Automated detection of inconsistent phraseology translation

Tom Vanallemeersch^{1*} and Hendrik Kockaert²

^{1, 2}Department of Applied Language Studies, Lessius University College (Katholieke Universiteit Leuven), Sint-Andriesstraat 2, Antwerpen 2000, Belgium

²Department of Language Management and Language Practice, University of The Free State, PO Box 339, Bloemfontein 9300, South Africa

*Corresponding author, e-mail: tom.vanallemeersch@lessius.eu

Abstract: We investigate the extent to which the detection of phraseological (in)consistency in the translation process can be automated. We describe the acquisition of a large corpus of Belgian legal documents consisting of French arrests translated into Dutch. We apply the sentence alignment tool GMA to the corpus, and extract phraseological unit candidates from the sentence pairs through the term candidate extraction tool TermCalc and the word alignment data produced by the GIZA++ tool. The candidates are compared to a reference set from a manual study of an MA student at Lessius/KULeuven. They appear to cover only 33% of the bilingual phraseological unit pairs and only four French units with more than one Dutch equivalent. This indicates the need for devising techniques specifically aimed at detecting multiple equivalence, hence potential phraseological inconsistency.

Background

Phraseological units are technical expressions, technical phrases or standard formulations (COTSOES, 2003). They underline the specialist nature of texts. One of the aspects in which these units differ from terms is the fact that their parts are less fixed. For instance, depending on the context, one may use the unit 'issue shares' or 'put shares on the market'. In the domain of legal texts, phraseological units are 'typically reproduced in certain recurrent contexts' and are 'characterized by a context-dependency ... in legal language' (Kjær, 1990:26). The context consists of both the co-text, which is an institutionalised and inherited use of archetypical legal language, and the extratextual context (Kockaert, Vanallemeersch & Steurs, 2007). When phraseological units are translated in an inconsistent way, they diminish the specialist nature of the target text.

The official languages of Belgium are Dutch, French and German. There is a lack of exhaustive resources, such as normative databases, for multilingual phraseology in the Belgian context. Most official legal documents (except for some older ones) are available online on Juridat, the portal of the judiciary in Belgium (Juridat, 2008). The documents are often available in both French and Dutch, sometimes in other language combinations, or in just one language. The user accesses the documents through a query window. Multilingual repositories of legal documents such as Juridat are primarily targeted towards legal and other specialists looking for specific articles, rather than towards translators. For instance, in the framework of an MA thesis at Lessius/KULeuven (Ingelbeen, 2008), the student spent several weeks finding out the Dutch translation equivalents of a mere 50 French phraseological units in the arrests of the Court of Cassation,¹ the highest court in Belgium. Figure 1 shows that the majority of units were assigned one to three equivalents by the student. The student lists a total of 110 bilingual unit pairs (i.e. the sum of all equivalents). Three examples are:

acquittement du prévenu ('acquittal of accused'): *vrijspraak van de beklaagde*
acte irrégulier ('irregular act'): *onregelmatige handeling/onregelmatigheid*
action publique ('criminal proceedings'): *strafvordering*.

In the last decades, more and more digital parallel corpora have become available, and researchers have become aware of their potential for several purposes, such as machine translation, translation practice, contrastive linguistics, multilingual terminology, and so on. In order to

Figure 1: Distribution of phraseological units across number of equivalents (Ingelbeen, 2008)

exploit the information in these corpora, tools have been devised for sentence alignment, which consists of linking source sentences to their target counterpart; see for instance the Geometric Mapping and Alignment (GMA) method by Melamed (2000). A number of tools have been developed for the detection of word (group) equivalences within sentence pairs. They detect such equivalences through word alignment (Tiedemann, 2003), that is, the linking of words across sentences, or through bilingual term candidate extraction, namely the association of term candidates that were extracted monolingually or the simultaneous extraction of source term candidates and their candidate equivalents from sentence pairs (Thurmaier, 2003). An example of a tool producing word alignment data is the statistical machine translation tool GIZA++ (Och & Ney, 2003). Bilingual term candidate extraction is performed by programs such as TermCalc (Vanallemeersch, 2003). Tools for the exploitation of information in parallel corpora are based on statistics, linguistic knowledge, or a combination of both.

