acquisition from parallel bilingual corpus, example database management (i.e. how examples are stored), example application (i.e. decomposition of a SL sentence into examples and the conversion of source texts into target texts based on the database of examples), synthesis (i.e. enhancing the readability of the target sentence after applying the previous task).

A restricted form of EBMT is known as **translation memory**, which adds translations to the database, and when the similar sentence occurs again, it includes the existing example from the database into the new translation. The examples of translation memories are such tools as OmegaT[4], MateCat[5], SDL Trados Studio[6] and DGT-TM[7] – a freely available translation memory of the European Comission's Directorate General for Translation in 22 languages. Thus, we can think of both example and translation memories databases as lexicographic resources resembling an extended bilingual dictionary not restricted to the level of lemma. One of the advantages of EBMT systems is the expandability of their databases. Nevertheless, quantity does not always guarantee quality, as a large corpus may contain less frequent examples, and limit the reusability of the database for further translation jobs. Moreover, acquisition of examples requires analysis and generation modules to produce the dependency trees needed for the example database and for analyzing the sentence, which makes the example database expansion difficult. The example-based techniques were applied to various MT tasks. Sato (1993) applied EBMT of computer technical terms with respect to the focus term and its surrounding contexts and reported an overall accuracy of 96%, with 92% for unknown terms. Therefore, the EBMT approach may be especially attractive for

4. http://omegat.org/
5. https://www.matecat.com/about/
6. https://www.sdltrados.com/products/trados-studio/
7. https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory

domain-specific terminologies and the development of special purpose systems rather than for general applications. Furthermore, EBMT performance may be boosted by the use of a controlled language. On the other hand, this would require expansion or modification of the example database to comply with the rules of the controlled language. As a result, databases cannot be easily extracted from parallel corpora. When the controlled language enters into play, the translation result is regulated by certain rules, which the developers or users define. Controlled EBMT is, thus, closer to rule-based MT or, at least, to a hybrid MT approach (Hutchins, 2005).

Neural Machine Translation (NMT) is a newly emerging approach to MT. Its main difference as compared to traditional phrase-based MT, which comprises different sub-components (i.e. language model, decoder, translation model), is that it aims at building and training a single neural network that reads a sentence and generates its translation. Läubli (2017) argues that the main strengths of NMT when compared to other corpus-based approaches are as follows:

- it captures the similarity of words (both SMT and NMT replace words with numbers and perform mathematical equations on those numbers to generate translations. While SMT will assign different numbers to related words, NMT will give them similar values if the training data shows that their use is similar, as in the case with "*but*" and "*except*" which are mutually intelligible in a range of contexts);
- it considers the entire sentence (whereas SMT is limited to the number of words of its N-gram model, NMT evaluates fluency for the whole sentence. It might be especially useful for languages like German with long-distance syntactic dependencies between words);
- the way NMT systems are trained allows them to capture much more interdependencies between the complex features of particular languages (the sub-components of SMT systems mentioned above are learnt independently of each other and, thus, disregard the fact that for some languages, translation model is more important than language model or v.v. On the contrary, all NMT system components are trained jointly).

Nowadays, machine translators provided by such companies as Google, Yandex[8], DeepL[9] are using NMT. Harvard NLP research group launched OpenNMT[10] – an open source NMT initiative in 2016.

8. https://translate.yandex.com/
9. https://www.deepl.com/translator
10. http://opennmt.net

Notwithstanding, among the weaknesses of NMT are (Bahdanau et al., 2015; Läubli, 2017):

- its slower training speed (much more translated texts and computational resources are needed for training an NMT system than for a SMT system);
- poor efficiency with rare words;
- translation on a sentence by sentence basis only (when the MT system does something to one sentence, it doesn't know about the others, in the end the translated output is always a sequence of sentences, and not a coherent text. Läubli (2017) gives an example of this problem, which is a translation performed by DeepL NMT system: *This organism has dual capability. It can grow with either phosphorous or arsenic* (English) vs. *Dieser Organismus hat eine doppelte Fähigkeit. Es kann entweder mit Phosphor oder Arsen wachsen* (German). In this case the agreement of gender is not observed, since the sentences are translated independently. The correct translation into German would be: *Er kann…,* as the pronoun "*it*" in the SL refers to the noun "*the organism*", which is, however, masculine in German, but translated literally by its dictionary equivalent "*es*", a pronoun which is used only for neutral gender nouns in German. Those are the issues to be tackled by researchers in MT field in the nearest future.

1.2.3. Hybrid approaches to MT.

Hybrid machine translation (HMT), as its name suggests, combines the strengths of different approaches in order to improve the quality of the output. It is essentially designed to integrate the core of MT engines. Although nowadays the most popular hybrids are RBMT & SMT, HMT may embrace features of different MT systems.

Translations may be performed using a rule-based approach at the first stage and corrected with the help of statistical data. In some cases, rules are used to pre-process the original text and to post-process the statistical result of a SMT system to achieve more flexibility and control in translation. We should also mention corpus-based hybrids, i.e. SMT & EBMT hybrids. These techniques both make use of aligned corpora and statistical methods for matching and retrieval, thus, more examples of their hybrids will emerge, for example, augmentation of SMT aligned corpora with EMBT-like phrases.

Hutchins (2005) argues that the most "hybrid-friendly" approach is EBMT due to its openness to various techniques and methods. He mentions many different versions of EBMT & RBMT hybrids (e.g. core example-based transfer with RBMT for syntactic analysis), EBMT-SMT (e.g. core example-based transfer with SMT language model for recombination processes), RBMT systems with EBMT components (e.g. templates for dealing with collocations and idioms) etc.

The hybrid METIS-II MT system (Carl et al., 2008) avoids parallel corpora by using a bilingual dictionary (similar to that found in most RBMT systems) and a monolingual TL corpus. MT systems of PROMPT[11] as well as SYSTRAN initially using rule-based techniques, had switched to hybrid approaches by the end of 2010, including: rules post-processed by statistics and statistics guided by rules.

1.3. Lexicographic resources in MT.

As we have already mentioned at the beginning of this paper, the definition of what constitutes a lexicographic resource, especially in the context of MT, goes beyond traditional monolingual or bilingual dictionaries, albeit those play a significant role in the generation of translations by RBMT systems. Notwithstanding, in the current paper we define corpora (both aligned parallel and comparable non-parallel corpora) as well as termbanks that support corpus-based MT engines as lexicographic resources. In a similar manner the abovementioned lexicographic resources contribute to human translation, they also make possible the generation of translations by a machine. They constitute a core element in all MT engines and provide raw lexical material for future translations. Compilation of lexicographic resources for MT is an expensive and time-consuming task. Nowadays many companies are allocating labour force and resources to make it more efficient and less costly.

1.3.1. Dictionaries and other resources for RBMT.

A rule-based translation engine normally needs a minimum of three dictionaries, including: a SL dictionary used by the SL morphological analyzer to run morphological analysis, a bilingual dictionary used by the translator to map each SL word to an equivalent TL word, and a TL dictionary used by the TL morphological generator to generate words in the TL. The dictionaries in existing RBMT systems vary in terms of formats, coverage, level of detail and formalism of lexical description (Arnold et al., 1994). More than that, the contents of the dictionary vary significantly depending on the type of MT approach within the rule-based paradigm. Thus, dictionaries in an interlingua system do not need any translation information as their only task is to associate words with the appropriate interlingua concepts. On the contrary, transfer-based systems need information about SL items and their translations as well as TL data necessary to perform the certain transformation (e.g. in case of translation into English the placement of particles in phrases like "*look it up"* and "*look up the answer"*). A common practice in a transfer system is to separate monolingual SL and TL

11. https://m.online-translator.com/

dictionaries (involved in analysis and generation phase) from bilingual ones (these facilitate the transfer proper, relating source and target lexical items).

A dictionary used by an MT system possesses some distinctive features as opposed to a traditional paper or electronic dictionary for human users (henceforth referred to as traditional dictionary). Some types of items found in a traditional dictionary such as pronunciation are of little value for MT dictionaries, unless we deal with speech to speech systems, of course. It is also useful to draw a distinction between the items concerning the headword itself (i.e. its properties) and the restrictions it puts on other words in its linguistic environment. This distinction is not explicitly reflected in the traditional dictionaries, yet they still provide the information of both types. Information about the grammatical properties of a headword includes, for instance, such items as *gender, number* etc. Information about the grammatical environment usually splits into subcategorization (syntactic environment a headword may occur in, for example, categories of *transitivity/intransitivity*) and selection restrictions (semantic properties of this environment). In this regard, the marking *transitive* implies that transitive verbs such as '*to warp"* require a subject and an object. As dictionary users we also know that this object (patient) has to be something that can be wrapped, for example, a present or a person, and that the subject (agent) of the verb is usually animate. These data constitute selection restrictions, and appear implicitly in traditional dictionaries, thus, the information about the object may be worked out from the marking "*wrap something/someone"* (the object may be both animate and inanimate) and the accompanying examples. The entry never describes the subjects and objects explicitly, as it is assumed that human users may decipher it themselves. On the other hand, these data have to be made explicit to be recognized by the machine for the processes of analysis, transfer and generation performed by MT systems (Arnold et al., 1994).

1.3.2. Compilation and application of lexicographic resources.

Let's consider the workflow of dictionary compilation used by Logos (Dillinger, 2001), SYSTRAN (Gerber and Yang, 1997) and Apertium (Tyers et al., 2010) RBMT systems. The current tendency is to automate this process as much as possible to increase efficiency, and to reduce time and money spent on their development.

According to Dillinger (2001), for a general-purpose MT system we need, roughly estimated, between 40 000 - 100 000 entries (unless we deal with a controlled language environment). As we see, this number is rather blurred. Thus, the decision of how many headwords should be actually included into the dictionary is still to be made. As a rule most

entries are open-class words (i.e. nouns, adjectives, verbs and adverbs), with special focus on nouns and noun phrases making up the core of the dictionary. The key task here is to avoid terms that are rarely or never used by the system. A good strategy followed by Logos is to establish domain-specific lexical requirements based on a representative sample of domain texts. This can be done via crawling the Web in the following ways: a) to look for monolingual glossaries of the given domain and to merge several glossaries creating a representative list of SL terms to be included into the dictionary; b) to generate automatically a domain-specific corpus and to extract open-class terms. There are five sources for RBMT dictionaries: human experts, parallel corpora, print and online dictionaries. It is to be mentioned, that even though the RBMT systems do not generate translations based directly on the corpus, they may also benefit from corpora as a lexicographic resource that serves to compile reliable and up-to-date dictionaries. For instance, SYSTRAN traditionally used to develop its dictionaries based on a combination of several types of resources: extensive live text examples, and published linguistic and grammar reference works with the special attention to live text examples. The following so-called "fast coding" methods are used by SYSTRAN specialists to automate dictionary building process depending on the resource used: word lists (with non-electronic resources) with the desired meanings, that can be further run through the system to be modified or extended; when electronic dictionary resources are available, the entries are coded directly from them; when working with corpora, frequency lists and concordances are generated to identify high-frequency items to be addressed first.

Since the entries may originate from different resources, they are to be parsed and reformatted to a single format. In the past lexicographers used to check each entry for redundancy and validity, manually adding the needed information. That was not very efficient, because a lexicographer had to check every entry thoroughly, including the ones further subject to deleting. Nowadays Logos made a significant step forward by automating the evaluation processes. Only entries with non-canonical forms (such as nouns in plural, inflected verbs and the like), various explanations, forms derived from verbs are separated for off-line processing. This makes possible to identify the entries that require special treatment, and to import the ones that do not require it automatically. The final step is to check the performance of the dictionary in translation. Taking into account the diversity of languages and geographic expansion, it is important to collaborate with terminology teams in different countries. Logos is also experimenting with concordancing tools to get example sentences for the entries from corpus or from the Web. Even though these examples are not always helpful for identifying the correct sense of a word, they still provide contrastive contexts for different

word senses. At SYSTRAN dictionary validation process is a two-step procedure. The first step “edit” is the most basic technical check that verifies the correctness of the dictionary format. The second check “audit” validates the logic and consistency of codes in each entry. For example, it is to be insured that noun syntactic codes are not coded on verb entries or that English proper nouns are capitalized (Gerber and Yang, 1997). Before adding new entries another check should be performed to avoid duplicate entries in the dictionary or entries with incorrect cross-references to existing headwords. In addition to the entry validation, quality check of the whole dictionary as a system is required. To estimate the impact of new terms and modifications on the RBMT performance, it is essential to perform some type of regression testing, taking into consideration the following most common risks: a) what seems to be the best translation equivalent in the bilingual dictionary, will not a priori be a good translation in the MT output sentence; b) traditional dictionaries sometimes assign to words the parts of speech different to those assigned by live texts and real language usage; c) improvements of some aspects of the dictionary may lead to poor functionality of other aspects. The regression testing is usually run on a large sample (between 1,000-6,000 sentences) of varied texts.

As we have already mentioned when describing MT dictionaries, the need exists to deal with word-specific rules that are different from general rules. The strategy of SYSTRAN in this respect is to keep word-specific rules separated from the general ones. Thus, the former are included into the expression dictionary, and to some extent into the transfer programs called “lexical routines” which disambiguate polysemous words, and process high-frequency words or word classes which require special transfer processing (Gerber and Yang, 1997). The strategy of putting word-specific rules into the expression dictionary provides a good basis for further elaboration of generalizations. There are basically two types of SYSTRAN dictionaries: stem and expression dictionaries. The former contain lexical information for linguistic analysis, the latter complement the general rules for linguistic analysis and transfer with word-specific rules. Most of these dictionaries are multi-target, i.e. they provide information only for the source word or expression and for multiple target languages.