In this paper, we investigate the extent to which phraseological (in)consistency in translation can be detected automatically. In the following sections, we describe the acquisition of a parallel legal corpus for French and Dutch and the sentence alignment process applied to it. We subsequently describe the extraction of French phraseological unit candidates and their Dutch equivalents from sentence pairs, and the comparison of the results with a reference set, that is, the bilingual unit pairs of Ingelbeen (2008). Finally, we present our conclusions and future proposed research.

Corpus acquisition

We compiled a parallel corpus consisting of all arrests of the Court of Cassation that are available through Juridat for the time span 2000 to 2006 and whose source language is French. As Juridat contains thousands of arrests, creating a corpus using the query window is not feasible. Therefore, we downloaded arrests in batch. By consulting some websites specialising in legal matters, we determined the template of web addresses which allowed us to gain direct access to the arrests. Such addresses contain keywords that indicate language and date (year, month and day). We generated automatically a list of potential addresses ranging over all days of the abovementioned

time span. We downloaded the list using wget,² a web crawler. We converted the downloaded HTML files into pure text using the HTMLAsText³ utility and converted each paragraph into a single line rather than a set of lines.

From the list of downloaded arrests, we removed the files that were only available in one language. From the bilingual file pairs, we only retained those whose original language is French, by checking whether the Dutch arrest mentions that it originates from the French-speaking section of the Court of Cassation, and that the text is a translation. We noticed that some of the Dutch files only provide a translation for the most important information of the French file. As this type of translation may lead to sentence alignment problems, we removed all bilingual file pairs showing a difference in length of more than 100% between the files. Our final corpus consists of 1 500 parallel arrests, ranging in size from 600 to 29 000 words (sum of words in source and target text). The high majority of the parallel arrests (88%) contain no more than 6 000 words. The corpus contains a total of 5 million words. The average parallel arrest size is 3 400 words.

Sentence alignment

We wrote a script that segments the texts in the parallel legal corpus into a list of sentences. The script disambiguates periods, for instance through its knowledge of abbreviations. We performed sentence alignment on the bilingual pairs of sentence lists using the GMA method. This method consists of two steps. First, it creates a list of anchors. These are potential points of correspondence between the source and the target text, which link identical words (digits, proper names), cognates (similar orthography), or lexical equivalents (using a bilingual lexicon). The method also uses stop word lists (e.g. function words) in order to avoid an excessive number of potential correspondences. The second step of GMA consists of linking one or more source sentences to one or more target sentences by grouping anchors. For the first step, we restricted language-specific knowledge to stop word lists for French and Dutch, as no extensive bilingual lexicon was at hand.

The sentence alignment on the 1 500 arrest pairs resulted in 92 000 one-to-one links. We ignored links that involve more than one sentence in at least one language (one-to-many or many-to-many links) and null links (sentences without equivalent), assuming that they may be the product of a lack of alignment evidence. We estimated the quality of the alignment result by selecting a small sample of aligned arrest pairs of different sizes. It turned out that 87% of the sentence links were correct; some erroneous links are due to text segmentation errors. Our goal being an investigation into the extent to which automated extraction allows for detecting phraseological (in)consistency, we did not undertake a manual correction of the alignment errors.

Phraseological unit candidate extraction

We extracted French phraseological unit candidates and their Dutch equivalents from aligned sentence pairs using TermCalc and GIZA++. In the section on evaluation, the extraction results will be compared to the reference set and to each other.

TermCalc was developed as an experimental tool in a translation agency.⁴ It is designed for processing a translation memory, which is a database updated automatically with sentences during the translation process.⁵ TermCalc simultaneously extracts source term candidates and their candidate equivalents from sentence pairs, without applying a monolingual step. It creates a list of n-grams (sequences of one or more words) in the source and target text, filtering out those who start or end with a stop word and those containing more words than a predefined maximum. It calculates an association score for each combination of source and target n-gram. This score is based on the formula in Figure 2.⁶ C_{freq} stands for the number of sentence pairs in which both S (the source n-gram) and T (the target n-gram) occur. S_{freq} stands for the number of source sentences in which S occurs and T_{freq} for the number of target sentences in which T occurs. As an example, there are 12 source sentences in the parallel legal corpus containing *réclusion à perpétuité* ('life imprisonment') and 11 target sentences containing *levenslange opsluiting*. C_{freq} is 11: all sentence pairs containing T also contain S. There is one source sentence which contains a translation equivalent of S which is different from T. Therefore, the association score is 0.96. The maximal score is 1; pairs of n-grams with a score below 0.5 are filtered out.