A typical entry contains a SL part, a TL part, and an extra field for transfer information on translation of prepositions. The stem dictionaries treat single words in the SL accompanied by detailed grammatical information on their morphology, semantics, syntactic behaviour, and homographs. The TL items also include morphology, syntactic behavior, part of speech, and translation of prepositions. Unlimited additional meanings may be assigned to headwords through linguistic programs or expression dictionaries. The stem dictionaries are to be used in

the analysis phase of translation. The items or, in MT terms, attributes included for the SL are: part of speech code, gender and number, homograph pattern, inflection or pattern code (to identify an entry with a table of valid inflections for matching inflected forms with the citation forms in the entry, parsing tense, person, number, and other information deduced from the inflection), semo-syntactic codes (such as: *ANSUB = Prefers animate subject (coded on verbs), ADVNN = Adverb may modify noun (coded on adverbs), GI = Word can govern infinitive (coded on various)* taken from Gerber and Yang, 1997), semantic tags, domain usage (vocabulary restricted to specific domains is tagged correspondingly) and document type (the same applies to vocabulary items restricted to specific document type, e.g. patents).

Let us consider semantic tagging by SYSTRAN stem dictionaries in more detail. The semantic codes are organized hierarchically in a way that permits lower modes to inherit semantic properties of the superior ones. For instance, the semantic category oral expression (ORALEX) is a lower node of the process (PROCES) and automatically inherits its properties. The semantic code for ORALEX is as follows: *SEM – ORALEX (SONIFY, TRANSM, GIVOUT, GIVE, PRPHY, ACT, PRGEN, PROCES).* The inherited codes are listed from the specific to the general:

SONIFY to make sound

TRANSM to transmit something

GIVOUT to give out or emit something

GIVE to give

PRPHY physical process

ACT action

PRGEN general process

PROCES process (Gerber and Yang, 1997).

The principal target language attributes in the stem dictionary are: preferences for translation of prepositions which govern or are governed by the SL headword (transfer field), TL meaning with morphology and syntax: part of speech, meaning identifiers (such as animate/inanimate nouns, transitive/intransitive verbs, reflexive/non-reflexive verbs etc.), inflection pattern, information on article use, gender and number, syntactic information (comparable with that of the SL, but without semantic part).

On the other hand, expression dictionaries may treat several types of entries, including:

- Idioms (multiword expressions which can then be integrated into the stem dictionaries as a single part of speech), for example: *out of the blue – out.of.the.blue (adverb).*
- Collocations (very useful for multiword noun terms), e.g. *learning disability, panic attack*

etc.

- Conditional expression (used when additional TL data should be added only under certain conditions depending on syntactic or semantic relationships, or semantic properties of the words), e.g. *If word is "eat" and direct object has semantic category METAL, translate as "corrode"*.
- Parsing expression: *If word is "turn" check right for "off". Force "off" to function as an adverb, resolve "turn" as a verb*.
- Homograph (these entries serve to disambiguate between homographs and assign the correct part of speech for each word).

As we have seen, dictionaries as such are not the only lexical resources used by RBMT systems. Apertium is an open-source platform for creating rule-based (transfer-based) MT systems, and containing stable data for multiple languages pairs. It makes quite a significant contribution to RBMT in general. Its operation is based on the following resources: finite-state morphologies for morphological analysis and generation, bilingual transfer lexica, probabilistic PoS taggers and a set of transfer rules (Tyers et al., 2010). All these resources are in a standardized format (as has been mentioned earlier, format is very important in the computer environment). Morphological dictionaries are generated in a format that enables encoding regularities as paradigms. Since there are many languages processed by the system, Apertium is not able to cover morphologies of each language to the full extent. Nevertheless, it is possible to integrate other free software. Some of morphological dictionaries are based on existing resources such as the Norsk Ordbank[12] (Norwegian wordbank), Eurfa[13] (the largest monolingual Welsh and bilingual bidirectional Welsh-Spanish dictionary under the free license), or Matxin[14] (public use MT system for the Basque language). The morphological dictionaries serve for morphological analysis of the SL text, i.e. segmenting the text in surface forms (words) and delivering one or more lexical forms for each word including lemma (citation form), lexical category, morphological inflection data. Morphological generator is also based on these resources. It basically does the same as the analyzer, but in the direction of the TL, i.e. delivering a TL surface form for each TL lexical form by suitably inflecting it. In addition to morphological dictionaries, Apertium also employs a number of bilingual lexica in XML format. These represent correspondences between headwords, including MWE, parts of speech, and sometimes morphological data such as changes in inflection formation from

12. https://www.hf.uio.no/iln/om/organisasjon/edd/forsking/norsk-ordbank/
13. http://eurfa.org.uk/
14. http://wiki.apertium.org/wiki/Documentation_of_Matxin_1.0

the SL to TL.

The lexica are used by the lexical transfer module of the system which reads its SL lexical form, delivers its TL correspondence, and looks it up in a bilingual dictionary. There are also statistical PoS taggers trained on corpus and a tagger definition file in XML which choose the most likely lexical form corresponding to an ambiguous surface form (Tyers et al., 2010). The tagger definition file serves to specify how the lexical forms resulting from the morphological analysis must be grouped into coarse PoS tags (such as N for nouns or J for adjectives). Each coarse tag is defined by a list of fine-grained tags based on the lemma and morphological information about it (such as NN for singular or mass nouns, NNS for plural nouns, NNP for proper nouns in singular, and NNPS for proper nouns in plural). Additionally, it is possible to set constraint rules – forbid and enforce rules. While the former specify restrictions such as sequences of two coarse tags that cannot occur together, the latter define the set of coarse tags that may occur after a particular coarse tag. There are also transfer rules used at three transfer stages: chunker (splits the sequence of lexical units into chunks, e.g. noun phrase or verb phrase), interchunk (performs more global operation inside the chunk and between different chunks) and postchunk (performs finishing operations on each chunk and produces a sequence of lexical forms). Chunker rules deal with, for instance, number and gender agreement in noun phrases, lexical changes of prepositions and other local phenomena. Interchunk rules are normally used for chunk merging and reordering. Finally, postchunk rules are applied for internal adjustment after the application of the rules mentioned above.

The dictionaries of RBMT systems such as Apertium may be further used to compile bilingual vocabulary lists, which added to parallel corpora, serve for training SMT systems. These lists aim at improving performance for languages with rich inflectional system, when using a small corpus or a corpus for a limited domain. Adding dictionaries to SMT systems also improves word alignment since one-to-one word mappings are provided there. There is a potential of performance improvement of Apertium and other MT systems via hybridization. For instance, one strategy is an integration of sub-sentential translation units (in other words bilingual chunks) into Aperitum RBMT engine (Sánchez-Martínez et al., 2009). Thus, bilingual chunks were generated automatically from parallel corpora with the help of chunkers and aligners used in the example-based MT system. These bilingual chunks may be also extracted by means of SMT. However, the integration of bilingual chunks into RBMT systems such as Apertium has its risks. There is always a danger of producing grammatically incorrect translations as a result of breaking structural transfer rules. The approach that may solve this problem offered in Sánchez-Martínez et al. (2009) is as follows: a) computing the

best translation of the source sentence given the collection of available bilingual chunks; b) usual translation of the source sentence by Apertium engine; c) applying language model to choose one of the possible translations for each of the detected bilingual chunks. Some improvements of the system's performance for English-Spanish language pair have been reported (Sánchez-Martínez et al., 2009).

1.3.3. Corpora for corpus-based MT systems.

As we have already mentioned in the preceding chapters, parallel corpora, also referred to as bitexts, are a core resource for corpus-based MT. As a rule, more bitexts lead to better performance of the system. A parallel corpus consists of bilingual texts aligned at the sentence level. In addition to MT tasks, these corpora are extensively used for other NLP applications such as automatic lexical acquisition, cross-language information retrieval, sense disambiguation, anaphora resolution etc. Alas, parallel corpora are a scarce resource, available only for specific language pairs and domains. The majority of available resources of parallel corpora (such as Europarl, EUR-Lex etc.) come from one domain (documents of multinational institutions such as the United Nations or the European Union, and multilingual countries such as Canada (English, French) or Hong Kong (English, Chinese)). This becomes a serious problem when SMT systems trained on such corpora are intended for general translations, since the language of Europarl, for example, is absolutely different from the one of daily life, and is inappropriate for use in other specific domains. Expanding the number and scope of parallel corpora via human translations would be an expensive and time-consuming option. Moreover, parallel corpora are difficult to obtain, as will be illustrated below, the acquisition of such corpora is a complex and expensive process which involves quite a few technical and legal issues. Some recent developments in the field focus on comparable non-parallel corpora to improve SMT performance and learning bilingual lexicons (e.g. Abdul-Rauf and Schwenk, Rapp et al., Munteanu and Marcu, Mikolov et al., Gamallo Otero etc.).

The classification of corpora is proposed by Wu and Fung (in Rapp et al., 2016):

- Parallel corpora: sentence-aligned corpora containing bilingual translations of the same document.
- Noisy parallel corpora: non-aligned sentences that are still bilingual translations of the same document.
- Comparable corpora: non-sentence-aligned, non-translated bilingual documents devoted to the same topic.
- Quasi-comparable corpora: non-aligned and non-translated bilingual documents that could deal with the same topic (in-topic) or with different topics (off-topic).

Having defined the dimensions we are interested in, we can compare any two corpora regardless of their size and format. These dimensions might be: language, dialect, domain, genre, content, author's sex, time, location of creation, purpose, text difficulty, text type (original, translation, summary etc.), vocabulary, modality (written, spoken). If the corpora coincide in almost all of these dimensions except of language, these are parallel corpora. The corpora should agree at least in the dimensions of domain, genre and modality to be considered comparable. The dimensions of language and content are of paramount importance for MT research.

Text collections used in a corpus research within the NLP community: Europarl (Koehn, 2005), Wikipedia[15], the International Corpus of English[16], the MLCC Corpus[17], WaCky[18] corpora. All of them differ in the degree of comparability mentioned above.

Europarl is a corpus of parallel texts in 21 languages dated back to 1996. It is based on the proceedings of the European Parliament available on the Internet. Originally SMT systems for 110 language pairs were trained on this corpus (in its first release in 2001 it covered 11 official languages of the EU). The size of the corpus is about 30-60 million words per language. Acquisition of a **parallel corpus** such as Europarl for the use in SMT is usually a five-step process (Koehn, 2005): 1) obtain raw data (crawling the web); 2) extract and map parallel chunks of text (document alignment); 3) break the text into sentences (sentence splitting); 4) prepare the corpus for SMT processing (tokenization, normalization); 4) map sentences in one languages to the sentences of the other language (sentence alignment).

This is how these steps were performed in the case of Europarl (Koehn, 2005):

1) Crawling the page of the European Parliament for the Proceedings in the form of HTML files. The URL for each file includes relevant identification information such as language, date, number of the discussion thread, number of the utterance. Crawling is a time-consuming task, given that the corpus consists of many small parts. Despite the poor speed of such crawls, it is generally easier than negotiating the transfer procedure directly with the technical staff of the web page. Copyright concerns normally pose another challenge for obtaining data, although to a lesser extent in the case of Europarl, which is based on the government sources. The European Parliament authorizes their reproduction as far as the source is indicated. In

15. https://corpus.byu.edu/wiki/
16. http://ice-corpora.net/ice/
17. http://catalog.elra.info/en-us/repository/browse/mlcc-multilingual-and-parallel-corpora/8bec17e2a9dc11e7a093ac9e1701ca0247ac6cf53a1d45f9b1d24f7ddc7bd444/
18. http://wacky.sslmit.unibo.it/doku.php?id=corpora

other cases, a much longer legal process may be expected to get a permission to access the data.

2) The next step is document alignment through identifying the topic of each text and matching these between languages. The extraction of relevant text pieces from noisy data is a challenging task. For Europarl a Perl program was employed to identify and extract the identity of speakers and their speech. Automatically learning systems to retrieve structured information from the Web have been an object of recent research.

3) Sentence splitting and tokenization also has its difficulties. When defining sentence borders, machine can never be sure whether a period *"."* acts as a sentence marker or an abbreviation. A possible solution is to create a list of known abbreviations followed by a period for each language, as it was done in Europarl. Another clue is a lowercased word after the period, which reveals that the period is not the end of the sentence, which, however, won't work in such cases as *"Mr."*, *"Dr."* etc. followed by proper nouns. Tokenization issues include merging of words such as English *"can't"*, *"wasn't"* (which should be transformed to *"can not"* and *"was not"* correspondingly).

4) Sentence alignment is easier in the case of Europarl, since the texts are usually available in the aligned format, each paragraph consisting of 2-5 sentences. The sentence alignment in Europarl is done with an algorithm that tries to match the sentences of similar length and merges the sentences if necessary (two short sentences in one language may correspond to one long sentence in another language). Since the paragraphs are short, the quality of alignment is quite high. A lot of work is being done on the improvement of sentence alignment algorithms.

5) Not only does the parallel corpus serve as a common training set, but also as a common test set to compare MT systems (a portion of the corpus). To be able to evaluate and to compare the system performance for different language pairs, a set of aligned sentences may be extracted for these languages.

Despite the success of SMT systems based on parallel corpora, the use of these corpora in MT has a number of disadvantages. The subtype of non-parallel corpora most promising for MT apart from parallel corpora are probably monolingual corpora covering roughly the same subject area in different languages without being exact translations of each other. An early example of such corpus may be the MLCC newspaper corpus. It embraces a number of financial editions for the period 1986-1994 such as *Financial Times* (English), *Le Monde* (French), *Handelsblatt* (German), *Expansion* (Spanish) etc. Although theoretically the publications of different newspaper do not depend on each other, their materials are, in

practice, related to the same world news/topics, provided by the news agencies.

1.4. Bilingual lexicon extraction based on non-parallel corpora.

The research based on comparable corpora has been ongoing for almost 20 years. The objectives of this research are mainly devoted to the following three aspects:

- development of MT systems based on comparable corpora;
- extraction of parallel segments from comparable corpora to provide training data for SMT systems;
- extraction of bilingual lexicon from comparable corpora and monolingual corpora.

In this paper we will concentrate on the extraction of parallel segments and bilingual lexica from comparable corpora. Many approaches have tried to automatically acquire translation equivalents from bilingual corpora. The approaches to automatic acquisition of translation equivalents from bilingual corpora can be organized in a continuum according to the type of bilingual input required for acquisition, ranging from aligned parallel, noisy parallel, comparable and unrelated non-parallel corpora (Gamallo Otero, 2007). We have already mentioned the problem of scarcity of parallel corpora. In this regard, mining parallel segments in comparable corpora may be a promising method. A pioneering research in this field is the work by Munteanu and Marcu (2002). They started from a small bilingual dictionary obtained from a parallel corpus and used bilingual suffix trees to extract parallel data from a comparable corpus. The suffix trees serve to compare strings of varying length making it possible to consider the full context of a word. For example, if we take a sequence of three words *abc* in the SL and *xyz* in the TL, in which according to the dictionary *a* is a translation of *x* and *c* is a translation of *z,* there is a high probability that *b* is a translation of *y* (Rapp et al., 2016). This assumption would be reinforced if *b* and *y* are encountered within other constructions in the middle positions. If the number of such examples is sufficient, the bilingual dictionary may be expanded by the word pair *b-y.* This approach is based on iterative expansions as the one explained above. One of its limitations is that it cannot be applied to languages with a significantly different word order.

The technique employed by Abdul-Rauf and Schwenk (2009) is quite similar to the previous one. The main difference is that they used proper SMT translations instead of a bilingual dictionary. They also used word error rate (WER) and translation error rate (TER) to define whether the extracted sentences are parallel or not. Starting from comparable corpora for English and French, French to English translation was performed with a SMT system. These translations were then used to retrieve data from the English corpus to be verified with

WER (measures the number of operations needed to translate the sentence including insertions, deletions and substitutions) and TER (considers the reordering of words and phrases in translation) metrics with a final objective to generate a parallel corpus for SMT uses. A zero WER would mean that two sentences are identical, and the sentences with lower WER score are, thus, most likely to be translations of each other. Nevertheless, two translations may differ in the word order, but still be correct. This issue is addressed by TER. The sentence extraction was based on the assumption that a news item that appears in the French corpus on the certain day, is likely to be reported in the English one within 5 days before or after this date. With the ID and date for each sentence of both corpora, they collected all sentences from SMT translations with the same date, and the corresponding articles from the English corpus, disregarding stop words such as articles. The retrieved sentences were filtered according to the certain criteria as well as WER and TER metrics to select candidate sentence pairs. Abdul-Rauf and Schwenk (2009) mention two types of common errors: 1) when the sentences share a lot of common words conveying different meanings, and 2) the sentences are parallel except of the sentence ends, where one sentence has less information than the other. The second issue was solved by detecting and deleting redundant information, showing a significant performance improvement.

Both approaches (Munteanu and Marcu, Abdul-Rauf and Schwenk) may be especially attractive for the language pairs with limited number of parallel corpora available. In addition to newspaper articles, other potential sources of comparable corpora include multilingual encyclopedias such as Wikipedia, domain specific comparable (potentially parallel) corpora such as translations of documents into national/regional languages (e.g. English in India, Galician in Galicia etc.), or the translations of academic publications and research papers.

The research at Google focuses on the method that would allow automatic extension and revision of lexicographic resources - dictionaries and phrase tables (a table of phrase pairs with associated scores which may come from probability distributions, in rough terms a type of a dictionary used by SMT systems). This method translates the missing entries by learning language structure from numerous monolingual data and mapping between languages from small bilingual data - a starting dictionary (Mikolov et al., 2013). This method is language-independent, that means it can be used for any language pair. The first step of the process is building a monolingual language model from a large monolingual corpus. Then a small bilingual dictionary is employed to learn a linear projection between the languages. As a result, any word from the monolingual corpus may be translated by projecting its vector representation (word with its context) from the SL to the TL. It serves to quantify and

categorize semantic similarities between linguistic items based on their distributional properties in large samples of language data. Once the vector in the TL is obtained, the most similar word vector is reproduced in translation. The language representations are learnt with the distributed Skip-gram or Continuous Bag-of-Words (CBOW) models. These systems learn the word representations and try to predict its neighbours with a learning rate of billions of words per hour. In the CBOW model, the training objective is to learn word vector representations of surrounding words to predict the word in the middle. Skip-gram model targets at learning word vector representations that are good to predict its context in the same sentence (Mikolov et al., 2013). The related words have similar vector representations, for instance, "*school"* and "*university"*, and tend to occur in similar contexts. It is also interesting that vectors can capture relations between words, e.g. *vector(Ukraine) - vector(Kiev)* is similar to *vector(Spain) - vector(Madrid).*

The use of distributional semantics is also explored in the method developed by Garcia et al. (2017) to learn bilingual word embeddings from lemmatized versions of both noisy parallel and comparable corpora that serve to extract equivalents of the elements of each collocation in the TL. Word-embeddings models capture distributional context of words in corpora and learn their bilingual representations, thus being able to predict words crosslinguistically. As the models learn the distribution of single words (lemmata), they deal with the semantic issues such as polysemy or homonymy. The perspective of this method is its ability to extract not only the equivalents with word-for-word translation across languages, but also cases when the collocation equivalent cannot be translated literally into the TL as for EN *"red vine"* - ES *"vino tinto"* (correct translation) vs. ES *"vino rojo"* (literal incorrect translation) (Garcia et al., 2017). The research has been focused on extracting the following syntactic patterns of collocations in three languages (Spanish, Portuguese and English):

Adjective - Noun: *serial killer;*

Noun - Noun (may include the preposition *"of"* in English or *"de"* in Spanish and Portuguese): *fit of rage;*

Verb - Object: *take care.*

Given monolingual collocations (ideally taken from the same resources) and a bilingual source-target model of word-embeddings, five most similar lemmata in the TL (*"trouble", "mess"* etc.) were obtained from the bilingual model for each SL collocation base (e.g. "*lío"*). Then collocations with the base equivalents were identified in the target list, the cosine distance between collocates was computed to select those whose similarity was higher than the empirically defined threshold. If those also were among the most similar words, the source

and target collocations were aligned. (Garcia et al, 2017). As mentioned above, this strategy was applied both to noisy parallel and comparable corpora. The performance of the method was compared to the same strategy that employed bilingual dictionary instead of the word-embedding models. The results revealed that the number of the extracted bilingual equivalents was much higher with word embeddings, which may be connected to the size constraints of the dictionaries and to the fact that collocates are not always direct translations of each other (as in the example above). However, the dictionaries normally provided better accuracy. The results with parallel corpora were almost the same as those with comparable corpora, which proved that the method was efficient with non-parallel corpora as well. During the experiments some challenges have come up. It is worth mentioning that the strategy of Garcia et al. (2017) works with lemmata, and not with tokens. As lemmatization is different in different languages, it may lead to incorrect translations in such cases as: EN *"lovely daughter"* is translated as ES "*hijo encantador"* (incorrect translation, male citation form or lemma) instead of ES *"hija encantadora"* (correct translation, female form).

The strategy proposed by Gamallo Otero (2007) aims at extraction of translation equivalents from comparable corpora without the use of external bilingual resources. Instead it made use of bilingual correspondences between lexico-syntactic templates extracted from small parallel texts as seed words. The reason for that is that even a rather small specialized parallel corpus contains more accurate information for extraction as compared to a general-purpose dictionary, as has been proven in the process of experiments. Furthermore, not all the words in the dictionary are reliable seed expressions. For example, polysemous words should be excluded from the list, since they add semantic noise. On the other hand, lexico-semantic templates represent unambiguous local contexts of words, which makes them discriminative seed expressions to extract translations from comparable data. Basic pattern matching techniques were used on PoS-tagged parallel and comparable corpora to identify potential binary dependencies. From each binary dependency, two lexico-syntactic templates were selected, e.g. from the following binary dependency:

modA (legal, document)

the lexico-syntactic templates are derived:

<legal [NOUN]>, <[ADJ] document>

which represents: 1) a set of nouns that can occur after *"legal",* for example: *"action", "advice", "dispute", "liability", "paper"* and the like; 2) a set of adjectives occurring before *"document"* such as *"legal", "official", "electronic", "historical"* etc. On the basis of the identified lexico-syntactic templates, bilingual correspondences were extracted from small

aligned parallel corpora by computing the similarity between template pairs with Dice coefficient, considering their co-occurrence in each aligned segment. Sparse (rare) and unbalanced bilingual templates (when one pair is very frequent, and the other one is very rare) were excluded from the seed templates. Filtered seed templates served to extract translation candidates by ranking them and with the help of the context-based algorithm. The main condition of the standard ranking was that lemma *L1* should occur in the same seed templates as lemma *L2* to be its translation. Thus, to compute the similarity between the lemma and its possible translation, the seed templates they share and do not share were compared. According to the context-based algorithm, a SL lemma *L1* in the seed template *T1* is compared to all the TL lemmata that occur within *T1*. The objective of this strategy is to capture translations in context-sensitive cases. The potential of this approach is that with many small parallel texts, it could be possible to generate big sets of seed templates taken as bilingual anchors to easily pseudo-align large amounts of comparable texts, thus relying on easily available non-parallel corpora to train the translations extractor (Gamallo Otero, 2007).

Finally, we would like to mention a research that is rather similar to the approach we describe in the practical part of this thesis. Grefenstette (1999) used electronic bilingual dictionaries to create a gold standard of compositional compounds for the evaluation of World Wide Web as a resource for EBMT tasks. The criteria for inclusion of the entries to the gold standard were as follows:

1) Compound. The entry was decomposable into two other words found in the dictionary.
2) Compositionality. The compound term was translated into the TL by two word phrases.
3) Transparency. The translation of the compound form could be deduced from the dictionary translations of its smaller components;
4) Ambiguity. There existed more than one possible translation of each translation candidate.

For the research purposes, the dictionary translation was considered a preferred translation in all cases. Although, the compounds were to be found in the TL dictionaries as entries, it was assumed that they were to be generated by the system from the dictionary translations of their components, as if the compound entries themselves were not in the dictionary. This strategy aimed at modeling a situation, when the human users encounter a term they are not familiar with. In each case, all the possible two word translations using the decomposed SL word and recombining the TL translations of these subparts using the same dictionary were created. The resulting translation candidates were then sent to AltaVista as phrasal queries, and the frequency of occurrence of the phrase was noted. The choices of the WWW were compared to the dictionary translations ignored at the first stage of the experiments. The

results were quite promising, showing that 86-87% of those choices were correct (Grefenstette, 1999).

1.4.1. Extraction of terminology from non-parallel corpora.

Terminologists have been extracting terminological knowledge from specialized texts to create termbanks since late 1980s. Corpus-based terminology has contributed to the systems of information retrieval and machine or computer-assisted translation. A lot of research has been dedicated to the task of automatic extraction of terminology during the last 10-20 years. Since no parallel corpora are available for most of specialized domains, especially for emerging domains (such as renewable energy), most projects have been exploiting the potential of monolingual and comparable corpora widely available on the Web. Gurrutxaga et al. (2013) worked on the automatic collection of corpora and extraction of terms from those comparable corpora for specialized dictionary making for English and a less-resourced language – Basque. Frantzi et al. (2000) focused on multiword terms, automatic recognition and extraction thereof from comparable corpora. A prominent EU project in the field of terminology TTC ("Terminology Extraction, Translation Tools and Comparable Corpora") aims at improving MT, CAT and terminology management tools by automatically generating bilingual terminologies in multiple languages: English, German, French, Latvian, Spanish, Chinese and Russian (Blancafort et al., 2010). Many researches argue that the context of a term provides the information about its meaning and usage. According to L'Homme et al. (in Faber et al., 2005), the context helps retrieve related terms, defining elements, relations of synonymy and explicit relations between syntactically related concepts. Furthermore, the context makes it possible to verify conceptual information and the grammar extracted from the reference works, to specify abbreviations and to add encyclopedic information not included into the definition.

With respect to terminology extraction from a corpus, two strategies are applied: linguistic and statistical. The former assumes that terms correspond to specific (morpho-) syntactic patterns. Therefore, its objective is to identify and extract the sequences of words whose structure matches one of these patterns. The latter relies on the idea that the statistical features of terms differ from those of general vocabulary. The majority of statistical techniques focus on extraction of multiword terms (MWT), mainly by calculating association measures (serve to quantify the strength and the direction of the relationship between two data sets). Those include: Mutual Information (Church and Hanks, 1989), Log-Likelihood Ratio (Dunning, 1994), Chi-square (is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories) etc.