$$\frac{\left(\frac{C_{freq}}{S_{freq}} + \frac{C_{freq}}{T_{freq}} \right)}{2}$$

Figure 2: Calculation of association score in TermCalc

We converted the GMA output into a translation memory in TMX,⁷ which is the input format used by TermCalc. We ran TermCalc on the file, imposing a maximum length of five words on term candidates.⁸ Figure 3 shows a screenshot of the output. The fields SentLinked, SentFreq1 and SentFreq2 correspond to C_{freq} , S_{freq} and T_{freq} , respectively. Below the term candidate table, context sentence pairs for a specific pair of S and T are listed. The program also lists sentence pairs containing S but not T, or vice versa (Figure 3 contains *réclusion à perpétuité*, *levenslange opsluiting* and *levenslange hechtenis*). The output of the program consists of 6 200 French n-grams with one or more Dutch equivalents. A total of 855 n-grams (14%) have two equivalents, the others have only one.

GIZA++ is a tool for the creation of a statistical machine translation system. It builds a probabilistic translation model by deriving statistical information from a large sentence-aligned corpus. To this purpose, it performs word alignment within sentence pairs. Based on the word alignment, it associates n-grams (called ‘phrases’ in statistical machine translation) with a certain probability (a number between 0 and 1). We ran GIZA++ on the parallel legal corpus and filtered out phrases starting or ending with a stop word or longer than five words, in order to make the tool’s output comparable to the TermCalc results. We did not filter out associations with low probabilities, as the latter cannot simply be considered as low scores in a probabilistic model. The output contains 11 500 French phrases with one or more Dutch equivalents. Examples of phrase associations are listed in Table 1. Figure 4 shows that a high majority (80%) of the phrases have one equivalent, the others have two or more.

Evaluation of extraction

We compared the automatically extracted phraseological unit candidate pairs, that is the bilingual term candidates of TermCalc and the phrase pairs of GIZA++, with the reference set. For each French or Dutch phraseological unit in the reference set, we looked for a matching candidate in the tools’ output, more precisely, a unit with an identical canonical form (singular for nouns, infinitive for verbs). As an example, the reference pair *juge du fond – bodemrechter* (both singular nouns) matches with the TermCalc pair *juges du fond – bodemrechters* (both plural nouns). We determined the recall of the tools’ output (i.e. the percentage of the reference set that matched the output). Table 2 shows the recall of each tool and of their combined output with respect to the 50 French units (irrespective of their Dutch equivalents), to all 110 French-Dutch unit pairs, and to the 31 French units with at least two equivalents. In the latter case, the output matches a French unit and at least two of its Dutch equivalents. (Table 3 expands on the GIZA++ output for French units with two or more Dutch equivalents).

Comparing the output of both extraction tools, it appears that only GIZA++ detects French reference units with multiple equivalents, as indicated in Tables 2 and 3. This is due to the fact that TermCalc, upon finding a strong association between a French n-gram and a Dutch n-gram, tends to ignore weaker associations with other Dutch n-grams. From the reference bilingual unit pairs detected by the tools (21 for TermCalc and 28 for GIZA++), 13 are in the output of both tools. There is no clear-cut explanation for the differences in detection of bilingual pairs; further investigation is needed, as well as a more exhaustive evaluation of the tools, as indicated below.

The test we carried out for the evaluation of the extraction tools is limited, in the sense that we only checked the output for the presence of French phraseological units and Dutch equivalents from

Figure 3: TermCalc output

Table 1: Phrase associations in GIZA++

French phrase	English translation	Dutch equivalent phrase	Probability
annulation de la décision	'annulment of the decision'	nietigverklaring van de beslissing	1.00
question préjudicielle	'preliminary question'	prejudiciële vraag	0.55
		prejudiciële vraag te stellen	0.05
		prejudicieel geschil	0.05
		prejudicieel	0.05
		dergelijk prejudicieel geschil	0.05
		dergelijk prejudicieel	0.05

a small reference set, and not for the correctness of all French unit candidates and their proposed Dutch equivalents. Had we wanted to test the performance of the extraction tools in an exhaustive manner, we would have needed to annotate French units and their Dutch equivalents in the parallel legal corpus by hand. This way, we would be able to build a complete reference set for the corpus, allowing us to calculate not only the recall of the extraction output, but also its precision (i.e. the percentage of units in the output which also occurs in the reference set)⁹. The ideal phraseology extraction tool lists all real phraseological units and their real (possibly inconsistent) translation equivalents (i.e. perfect recall), and lists only those and no others (perfect precision). In practice, however, extraction tools become less precise the higher their recall becomes (i.e. an increase in quantity leads to a decrease in quality); therefore, a trade-off between recall and precision is a more realistic aim.