In order to determine the context of a word, distance-based and/or syntax-based windows may be considered. It is also possible to represent a word context by using statistical language models, i.e. a probability distribution over a sequence of words. In order to compute the similarity, contexts should be put into one space by translating one of them with MT, parallel corpora or a dictionary-based approach. The challenges of the latter are: 1) lexical ambiguity of words (the simplest method would be to use the first meaning of the entry as it is normally the most frequent one, which completely disregards the domain in which the word is used. Normally different statistical approaches based on the degree of cohesion or association between the translation candidates are employed. As a rule, a TL corpus is used to compute the association scores); 2) words that are not in the dictionary (here the cognate detection method might be useful, especially for the domain of science, where many word share the same origin).

Gurrutxaga et al. (2013) applied a hybrid approach for monolingual candidate term extraction in both English and Basque, i.e. looking for some linguistic patterns, then processing them statistically. For instance, for Basque they used Eustagger[19] for lemmatization and PoS tagging, and a grammar to identify the terms corresponding to most typical noun phrase structures of Basque terms (they project dealt only with noun phrases as terminological units). Afterwards, they measured the termhood of the extracted terms with Log-Likelihood Ratio (LLR) for one-word terms calculating the domain relevance of the terms with respect to an open-domain corpus, and unithood (degree to which a sequence of words is able to form a stable lexical unit) also measured with LLR as the association measure. The ones with the highest measures were taken as the final term candidates. With regard to context modeling, only content words such as adjectives, nouns and verbs were considered as contexts, adverbs were found to produce much noise. A distance-based window was established to delimit the contexts of terms. The window size for the Basque language was bigger than for English due to agglutinative nature of the former. The experiments showed that it was also useful to use punctuation marks to delimit the context window. To translate the contexts of a Basque word to make them comparable to the English contexts, a bilingual general dictionary was used. However, the results were not very good due to a large number of words not found in the dictionary. To improve the results the researchers tried to find equivalents not included into the dictionary by means of cognate detection. The context similarity and cognate detection techniques were applied for the extraction of translations for

19. http://ixa2.si.ehu.es/eustagger/

the source term candidates.

The method of Frantzi et al. (2000) for automatic extraction of MWT, as the previous one, combines linguistic and statistical information. It consists of two parts: 1) the *C-value* for the improved extraction of nested MWT (those that appear within other longer terms, and may or may not occur in the corpus by themselves) and collocations; 2) the *NC-value* that adds the context information to the *C-value* for improved MWT extraction in general. The *C-value* puts an emphasis on statistical information defining the termhood of MWE. The linguistic part embraces PoS tagging of the corpus, the linguistic filter constraining the type of terms extracted, and the stop-list. The linguistic filters serve to avoid the extraction of irrelevant strings such as "*of the", "is a"* etc. Instead only the most common syntactic patterns of terms were considered: *N - ADJ, N - N, N - Prep - N* etc. There are several types of filters that may be applied. A "closed" filter permits only certain strings, for instance, the filter *Noun+* is limited to the sequences of nouns and provides more precise output since noun sequences in the corpus are most likely to be terms. However, it overlooks the terms that consist of adjectives and nouns, and the output is not complete. On the contrary, an "open" filter for noun phrases such as *((Adj|Noun)+|((Adj|Noun)∗, (NounPrep)?)(Adj|Noun)∗)Noun* has an opposite effect. It will extract more terms than the previous one, for it considers both adjectives and prepositions, but the output will contain more MWE that are not terms. The choice of the filter depends on the research objectives. Frantzi et al. (2000) used both "closed" and "open" filters. The stop-list included the words that were not expected to be terms of the domain. For this method, the stop-list consisted of 229 function and general vocabulary words that had high corpus frequency. The statistical measure C-value assigns termhood score to candidate terms from the linguistic part using the statistical data such as corpus frequency (terms tend to be more frequent in the specialized corpus). However, raw frequency doesn't always work for nested terms, whose parts are not terms themselves. For instance, in the example "*soft contact lens"*, the substring *"soft contact"* is not a term (Frantzi et al., 2000). On the one hand, the simplest solution would be to compare the frequency of occurrence of the substring within the longer term and its independent occurrence in the corpus. On the other hand, there is always a risk to leave out the terms with low corpus frequency. The *C-value* approach suggests considering the number of times the substrings appear within the longer terms. The higher this number is, the more chances the substring is a term. The *NC-value* is an extension of a *C-value* that uses context information for the term extraction. Thus, adjectives, nouns and verbs frequently occurring with terms were used to distinguish between terms and non-terms. For this purpose, the ranked list of important context words was created,

the criterion for the word to appear on this list was the number of terms it appears with (Frantzi et al., 2000). The assumption is: the higher is the number, the higher is the possibility to appear with other terms in the domain. The list of term candidates ranked with the *C-value* should be re-ranked taking into account the context information so that real terms appear even closer to the top of the list. The given method showed good performance results on a well-structured medical corpus.

We have already mentioned the EU project TTC aimed at using the potential of comparable corpora for extraction of monolingual terminologies and bilingual alignment of the extracted terms from the domain corpora in different languages. When it comes to multilingual terminology, the same monolingual term extraction program and method should be applied to each language. It is also important to identify morphological features of terms with the tokenization tool. Here the difference between languages may be an issue. For instance in German, single word terms are more similar in their structure and meaning to multiword terms. The translation of MWT is of paramount importance in terminology, as the latter constitute up to 80% of domain-specific terms. Nevertheless, for the German language morphological compounds tend to be much more frequent than MWT as such (Daille, 2012). The MWT that are extracted correspond to specific syntactic patterns. For all languages, the most common SWT pattern is *ADJ* or *N*. For Spanish, the main MWT patterns are *N - N, N - Prep - N, N - ADJ*. The bilingual terminology alignment was performed with a bilingual dictionary and a method similar to the one used in Gurrutxaga et al. (2013). The major difference to the latter is the compositional translation approach for MWT, up to 48,7 % for English and Japanese *N - N* compound terms are compositional by nature (in Daille, 2012). The compositional approach at the word level implies translating each component of an MWT individually using a bilingual dictionary, and then putting together the translated components. TTC TermSuite was designed to perform bilingual term extraction in the project languages. It has a three-step functional architecture which is based on required inputs and the resulting outputs of each tool. Thus, the result of text-processing (spotter: performs PoS-tagging, stemming and lemmatization) becomes an input for monolingual term extractor (indexer: recognizes and indexes the terms, computes their relative frequency and domain specificity, filtering of candidates etc.), which in its turn provides the data for bilingual term alignment (aligner: uses either the context-based or the compositional translation approach). The aligned bilingual terminologies are to be used in both computer-assisted and machine translation.

Chapter 2. Building a bilingual lexicon from a monolingual corpus.

In the practice-oriented second chapter we suggest a workable approach that aims at automatic generation of the TL translations from a monolingual corpus of the TL texts. These translations may be further used to build new lexicographic resources – bilingual lexica, terminological databases and translation memories or/and to expand the existing ones. These, in their turn, might serve to improve the quality of MT systems. Due to the fact that our research is not based on parallel corpora, this approach may come in handy for the domains and languages where parallel corpora are not easily available or scarce.

This method roots in the experiments of Grefenstette (1999). We also aim at translating multiword terms by looking up their components in a bilingual dictionary and generating all possible translation candidates by combining the suitable translations of smaller parts of the MWT. The principal difference lies in the usage of a monolingual corpus and not the World Wide Web to compute the frequencies of translation candidates.

Given a corpus in Spanish (as a TL), English LSP glossaries (as a SL), and a bilingual dictionary (English-Spanish), our strategy aims at mining the corpus to extract Spanish translations of the English terminological entries. The translation strategy consists of two main tasks:

1) Generation of translation candidates: for each multiword term, TL candidates are generated making use of a bilingual dictionary and a set of basic syntactic patterns (Table 1):

MWT	**Syntactic pattern**	**Dictionary translations for each MWT component**	**Possible translation candidates**
sexual harassment	ADJ - N	*harassment - acoso, hostigamiento;* *sexual - sexual*	*hostigamiento sexual, acoso sexual*

Table 1. Generation of translation candidates.

2) Corpus-based selection: given the set of candidates generated at the previous stage, the system selects the candidates that occur in the Spanish monolingual corpus and ranks them by frequency:

"***178*** *hostigamiento sexual*", "***58*** *acoso sexual*"

where ***178*** and ***58*** is the number of times, the MWT occurs in the corpus, i.e. corpus frequency. In order to make it possible to search the corpus, it is transformed into a word-context model.

Our approach is based on an extraction tool elaborated by the natural language processing research group ProLNat@GE at the University of Santiago de Compostela within the frames of the EXTRA-LEX project[20]. The aims of the project were: a) automatic extraction of bilingual lexica for the Galician-Spanish language pair with the focus on multiword translation equivalents and non-parallel, comparable corpora;
b) actualization of lexicographic resources for machine translation engines, especially rule-based ones.

In the current research the extraction tool, which was implemented in Perl, has been adapted to be applied for English (as a SL) and Spanish (as a TL). The adaptation includes adjustment of PoS tagger and syntactic analyzer to the English language and tailoring transfer rules for the English-Spanish language pair. We have performed evaluation of the system with the *precision and recall metrics*. The output of the experiment is an updated English-Spanish legal terminological database.

2.1. Lexicographic resources.

Our research is based on the following lexicographic resources: 1) Spanish corpus of legal texts compiled by us for the purposes of the current research; 2) bilingual English - Spanish dictionary, 3) reference legal dictionaries.

The extraction tool integrates Collins[21] bilingual dictionary to carry out the lookup procedure of terms we would like to translate. The lookup results are used to generate the translation candidates. Additionally, we consult a number of LSP dictionaries as reference sources. The latter fulfill several functions. Firstly, they provide for seed terms to build a specialized corpus. Secondly, they are a source for the SL terms to be translated by the system. Finally, they help us check the system output and compare the dictionary translations to the ones produced automatically. The reference lexicographic resources are described in more detail in the next chapter devoted to evaluation.

The specialized corpus was compiled with the help of WebBootCaT tool provided by a corpus manager and text analysis software Sketch Engine[22]. In short, that is how WebBootCaT works (Baroni et al., 2006):

- 20. "EXTRA-LEX: Extracción automática de léxicos bilingües Galego-Español e actualización dos recursos lexicográficos de motores de tradución automática". (Ref: PGIDIT07PXIB20401PR). Consellaría de Economía e Industria de la Xunta de Galicia. Inicio: 01/01/2007. Final: 31/10/2010. https://gramatica.usc.es/pln/projects/extralex.html
21. https://www.collinsdictionary.com/dictionary/english-spanish
22. https://www.sketchengine.eu/

- the user selects the language and inputs so-called seed terms, i.e. terms that are expected to be representative of the domain (up to 20, in our case legal terms) into the system These may be either SWT or MWT in double quotes;
- the system then crawls the web for pages with seed terms. The default settings are that ten queries are sent to Google, each containing a randomly-selected triple of the seed terms;
- each Google query returns up to 100 hits, by default the top ten ones are selected, the duplicate web-pages are disregarded. Very long and very short pages are normally filtered out as well, as it is widely accepted that those do not contain useful samples of text;
- the output is further processed by the system to filter out unwanted information such as HTML and javascript; the text is tokenized; the corpus is lemmatized and PoS-tagged (for the languages where lemmatizer and PoS-tagger are available, which is the case of our Spanish corpus, i.e. FreeLing tagset[23] is used) and ready for use.

The metadata about the corpus "**derecho_ES**" built with WebBootCaT are given in the Figure 1:

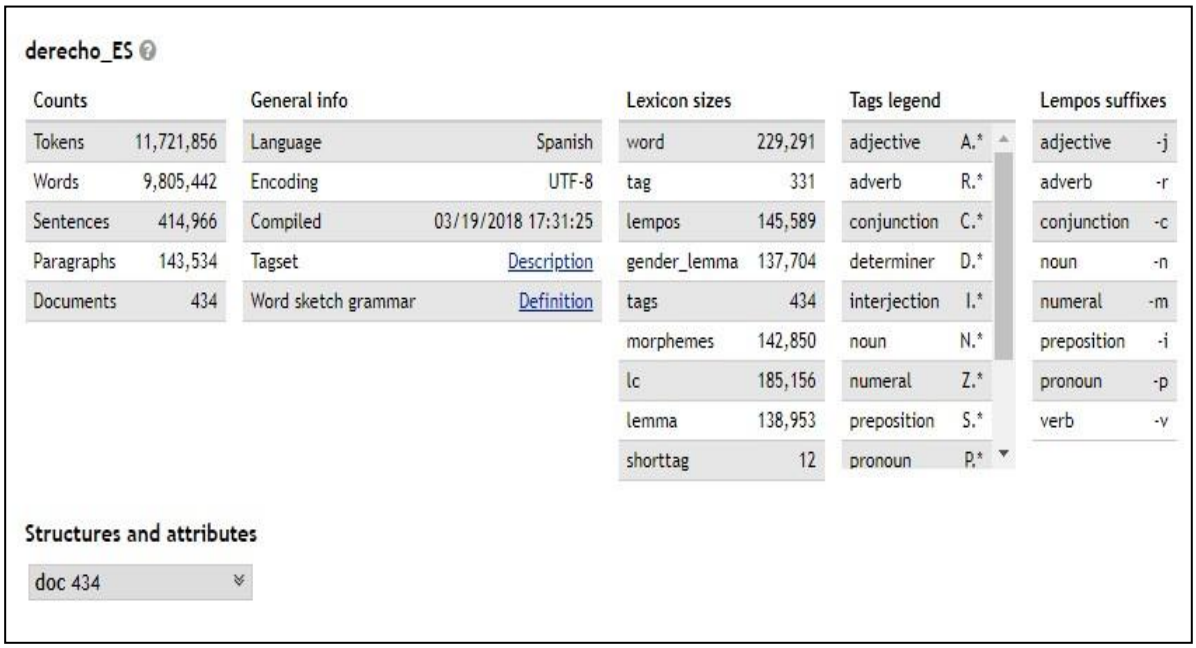

Figure 1. Corpus metadata

Thus, with the seed terms taken from a reference dictionary we have built a specialized corpus of Spanish legal texts of about 10 mln. words based on some 434 documents. Our hypothesis is that this size is sufficient for this type of research as far as it doesn't contain too much noise and still provides enough sublanguage data to remain representative.

23. https://talp-upc.gitbooks.io/freeling-4-0-user-manual/content/tagsets/tagset-es.html

2.2. System description.

As we have already mentioned, the translations are generated automatically by the term extractor based on Perl pipelines, where commands written for the output of one operation become an input for the following operation automatically. Linguakit toolkit[24] for NLP dependency parser and PoS-tagger also developed by ProLNat@GE group are applied to the corpus to generate the desired output. The system architecture (see Figure 2) includes two processes: language model construction and translation. The former takes the SL corpus as an input and performs a complex syntactic analysis that includes the following steps: sentence segmentation, tokenization, splitting, lemmatization, PoS tagging and dependency analysis. The next stage — vectorization generates a basic model of frequencies representing each word and its syntactic contexts; to be precise, the corpus is modeled as a "word-context-frequency" matrix.

The translation module (Translation 1 in Figure 2) takes as input a TL compound term (MWT we want to translate). The translation candidates generator uses the parser and a bilingual English-Spanish dictionary to generate translation candidates for the input term. The term selector performs the process of translation proper, i.e. with the language model selects the candidates occurring in the corpus and produces the ranking of translation candidates (Translation 2 in Figure 2). The final output offers the TL terms ordered by frequency.

24. https://github.com/citiususc/Linguakit

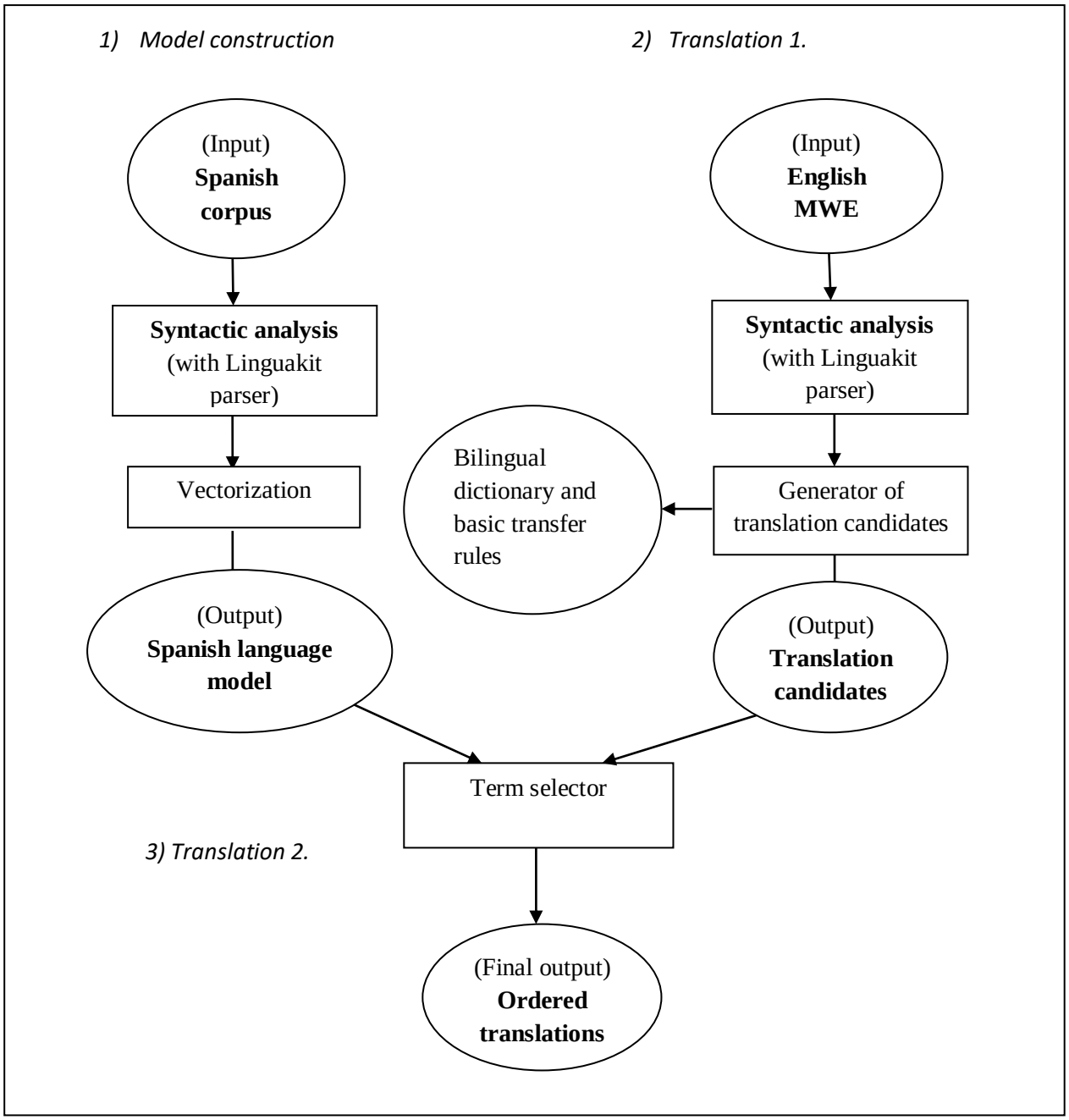

Figure 2. System architecture

The system handles the following English-Spanish transfer rules:

[Nsubj - V] → [Nsubj - V]

[V - N_{CDIR}] → [V - N_{CDIR}]

[ADJ - N] → [ADJ -N] or [N - ADJ]

[N_1 - N_2] → [N_2 - Prep de - N_1] or [N_2 - N_1]

[V - Prep - N] → [V - Prep - N]

[N - Prep - N] → [N - Prep - N]

As we see, these transfer rules guide how the MWT of the certain syntactic patterns in the SL may be transformed into the TL at the structural level. These may coincide in terms of order and syntax in both languages as in the case of MWT consisting of a noun (as subject or as a direct object) and a verb (*Nsubj - V, V - N_{CDIR}*); of a noun or a verb with a preposition and a noun (*N - Prep - N* and *V - Prep - N* correspondingly). On the other hand, some SL patterns may correspond to several syntactic patterns in the TL, for example: English adjective and noun (*ADJ - N*) multiword expressions may keep their order in Spanish, but the Spanish multiword terms with a noun preceding an adjective are far more common (*N - ADJ*). Thus, there exist two translation options for this syntactic pattern. Similarly, MWT that contain two nouns (N_1 - N_2) correspond to the Spanish N_2 - *Prep de* - N_1 or N_2 - N_1 constructions, in both cases the noun that occurs first in English, will be placed second in Spanish. Therefore, the transfer rules are designed to anticipate such syntactic changes and to provide for the coherent output in the TL.

In the same way there are different variants for translation of some syntactic patterns, some prepositions within the patterns that include a preposition may also have several equivalent translations in the TL. For instance, the Spanish preposition "*de*" may be translated into English as *"from", "of", "in"* or *"with"*. The system includes a special function to deal with the prepositions.

As we have already mentioned above, the system is capable of generating translation candidates from the bilingual dictionary and the aforementioned syntactic transfer rules. It looks up each component of the MWT in the bilingual dictionary, producing all potential translation variants, including translations for different parts of speech (e.g. "*search*" as a verb and a noun, *"criminal"* as an adjective and a noun). The PoS tagger has already defined the syntactic pattern of the English term. Consequently, the generator of translation candidates will not use the translations that do not match the corresponding PoS tags. We would like to analyze the following example for the input term "*criminal act*" to understand how the lookup process works:

N:#act# -- acto
-->V:#act# -- actuar
-->V:#act# -- agringarse
-->V:#act# -- fungir
N:#act# -- ley
-->V:#act# -- obrar
-->V:#act# -- representar

-->V:#act# -- seguir

A:#criminal# -- criminal

A:#criminal# -- delictivo

N:#criminal# -- delincuente

N:#criminal# -- facineroso

N:#criminal# -- malhechor

The system looks up all translations for both *"criminal"* as an adjective and a noun and *"act"* as a verb and a noun. The generation of translation candidates, however, is limited to the syntactic pattern defined via syntactic analysis, i.e. **ADJ - N**. These are the relevant translations:

ADJ: *criminal, delictivo* - N: *acto, ley*

In the list of syntactic transfer rules processed by the system we can see that the given syntactic pattern has two possibilities in Spanish: ADJ - N and N - ADJ. As a result, the system produces 8 translation candidates: *acto delictivo, delictivo acto, acto criminal, criminal acto, criminal ley, ley criminal, ley delictivo, delictivo ley.* Nevertheless, only the translation candidates found in the corpus will appear in the final output:

23 Rmod_down_acto criminal

41 Rmod_down_acto delictivo

4 Rmod_down_ley criminal

1 Lmod_down_acto delictivo

1 Lmod_down_ley criminal

The codification of the results is as follows: "**Rmod_down**" or "**Lmod_down**". This means that modifier should appear on the right or on the left to the noun it modifies, correspondingly. The marking "**down**" merely indicates that an adjective appears as a context of a noun as its head. The marking "**up**" would mean the opposite, i.e. that noun appears in the context of its modifier. Thus, "**Rmod**" means that the term is "*acto delictivo*". On the other hand, "**Lmod**" implies that the same term should be read as "*delictivo acto*", which in this case will not be a correct translation. This result may be a false positive or an error produced by the system. However, such cases are marginal and are not considered due to their low corpus frequency (in this case only one occurrence). We check if the translations are correct with the help of lexicographic resources: main and secondary reference dictionaries, or Web as corpus (for translations not found in the dictionaries), and evaluate the results with the help of the proposed methodology. Tendencies are detected in the types of translation errors the system produces to be analyzed in the "Problems and limitations" section.

Chapter 3. Evaluation.

We have taken the *Superior Court of California glossary (2006)* and *Law Firm Douglas S. Smith dictionary*[25] as a basis for this experiment and the source for the terms subject to evaluation. The aim is to check whether and how the existing resources may be enriched and/or improved with the help of our tool.

We have deliberately chosen institutional non-exhaustive dictionaries to illustrate how, whether and to which extent an existing glossary may be expanded with the help of the system of automatic term extraction. Further refinement of the corpus creation process would be extracting keywords/terms from a comparable in terms of size and domain English corpus to make the seed terms selection more objective. A number of other legal dictionaries, glossaries and termbases as well as Web as corpus technique were employed to verify translations automatically generated by the system. We included smaller dictionaries of legal institutions (such as *Douglas S. Smith dictionary, the glossary of the State of Connecticut Judicial Branch Superior Court Operations Division*[26]*, Superior Court of California glossary* etc.) into the list of the reference works for this research, for we argue that those tend to be more up-to-date than the paper ones. Moreover, for the selection of terms to be translated by the system we do not really need a very large dictionary, but rather a more compact one treating general law terms of high frequency, which are normally the ones included into the glossaries compiled by and for practicing lawyers or legal translators and interpreters. However, we are using some paper dictionaries to check the generated translations: bilingual dictionaries with encyclopedic features - *Diccionarios de Términos Jurídicos inglés-español, español-inglés* (Bossini, Gleeson 1997; Alcaraz Varó, Hughes 2007), *Diccionario ESPASA Términos jurídicos* (2002) and a bilingual dictionary *Essential Legal dictionary English-Spanish and Spanish-English* (Kaplan, 2008). Due to the space limitations of paper dictionaries, we use online resources such as *WordReference*[27] and *Linguee*[28], which provides access to large amounts of bilingual sentence pairs found online. We also consult *Inter-Active Terminology for Europe (IATE)*[29] - the terminology database of the European Union.

25. https://www.dcsmithpllc.com/Spanish-To-English-Dictionary.shtml
26. https://www.jud.ct.gov/external/news/jobs/interpreter/Glossary_of_Legal_Terminology_English-to-Spanish.pdf
27. http://www.wordreference.com/
28. https://www.linguee.es/
29. http://iate.europa.eu/SearchByQuery.do

3.1. Methodology.

The **precision and recall metrics**[30] were employed to evaluate the final output. Precision and recall have been used regularly to measure the performance of information retrieval and information extraction systems (that is basically what our system does: extracting terms from the corpus). Precision or positive predictive value is the fraction of correct translations among all system translations, while recall or sensitivity is the fraction of correct translations that have been generated by the system over the total amount of correct translations (gold standard). Thus, if the system generates four translations for a term that has, let's say, three translations in the gold standard, and two of them are correct, its precision will be 2/4 and its recall — 2/3. In this respect, precision can be seen as a measure of exactness or quality, whereas recall is a measure of completeness or quantity. A system with high recall but low precision generates many translations, but most of them are incorrect. On the contrary, a system with low recall and high precision produces few translations, but most of them are correct. An ideal system with high precision and high recall will generate many translations with none of them labeled as false.

As a rule, precision and recall scores are not discussed in isolation, but combined into a single measure. An example of measures that are a combination of precision and recall is the F1-score (or f-measure), which is calculated with the help of the following formula: $F1 = 2\frac{P \times R}{P+R}$. , where *P* stands for *precision* and *R* for *recall*.

In general, the f-measure is the balanced harmonic mean of recall and precision, giving both metrics equal weight. The higher it is, the better is the system.

We considered 50 English legal terms from the abovementioned glossaries corresponding to the patterns: **N_1 - N_2, ADJ - N, N - Prep - N and V - N_{CDIR}** for the evaluation of the system performance. We have created a table (see Annex 1) with the following information:

1) Input terms in the SL;
2) translation(s) offered by the reference dictionary;
3) gold standard (Spanish translations taken from the reference dictionary, and automatically generated translations labeled by us as correct);
4) system output (all translations generated by the system);
5) the most frequent translations produced by the system;

30. https://en.wikipedia.org/wiki/Precision_and_recall

6) whether the most frequent translation(s) of the system is (are) correct or not.

In the course of research we have come across a significant limitation. Out of the 300 English MWT we evaluated, the tool was able to generate translations only for 50, which reveals a low recall of the system. The failure of the tool to produce translations is connected to insufficient coverage by the lexicographic resources used by the system (for more details see the next section), which led to poor recall score. We have not included the untranslated terms into a qualitative table, i.e. evaluation data set. However, these terms have been considered in the final evaluation and represented in the quantitative table. In order to calculate the system recall we assume that for each of the 250 untranslated terms there is one translation in the gold standard (only a dictionary translation), which is added to the gold standard of translated terms (dictionary translation + correct system translations). The sample of the evaluation data set may be found in the Table 2 below:

English term	**Dictionary translation(s)**	**Gold standard**	**System translation(s)**	**System (most frequent)**	**Most frequent correct**
blood test	examen de sangre	examen de sangre, prueba de sangre	prueba de sangre	prueba de sangre	1
breach of peace	alteración del orden público	alteración del orden público, ruptura de paz	ruptura de paz	ruptura de paz	1
case law	precedentes	precedentes	derecho de caso, ley de pleito, ley caso, ley pleito	derecho de caso, ley de pleito, ley caso, ley pleito	0
cause of action	acción del litigio	acción del litigio, causa de acción, causa de acto	causa de acción, causa de acto, causante de acción, desencadenante de acción	causa de acción	1
chain of custody	cadena de custodia	cadena de custodia	cadena de custodia	cadena de custodia	1

Table 2. Sample of the evaluation data set

We would like to give the following important clarifications concerning our evaluation decisions:

1) In some cases the column “most frequent” includes several translations. That means

that the system either assigned the same frequency to those translations or there is a large gap between the translations with high frequency (and insignificant difference in frequency between them) and other marginal translations.

2) One of the biggest limitations of the system is that it translates lemmata and not word forms (e.g. *"ley delictivo", "exclusion of witnesses - exclusión de testigo", "leading question - dirigir pregunta"*). We have decided to evaluate as correct the cases where the agreement in gender between the noun and adjective is not kept and the cases where plural is translated as singular (including cases, where plural in the TL is more typical than the singular for the terms that are used in singular in the SL), since these errors do not lead to content distortion. On the contrary, the case of the third example and similar instances, i.e. translation of gerund or participle as infinitive, have been marked as incorrect and leading to misinterpretation of the term. The limitations of the system are discussed in more detail in the section "Problems and limitations".

Evaluation results with the precision and recall metrics are presented below in the Table 3:

System positives		**Un-translated**	**Total correct**	**Precision**		**Recall**		**F1-score**
All	Most frequent	250	342	50/59	84.75 %	63/342	18.42 %	30.26 %
63	50							

Table 3. Evaluation results

3.2. Problems and limitations.

In this research we have encountered various difficulties that are mostly triggered by four factors that turn out to be interconnected in the majority of cases: 1) inherent challenges of MT as such (e.g. too literal translations), 2) problems of terminology (e.g. synonymy), especially legal terminology (e.g. different legal systems in different countries); 3) quality of the lexicographic resources employed by the system (e.g. general dictionary and rather small specialized corpus), 4) limitations of the system itself (e.g. works with lemmata, and not with word forms resulting in ungrammatical translations).

We have taken some examples from the evaluation table to illustrate these difficulties and, where possible, suggest strategies for improvement. It is to be mentioned, that the solutions are suggested from the point of view of linguistics and translation, the technical side is not the main focus of the current research and has been touched upon only briefly.

The principal challenge of MT is a word-for-word translation. Indeed, in terminological field this drawback may not be that evident as in other text types (literature etc.), as cases of polysemy, idiomatic meanings and other semantically confusing factors are not that frequent. However, the translations produced by the system still may sound too literal in cases when the officially accepted equivalent TL terms differ significantly in their syntactic and semantic structure, when an additional explanation or a descriptive translation of the term is required (for instance, for the terms peculiar to the SL culture). A good example of a word-for-word incorrect translation would be "*derecho de caso, ley de pleito*" for the English term *"case law*". The reference glossary suggests "*precedentes*", in IATE we have also found "*jurisprudencia*" and *"doctrina legal",* in Linguee - "*jurisprudencia"* and "*derecho jurisprudencial*" (the latter marked as less frequent). First of all, the problem of "*precedents"* and "*jurisprudencia"* is that the system is not capable of working with terms other than MWE, which, however, belongs to the limitations of the system as such (see below). On the other hand, automatic translations still tend to be inaccurate in cases of compound terms with non-transparent meaning, i.e. the translation cannot be generated only from the meanings of their components. Thus, the component *case* in the MWE "*case law*" cannot be merely substituted by its direct dictionary equivalent *caso*. Furthermore, we cannot expect from the system to generate "*igual protección ante la ley"* as a translation for the MWE *"equal protection",* for the component *"ante la ley" (EN "under the law")* is absent in the SL input term. The best solution in this case, would be revision (post-editing) performed by a human translator and/or terminologist, who could decide whether such additions or explanations are necessary in each particular case. The resulting revised translation memory or example base of high quality may, consequently, reduce human participation in MT, when used by MT engines for further translation tasks. If we look for purely automatic solutions at the stage of terminology extraction and lexicon building stage, we may still try improving the lexicographic resources which provide the basis for the system, i.e. including corpora of higher specialization, using more advanced corpus-building tools and methods, substituting the system dictionary which is a general one for an LSP dictionary that will treat field-specific translations for each word (e.g. "*case*" - "*jurisprudencial" etc.*).

Quite logically, the system performs best with the MWT with a transparent meaning of both components that follow the same syntactic pattern in both languages, where even word-for-word translations would be correct, for example: *"domestic violence - violencia doméstica", "constitutional right - derecho constitucional", "hostile witness - testigo hostil"* and the like. We would suggest the method of word embeddings proposed by Garcia et al.

(2017) for the MWT which cannot be translated by merely looking up dictionary equivalents of their components.

The terminological field, however attractive it may seem for MT tasks, is not devoid of ambiguity. Thus, IATE offers a number of synonyms, and LSP dictionaries offer different translations for the same term. In the case of print LSP dictionaries, which due to the space limitations cannot treat all the synonyms, the criteria for inclusion or exclusion of certain equivalent terms remain blurred. Therefore, each reference dictionary we have consulted to check the translations suggests different Spanish translations for "*criminal record*" - *"ficha delictiva", "antecedentes penales"* (Varo, Hughes, 2007)*; "antecesdenetes penales", "antecedentes criminales", "historia criminal"* (Kaplan, 2008), "*antecedents"*, *"antecedentes penales"* (Linguee) and "*direct evidence*" - *"testimonio de primera mano"* (Espasa, 2002), *"prueba directa"* (Kaplan, 2008), *"prueba directa", "evidencia directa"* (Bossini, Gleeson, 1997), *"testimonial directa", "interrogatorio"* (Varo, Hughes, 2007) etc. More than that, the system itself produces several synonymic translations quite often (as in the case of "*deadly weapon*" - "*arma mortífera, arma mortal, arma letal*"), which are all formally correct. The main subject of discussion when it comes to terminology is whether the terminological resources should include all existing synonyms or rather limit themselves to the most frequents ones/ the ones used within a certain institution/ the ones used in the certain geographical region etc. On the one hand, inclusion of all translation variants aims at giving the translator (or any other target user) as much information as possible. On the other hand, it makes the translation task more complicated. We should also take into account the existence of near synonyms that may reduce, expand or in any other way modify the meaning of the original term. The situation is even more confusing for the legal sphere, where each country has its own judicial system and it own country-specific terms. To sum up, it is up to the translator to make a decision and to select an appropriate equivalent depending on frequency, standardization norms, geographical distribution etc. In this research we do not aim at tackling the dilemma of synonymy. We just label translations that have been attested by reference resources as acceptable and correct. In our opinion, the cases of synonymy should be subject to human revision.

The low system recall score (18.42%) points out the defects of the lexicographic resources used by the tool. These may be of two kinds: a) the term is not found in the dictionary, b) the term is not found in the corpus. In the first case, either both or one of the MWT components is absent in the dictionary. Consequently, the system fails to produce translation candidates to look up in the corpus. This problem might be solved by various

means: 1) manual revision of the system dictionary and addition of missing entries, which is a very time- and labour-consuming task; 2) substitute the general dictionary Collins with an LSP dictionary that is more likely to contain more specific legal terms, which still might not fill all the gaps, taking into account the insufficient coverage of terms even by LSP dictionaries; 3) merge several dictionaries, which, to our mind, would be the best solution. On the other hand, even when all the MWT components are found in the dictionary, it may still be left out by the system, when the MWT is absent in the corpus. Here expanding the specialized corpus size might seem a good solution, yet, as we know, quantity does not always guarantee quality. More effective strategy would be improving the specialized corpus quality at the stage of corpus creation. This may be achieved with the following techniques: 1) filtering web documents by their size. As it was proposed by Fletcher (2004), to leave out the pages, that do not reach a threshold of 5 KB as having little textual content once page headers, menus and the like are removed; as well as the ones that are larger than 200 KB, since these tend to be lists, catalogues, spam etc.; 2) filtering out spam and "boilerplate" material (ads, headers, copyright notices), for example, with the list of stop words; 3) near-duplicate detection, for instance, with Broder's algorithm that takes all n-grams of one document and compares them to the n-grams of another document, as it was done in the Basque language project of Gurrutxaga et al. (2013). The latter even suggest another step to corpus refinement. Instead of seed words as a starting point they take a bunch of seed documents representative of the domain and covering as many of its subfields as possible (so-called sample mini-corpus, 10-20 documents depending on the domain). It serves to extract seed words automatically taking the ones with the highest LLR between the word frequency in general and in mini-corpus. These seed words are used in a procedure similar to WebBootCaT (crawling the web and downloading the pages applying the aforementioned filters). The last step will be final domain filtering: each downloaded document and each sample corpus document is represented with a vector of the most relevant keywords, i.e. nouns, adjectives and verbs. The keywords are again weighted with LLR. A threshold for similarity between the documents from the web and the mini-corpus is established empirically, and only the documents that reach this threshold are accepted in the specialized corpus (Gurrutxaga et al., 2013). In the process of specialized corpus compilation, we also suggest combining texts extracted from the Internet with encyclopedias, legal documents and publications, texts rich in core legal vocabulary (materials for law students and lawyers) selected according to their relevance established by experts in the field.

It is to be noted, that the system is not designed for SWT such as *"parole", "mediator"* and the like as well as for hyphenated terms as *"self-defense"*. In this regard, the system is not the best solution for the terms which are not MWT in the target language (e.g. "*case law*" - "*precedentes*", "*jurisprudencia"*). This problem may become even more significant for the languages such as German, where compounding is the most productive means of word formation. Moreover, in its current design the system is not suitable for translating:

1) MWT with a determiner: definite or indefinite article such as *"reverse the decision"* vs. *"reverse decision", "dismiss the case"* vs. *"dismiss case", "commit a crime"* vs. *"commit crime"* etc. Although the system analyzes the determiners, it does not consider them. What it does is generalizing based on the syntactic analysis. Thus, if the input is to "*commit a crime"*, the system is only interested in the V - **N**$_{\text{CDIR}}$ relation between the verb *"commit"* and the noun *"crime"*. As a consequence, if "*cometer un crimen" (commit a crime), "cometer el crimen" (commit the crime), "cometer crimenes" (commit crimes)* occur in the Spanish corpus, all of them would be counted as the same MWT. The problem of such generalizations is that in some cases the usage of definite/indefinite article (or its absence, i.e. zero article) may change the meaning of the term giving rise to different translations. Moreover, in some cases the term in its standardized form should include a determiner, and omitting it would be inappropriate (e.g. "*commit a crime"* is usually used with an article).

2) Nouns with several modifiers (compound modifiers), e.g. *"supreme legal body"* vs. *"legal body", "prior criminal record"* vs. *"criminal record", "general jurisdiction court"* vs. *"general court"*; expressions containing a noun, a verb and a modifier such as: *"overcome circumstantial evidence"* vs. *"circumstantial evidence", "give expert opinion"* vs. *"give opinion"* or *"expert opinion",* expressions with a conjunction *"and": "search and seizure", "pain and suffering", "costs and fees"* and others. As a rule, the system fails to generate translations for MWT consisting of three and more components except of terms with a preposition such as *"of"*, e.g. *"rule of law", "division of power", "certificate of birth".* In order to be able to treat all or some of these patterns, the system technology should be modified at the level of coding.

However, the biggest limitation of the system is its inability to translate word forms, including:

a) plurals ("*exclusion of witnesses*" - "*exclusión de testigo"*),

b) adjectival endings for feminine nouns ("*arma mortífero"*, "*conducta delictivo"*),

c) verbals acting as noun modifiers ("*concealed weapon*" - "*ocultar arma*", "*leading question*" - "*dirigir pregunta*").

The same challenge was mentioned for other approaches that work with lemmata (Gracia et al, 2007). As we may see, the system treats lemmata, i.e. citation forms:

a) masculine form for adjectives,

b) singular for nouns,

c) infinitive form for verbs.

In the last case, verbals, albeit in the role of modifiers, are PoS-tagged as **VBG** (gerund) and **VBN** (participle). The system identifies the MWT syntactic pattern as **V - N**$_{\text{CDIR}}$ (instead of **ADJ -N**). Thus, the generated translation includes the citation form of the verb and the noun. Whereas in the instances *a)* and *b)* the given limitation merely leads to ungrammatical translations, the case of *c)* is more complex, since it generates inadequate translations. The decision may be creating other programs to automate the processes of translating plural and of checking noun-adjective agreement, establishing it where necessary. The solution for the problem of verbals may be attributing **VBN** and **VBG + N** terms to **N - ADJ** syntactic pattern (being a part of the compound term a gerund and a participle will always fulfill the role of a modifier similar to that of an adjective).

To sum up, the most significant system limitations are rooted not in the system itself, but in the insufficient coverage of legal terminology by the lexicographic resources it is based on. However, the system can and should be improved, for instance, by adapting it to working on the word form level. Even though we mainly disregarded gender and number agreement problems for the purposes of quantitative assessment, we could not leave them out in the process of qualitative evaluation. On the other hand, the innate challenges of terminology and machine translation were quite predictable. Nowadays, the whole scientific community is working on strategies, tools, and standardization norms (in case of terminology) to overcome these challenges.

Conclusions

In the theoretical chapter of this paper we have investigated the principal historical developments and challenges of MT. In this respect, the emphasis was put on the ways lexicographic resources have been used for MT purposes in different periods of its history. Taking into account the lack of parallel corpora for many language pairs and specific domains, we have explored different approaches to automatic terminology extraction from non-parallel comparable corpora. In the practical part, we experimented with WebBootCat tool for specialized corpus creation and employed an extraction tool to compile a bilingual lexicon from a monolingual corpus and a bilingual dictionary.

The evaluation of system performance with the precision and recall metrics allowed us to detect its weaknesses, to analyze their causes and to look for possible solutions to improve the system's potential. On the one hand, the system demonstrated poor recall, i.e. it failed to generate translations for most input terms. This result reveals to which extent the performance of a MT tool depends on the quality and coverage of the lexicographic resources it is based on. In this regard, even the most technically advanced and flawless systems will still demonstrate poor performance if their lexicographic resources are not good or complete enough. We argue that the recall percentage may be increased significantly by improving the quality and coverage of the lexicographic resources (dictionary and corpus) used by the system. On the other hand, most of the high frequency translations generated by the system are correct, i.e. its precision is quite good. Moreover, the system generated correct translations distinctive from the dictionary ones in 53% of cases (for translated terms). In general, we claim that the system may, indeed, help lexicographers and terminologists expand or revise the existing glossaries. We labeled the translations whose existence and use was attested either by reference dictionaries or Web as corpus as correct. Passing judgment on the new translations (whether they are more frequent/appropriate/correct than the ones offered by the glossary), however, was not the task of this research. These decisions should be made by a lexicographer/terminologist specialized in legal terminology taking into account the features of the given legal system and the target user of the database (internal use by a certain institution vs. more general use by different institutions and users). Thus, we cannot fully liberate this process from human post-editing. The main conclusion is that the system is an efficient machine translation aid that can potentially decrease the labour and time spent on the creation of new lexicographic resources and revision of the existing ones.

It is also worth mentioning that checking the system translations we came across the problem of insufficient lexical coverage of paper LSP dictionaries. As a rule, those included

just one translation variant which was in many cases different from the translations found in other dictionaries. Therefore, we had to consult online resources or Web as corpus almost in each case. This problem gives us a hint to the features of the lexicographic resource we should work on — devoid of space limitations, up-to-date, easy to revise on a daily basis, adaptable to different geographical contexts, needs and clients. These are the key features of an electronic medium and a clear shift to E-Lexicography. The drawbacks of the existing LSP dictionaries have confirmed our hypothesis about the usefulness and practical value of the tool for automatic generation of translations from a specialized monolingual corpus for further use in the lexicographic and MT tasks.

In general, we are quite optimistic about the future of the suggested method due to its domain flexibility and openness to expansion with different types of corpora and dictionaries. The future work on the given method may focus both on improvement of the lexicographic resources for the tool or upgrading the tool to improve grammaticality and cohesion of its translations. This type of research may give rise to interdisciplinary projects uniting experts in NLP and lexicography. To conclude, nowadays we can no longer speak about lexicography, translation and information technologies as independent disciplines. In the same way, the borders between rule-based and corpus-based approaches to MT and CAT are becoming less evident, giving way to hybrids which are to combine the strengths of all approaches and methods suggested throughout the history of MT starting from the first “translation machines” by George Astrouni and Petr Smirnov-Troyanskii.

BIBLIOGRAPHY

Reference works:

ABDUL-RAUF, Sadaf; SCHWENK , Holger (2009): *On the use of Comparable Corpora to improve SMT performance.* Published in Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Athens, Greece. pp. 16-23. Available at: https://pdfs.semanticscholar.org/e5b2/7789409eb2cd176d4ef43aee4ecf731397f0.pdf

ARMENTANO OLLER, Carme; CORBÍ BELLOT, Antonio Miguel; FORCADA, Mikel L.; GINESTI ROSELL, Mireia; MONTAVA BELDA, Marco A.; ORTIZ ROJAS, Sergio; PEREZ-ORTIZ, Juan Antonio; RAMÍREZ SÁNCHEZ, Gema; SÁNCHEZ-MARTÍNEZ, Felipe (2007): *Apertium, una plataforma de código abierto para el desarrollo de sistemas de traducción automática.* Universitat d'Alacant, pp. 3 - 6. Available at: http://rua.ua.es/dspace/handle/10045/27531

ARNOLD, D.J. et al. (1994): *Machine Translation: an Introductory Guide.* NCC Blackwell, London, Available at: http://promethee.philo.ulg.ac.be/engdep1/download/bacIII/Arnold%20et%20al%20Machine%20Translation.pdf

BAHDANAU, Dzmitry; CHO, Kyunghyun; BENGIO, Yoshua (2015): *Neural Machine Translation by jointly learning to align and translate*. Conference paper at ICLR 2015. Available at: https://arxiv.org/pdf/1409.0473.pdf

BARONI, Marco; KILGARRIFF, Adam; POMIKÁLEK, Jan and RYCHLÝ, Pavel (2006): *WebBootCat: a Web Tool for Instant Corpora.* In Proceeding of the EuraLex Conference, Italy: Edizioni dell'Orso s.r.l., pp. 123-132. Available at: http://www.euralex.org/elx_proceedings/Euralex2006/016_2006_V1_Marco%20BARONI,%20Adam%20KILGARRIFF,%20Jan%20POMIKALEK,%20Pavel%20RYCHLY_WebBootCaT_a%20Web%20Tool%20for%20instant%20corpora.pdf

BLANCAFORT, Helena; DAILLE, Béatrice; GORNOSTAY, Tatiana; HEID, Ulrich: MECHOULAM, Claude; SHAROFF, Serge (2010): *TTC: Terminology Extraction, Translation Tools and Comparable Corpora.* Available at: https://www.researchgate.net/publication/236616235

BROWN, Peter F. ; COCKE, John ; DELLA PIETRA, Stephen A. ; DELLA PIETRA, Vincent J.; JELINEK, Fredrick; LAFFERTLY, John D.; MERCER, Robert L. and ROOSSIN, Paul S. (1990): *A statistical approach to machine translation.* In Computational Linguistics. Vol. 16, No 2, pp.79-84. Available at: http://www.aclweb.org/anthology/J90-2002

CARBONELL J.G.; BROWN R.D. (2004): *Example-based machine translation*. Available at: http://www.cs.cmu.edu/~ralf/ebmt/ebmt.html

CARL, Michael; MELERO, Maite; BADIA, Toni; VANDEGHINSTE, Vincent; DIRIX, Peter; SCHUURMAN, Ineke; MARKANTONATOU, Stella; SOFIANOPOULOS, Sokratis; VASSILIOU, Marina; YANNOUTSOU, Olga (2008): *METIS-II: low resource machine translation.* Available at: http://openarchive.cbs.dk/bitstream/handle/10398/8037/METIS-II.pdf?sequence=1

CHURCH, K.; HANKS, P. (1989): *Word association norms, mutual information and lexicography.* In Proceedings of the 27th Annual Meeting of the ACL. Vancouver, Canada. pp. 76 - 83. Available at: http://www.aclweb.org/anthology/P89-1010.pdf

COSTA-JUSSA, Marta R. ; FONOLLOSA, José A.R. (2015): *Latest trends in hybrid machine translation and its applications.* In Computer Speech & Language. Vol. 32, pp. 3 - 10. Available at: https://www.sciencedirect.com/science/article/pii/S0885230814001077

DAILLE, Béatrice (2012): *Building bilingual terminologies from comparable corpora: The TTC TermSuite*. In Proceeding of the 5th Workshop on Building and Using Comparable Corpora with special topic ”Language Resources for Machine Translation in Less-Resourced Languages and Domains”. Instanbul, Turkey. pp. 29 - 32. Available at: http://www.lrec-conf.org/proceedings/lrec2012/workshops/16.BUCC2012%20Proceedings.pdf

DILLINGER, Mike (2001): *Dictionary Development Workflow for MT: Design and Management.* In Proceedings of MT Summit VIII. pp.83-88. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.497.3526&rep=rep1&type=pdf

DUNNING, T. (1994): *Accurate methods for the statistics of surprise and coincidence.* In Computational Linguistics. Vol. 19(1), pp. 61-74. Available at: http://aclweb.org/anthology/J93-1003

FABER P.; LOPEZ-RODRIGUEZ C.; TERCEDOR M. (2005): *La utilización de técnicas de corpus en la representación del conocimiento médico.* In Terminology. January 2005, pp. 167-198. Available at: https://www.researchgate.net/publication/313425267

FLETCHER, W. (2004): *Making the web more useful as a source for linguistic corpora*. In Corpus Linguistics in North America 2002: Selections from the Fourth North American Symposium of the American Association for Applied Corpus Linguistics. Amsterdam: Rodopi. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.72.1297

FRANTZI, Katerina; ANANIADOU, Sophia; HIDEKI, Mima (2000): *Automatic recognition of multi-word terms: the C-value/NC-value method.* In International Journal on Digital Libraries. Vol. 3. Springer. pp. 115 - 130. Available at: https://www.researchgate.net/publication/220387502_Automatic_Recognition_of_Multi-word_Terms_The_C-value_NC-value_Method

GAMALLO OTERO, Pablo (2007): *Learning Bilingual Lexicons from Comparable English and Spanish Corpora*. In Proceedings of MT Summit XI, pp. 191-198. Available at: http://www.mt-archive.info/MTS-2007-Gamallo-Otero.pdf

GARCIA, Marcos; GARCÍA-SALIDO, Marcos and ALONSO-RAMOS, Margarita (2017): *Using bilingual word-embeddings for multilingual collocation extraction*. In Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017) at the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017), Valencia, pp. 21-30. Available at: http://multiword.sourceforge.net/mwe2017/proceedings/MWE201703.pdf

GERBER, Laurie; YANG, Jin (1997): *Systran MT dictionary development*. In Machine Translation: Past, Present, and Future: Proceedings of Machine Translation Summit VI. pp. 211 - 218. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.128.7109&rep=rep1&type=pdf

GREFENSTETTE, Gregory (1999): *The World Wide Web as a Resource for Example-Based Machine Translation Tasks*. In Translating and the Computer 21. Proceedings. London: Aslib. Available at: http://www.mt-archive.info/Aslib-1999-Grefenstette.pdf

GURRUTXAGA, Antton; LETURA, Igor; IÑAKI, San Vicente (2013): *Automatic comparable web corpora collection and bilingual terminology extraction for specialized dictionary making*. In Building and Using Comparable Corpora, Springer, pp.51-75. Available at: https://www.researchgate.net/publication/264207424

HUTCHINS, William John (1995): *Machine Translation: A brief history.* In Concise History of the Language Sciences: From the Sumerians to the Cognitivists. Oxford: Pergamon Press. pp. 431-444

HUTCHINS, William John (2005): *Example-based machine translation: a review and commentary*. In Machine Translation. Vol. 19, pp. 197–211. Available at: https://link.springer.com/content/pdf/10.1007%2Fs10590-006-9003-9.pdf

HUTCHINS, William John (2006): *Machine translation: a concise history.* Available at: http://www.hutchinsweb.me.uk/CUHK-2006.pdf

HUTCHINS, William John (2014): *The history of machine translation in a nutshell.* Available at: http://www.mt-archive.info/10/Hutchins-2014.pdf

KILGARIFF, Adam and GREFENSTETTE, Gregory (2003): *Introduction to the Special Issue on the Web as Corpus*. In Computational Linguistics, Vol. 29, Issue 3, pp. 333-347. Available at: https://doi.org/10.1162/089120103322711569

KILGARIFF A.; RYCHLY P.; SMRZ P. and TUGWELL D. (2004): *The Sketch Engine.* In Proceedings of Euralex, Lorient, France, pp. 105-116. Available at: https://www.researchgate.net/publication/260387608_ITRI-04-08_the_sketch_engine

KOEHN, Philipp (2005): Europarl*: A Parallel Corpus for Statistical Machine Translation*. In Conference Proceedings: the tenth Machine Translation Summit, Phuket, Thailand; pp. 79-86. Available at: https://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf

LÄUBLI, Samuel (2017): *3 Reasons Why Neural Machine Translation is a Breakthrough*. Invited keynote at SlatorCon Zurich. Zurich, Switzerland. Available at: https://slator.com/technology/3-reasons-why-neural-machine-translation-is-a-breakthrough/

MIKOLOV, Tomas; LE, Quoc V.; SUTSKEVER, Ilya (2013): *Exploiting Similarities among Languages for Machine Translation.* Available at: https://arxiv.org/abs/1309.4168

MUNTEANU D. S.; MARCU, D. (2002): *Processing comparable corpora with bilingual suffix trees.* In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Philadelphia. pp. 289 - 295. Available at: https://aclanthology.info/pdf/W/W02/W02-1037.pdf

OKPOR M.D. (2014): *Machine Translation Approaches: Issues and Challenges.* In IJCSI International Journal of Computer Science Issues, Vol. 11, Issue 5, No 2, pp. 159-165. Available at: http://www.ijcsi.org/papers/IJCSI-11-5-2-159-165.pdf

RAPP, Reinhard; SHAROFF, Serge; ZWIEGENBAUM, Pierre (2016): *Recent advances in machine translation using comparable corpora.* In Natural Language Engineering. Vol. 22 (4), Cambridge University Press, pp. 501–516. Available at: https://doi.org/10.1017/S1351324916000115

SAG, Ivan A.; BALDWIN, Timothy; BOND, Francis; COPESTAKE, Ann; FLICKINGER, Dan (2002): *Multiword Expressions: A Pain in the Neck for NLP*. Computational Linguistics and Intelligent Text Processing. Available at: http://lingo.stanford.edu/pubs/WP-2001-03.pdf

SÁNCHEZ-MARTÍNEZ, Felipe; FORCADA, Mikel L. and WAY, Andy (2009): *Hybrid rule-based – example based MT: Feeding apertium with sub-sentential translation units.* In Proceedings of the 3rd Workshop on Example-Based Machine Translation, Dublin, pp. 11–18. Available at: https://rua.ua.es/dspace/handle/10045/14024

SATO, S. (1993): *Example-based translation of technical terms*. In TMI-93. Kyoto. pp. 58 - 63. Available at: http://mt-archive.info/TMI-1993-Sato.pdf

SOMERS H.I. (2003): *An Overview of EBMT*. In Recent Advances in Example-Based Machine Translation. Pp. 3-57. Available at: http://utkl.ff.cuni.cz/~rosen/public/somers_EBMT.doc

TRIPATHI, Sneha; SARKHEL, Juran Krishna (2010): *Approaches to machine translation.* In Annals of Library and Information Studies. Vol. 57, pp. 388-393. Available at: https://www.researchgate.net/publication/228574546_Approaches_to_machine_translation

TYERS, Francis M.; SÁNCHEZ-MARTÍNEZ, Felipe; ORTIZ-ROJAS, Sergio; FORCADA , Mikel L. (2010): *Free/Open-Source Resources in the Apertium Platform for Machine Translation Research and Development*. In The Prague Bulletin of Mathematical Linguistics. pp. 67-76. Available at: https://ufal.mff.cuni.cz/pbml/93/art-tyers-et-al.pdf

XUAN H. W.; LI W.; TANG G. Y. (2014): *An Advanced Review of Hybrid Machine Translation (HMT)*. In Procedia Engineering. Vol. 29, pp. 3017-3022. Available at: https://www.researchgate.net/publication/271891394_An_Advanced_Review_of_Hybrid_Machine_Translation_HMT

WITKAM, Toon (2006): *History and heritage of the DLT (Distributed Language Translation) project.* Utrecht, Netherlands. Available at: http://www.mt-archive.info/Witkam-2006.pdf

WU, Yonghui; SCHUSTER, Mike; CHEN, Zhifeng ; LE, Quoc V.; NOROUZI, Mohammad et al. (2016): *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. Available at: https://arxiv.org/pdf/1609.08144.pdf

Reference Dictionaries:

Collins English-Spanish dictionary. Electronic resource. Available at: https://www.collinsdictionary.com/dictionary/english-spanish

Diccionario de términos jurídicos : inglés-español = A dictionary of legal terms : Spanish-English / Enrique Alcaraz Varó y Brian Hughes ; prólogo de Ramón Martín Mateo. Barcelona : Ariel, 2007. 1060 p.

Diccionario de términos jurídicos : inglés-español, español-inglés / Francisco Ramos Bossini, Mary Gleeson. Granada : Comares, 1997. 535 p.

Diccionario Espasa términos jurídicos Español-inglés, English-Spanish. Madrid : Espasa, D.L. 2002 . 414 p.

English/Spanish Legal Glossary/Glosario Legal. Superior Court of California, County of Sacramento. Translated from English into Spanish by Rodrigo Mayorga, Esq.2006

Essential english/spanish and spanish/english legal dictionary / Steven M. Kaplan. Alphen aan den Rijn (The Netherlands) : Kluwer Law International, cop. 2008. 521 p.

Glossary of legal terminology English to Spanish. State of Connecticut Judicial Branch Superior Court Operations Division. Edited by LOMBARDI, John. Electronic resource: https://www.jud.ct.gov/external/news/jobs/interpreter/Glossary_of_Legal_Terminology_English-to-Spanish.pdf (Accessed in April and May, 2018)

IATE. Inter-Active Terminology for Europe. Electronic resource: http://iate.europa.eu/SearchByQueryLoad.do?method=load (Accessed in April and May, 2018)

Linguee. Electronic resource: https://www.linguee.es/ (Accessed in April and May, 2018)

Spanish to English dictionary - Law Firm Law Office of Douglas C. Smith Attorneys El Paso, Texas. Available at: https://www.dcsmithpllc.com/Spanish-To-English-Dictionary.shtml

WordReference. Electronic resource: https://www.wordreference.com/es/ (Accessed in April and May, 2018)

ANNEXES

Annex 1. Evaluation data set

English term	Dictionary translation(s)	Gold standard	System translation(s)	System (most frequent)	Most frequent correct
abuse of process	abuso del proceso	abuso del proceso, abuso de procedimiento	abuso de procedimiento	abuso de procedimiento	1
adhesion contract	contrato de adhesión	contrato de adhesión	contrato de adhesión	contrato de adhesión	1
affirmative defense	defensa afirmativa	defensa afirmativa	defensa afirmativo	defensa afirmativo	1
assumption of risk	presunción de riesgo	presunción de riesgo, asunción de riesgo	asunción de riesgo, supuesto de riesgo	asunción de riesgo	1
blood test	examen de sangre	examen de sangre, prueba de sangre	prueba de sangre	prueba de sangre	1
breach of peace	alteración del orden público	alteración del orden público, ruptura de paz	ruptura de paz	ruptura de paz	1
case law	precedentes	precedentes	derecho de caso, ley de pleito, ley caso, ley pleito	derecho de caso, ley de pleito, ley caso, ley pleito	0
cause of action	acción del litigio	acción del litigio, causa de acción, causa de acto	causa de acción, causa de acto, causante de acción, desencadenante de acción	causa de acción	1
chain of custody	cadena de custodia	cadena de custodia	cadena de custodia	cadena de custodia	1
chief judge	primer magistrado	primer magistrado, juez principal	juez principal	juez principal	1
child abuse	abuso de niños	abuso de niño(s), abuso de menor	abuso de menor, abuso de niño, abuso	abuso de menor	1

			menor		
child support	sustento de menor	sustento de menor, sostenimiento de hijo, sustento de hijo, sustento de menor	sostenimiento de hijo, sostenimiento de menor, sustento de hijo, sustento de menor	sostenimiento de hijo, sustento de menor	2
civil bail	caución civil	caución civil, fianza civil	fianza civil	fianza civil	1
civil case	caso civil	caso civil, pleito civil	caso civil, pleito civil	caso civil	1
closing argument	argumento final	argumento final	cerrar discusión	cerrar discusión	0
common law	derecho consuetudinario	derecho consuetudinario, derecho común	derecho común, ley común, común derecho	derecho común	1
concealed weapon	arma oculta	arma oculta	ocultar arma	ocultar arma	0
concurrent jurisdiction	jurisdicción simultanea	jurisdicción simultanea, jurisdiccón concurrente	jurisdiccón concurrente	jurisdiccón concurrente	1
conflict of interest	conflicto de intereses	conflicto de interes(es)	conflicto de interés, pugna de interés	conflicto de interés	1
constitutional right	derecho constitucional	derecho constitucional	derecho constitucional	derecho constitucional	1
court costs	costos del tribunal	costo(s) del tribunal, costo de tribunal	costo de corte, costo de tribunal	costo de corte, costo de tribunal	1
court of appeals	tribunal de apelaciones	tribunal de apelacion(es), tribunal de apelación	tribunal de apelación, corte de apelación	tribunal de apelación	1
criminal case	caso penal	caso penal, caso criminal	caso criminal, delictivo caso	caso criminal	1
criminal conduct	conducta delictiva	conducta delictiva, conducta criminal	conducta delictivo, conducta criminal, delictivo	conducta delictivo	1

			conducta		
criminal record	registro criminal, antecedentes penales	registro criminal, antecedentes penales	registro criminal	registro criminal	1
deadly weapon	arma mortífera	arma mortífera, arma mortal, arma letal	arma mortífero, arma mortal, arma letal	arma mortífero, arma mortal	2
defense attorney	abogado defensor	abogado defensor, abogado de defensa	abogado de defensa	abogado de defensa	1
degree of crime	grado del crimen	grado del crimen, grado de crimen, grado de delito	grado de crimen, grado de delito	grado de crimen, grado de delito	2
direct evidence	evidencia directa	evidencia directa, testimonio directo	testimonio directo, evidencia directo	testimonio directo	1
direct examination	interrogatorio directo	interrogatorio directo, examen directo	interrogatorio directo, examen directo, directo interrogatorio, inspección directo	interrogatorio directo	1
domestic violence	violencia doméstica	violencia doméstica	violencia doméstico	violencia doméstico	1
drug dealer	narcotraficante	narcotraficante, traficante de droga	traficante de droga	traficante de droga	1
equal protection	igual protección ante la ley	igual protección ante la ley, igual protección	igual protección	igual protección	1
exclusion of witnesses	exclusión de testigos	exclusión de testigo(s)	exclusión de testigo	exclusión de testigo	1
expert testimony	testimonio de experto	testimonio de experto, testimonio pericial	testimonio pericial	testimonio pericial	1

hostile witness	testigo hostil	testigo hostil	testigo hostil	testigo hostil	1
incurable insanity	enfermedad incurable	enfermedad incurable, demencia incurable	demencia incurable	demencia incurable	1
joint tenancy	condominio	condominio, tenencia conjunta	tenencia conjunto	tenencia conjunto	1
juvenile court	tribunal juvenil, tribunal de menores	tribunal juvenil, tribunal de menores, corte juvenil	corte juvenil, juzgado juvenil, juvenil juzgado	corte juvenil	1
leading question	pregunta sugestiva	pregunta sugestiva	dirigir pregunta	dirigir pregunta	0
legal aid	ayuda legal	ayuda legal	ayuda legal	ayuda legal	1
material fact	hecho material	hecho material	hecho material	hecho material	1
opening argument	alegato inicial	alegato inicial	abrir discusión	abrir discusión	0
penalty of perjury	pena de perjurio	pena de perjurio	pena de perjurio	pena de perjurio	1
personal property	bienes muebles	bienes muebles, propiedad personal, inmueble personal	propiedad personal, inmueble personal	propiedad personal	1
power of attorney	por poder	por poder, poder de abogado	poder de madatorio, poder de abogado	poder de mandatorio	0
preliminary hearing	audiencia preliminar	audiencia preliminar, vista preliminar	vista preliminar	vista preliminar	1
presumption of law	presunción de ley	presunción de ley, presunción de derecho	presunción de ley, presunción de derecho	presunción de derecho	1
sexual harassment	acoso sexual	acoso sexual, hostigamiento sexual	hostigamiento sexual, acoso sexual	hostigamiento sexual	1

statutory law	ley reglamentaria	ley estatuaria, ley reglamentaria, derecho estatuario	ley estatuario, ley reglamentario, derecho estatuario	ley estatuario, ley reglamentario, derecho estatuario	3

Annex 2. The list of abbreviations

CAT - Computer-assisted translation

CBOW - Continuous Bag-of-Words

EBMT - Example-based Machine Tranlslation

HMT - Hybrid Machine Translation

LLR - Log-likelihood ratio

LSP - Language for specific purposes

MT - Machine Translation

MWE - Multiword expression

MWT - Multiword term

NLP - Natural Language Processing

NMT - Neural Machine Translation

PoS - Part of speech

RBMT - Rule-based Machine Translation

SL - Source language

SMT - Statistical Machine Translation

SWT - Single-word term

TER -Translation Error Rate

TL - Target language

TTC - Terminology Extraction, Translation Tools and Comparable Corpora

WER - Word Error Rate