As a preliminary, basic test for evaluating precision, we looked at a sample of the bilingual n-gram pairs in the output of the tools, ignoring their phraseological status and checking whether the Dutch n-gram is a true translation equivalent of the French n-gram. In the case of TermCalc, we looked at a sample of 400 bilingual term candidates with high scores. It appeared that 88%

Figure 4: Distribution of French phrases across number of equivalents in GIZA++

Table 2: Recall of each tool's output and of combined output with respect to the reference set

Tool	French units	French-Dutch unit pairs	French units ≥ 2 equivalents
TermCalc	21 out of 50 (recall = 42%)	21 out of 110 (recall = 19%)	0 out of 31 (recall = 0%)
GIZA++	24 out of 50 (recall = 48%)	28 out of 110 (recall = 25%)	4 out of 31 (recall = 13%)
Combined	32 out of 50 (recall = 64%)	36 out of 110 (recall = 33%)	4 out of 31 (recall = 13%)

Table 3: French units with 2 equivalents in the GIZA++ output, matching the reference set

French unit	English translation	Dutch equivalent
dispositions légales violées	'violated legal provisions'	geschonden wettelijke bepalingen geschonden wetbepalingen
effet dévolutif	'transfer of jurisdiction'	devolutieve kracht devolutieve werking
question préjudicielle	'preliminary question'	prejudicieel geschil prejudiciële vraag
critique	'criticised'	opkomen (tegen) bekritiseerd

of the bilingual candidates are correct according to our criterion. In the case of GIZA++, we looked at a sample of 400 pairs in which the Dutch equivalent has a high probability and found that 78% of the pairs were correct. Hence, the precision of at least a part of the output is acceptable, as long as we abstract from the question whether the bilingual candidates actually consist of phraseological units.

Conclusions and future research

In this paper, we have investigated the extent to which the detection of phraseological (in)consistency in translation can be automated. A first step towards automation consists of creating a large sentence-aligned corpus. This provides a basis for further automation, and also makes the manual retrieval of information more efficient in comparison to a web interface such as Juridat. We compiled a parallel legal corpus containing 92 000 sentence pairs from French documents translated into Dutch which were downloaded from Juridat. A second step towards automation consisted of extracting phraseological unit candidates in the source language and their translation equivalents from the sentence-aligned corpus. When combining the term candidate output from the tool TermCalc with the word alignment output produced by the tool GIZA++, we found a rather limited percentage (33%) of the French-Dutch phraseological unit pairs to be present in a reference set from a manual study of an MA student at Lessius/KULeuven. The output contains only a few multiple equivalences (four French units with two equivalents) from the reference set, which indicates that detecting potential inconsistencies in phraseology translation is a task difficult to automate.

Future research will involve an enhanced sentence alignment procedure, manual corpus annotation, and techniques or tools specifically aimed at detecting multiple equivalence. We estimate the sentence alignment quality to be 87%, a score which may be improved by comparing output from different alignment tools (Rura *et al.*, 2008). We will manually annotate French phraseological units and their equivalents in a part of the parallel corpus in order to calculate not only recall but also precision. As for detection of multiple equivalence, we will test or devise additional techniques. For instance, French phraseological units may be translated into Dutch equivalents which are similar to one another (which is also illustrated by the four cases of multiple equivalence in the extraction output). A possible approach could consist of the monolingual clustering of terms described in Bourigault and Jacquemin (1999).

Notes

- ¹ In Dutch, this court is called *Hof van Cassatie*, in French *Cour de Cassation*.
- ² The GNU web crawler is available at: <http://www.gnu.org/software/wget>.
- ³ The HTMLAsText utility is available at: <http://www.nirsoft.net>.
- ⁴ At the time of development, it was called LNeTerm, on the analogy of the name of the agency, LNE International.
- ⁵ Upon translating a sentence, the user of a translation memory may be provided with a perfect match (a source sentence in the database identical to the sentence, together with its translation) or a fuzzy match (a source sentence in the database similar to the sentence).
- ⁶ The formula is similar to the Dice score (Pazienza *et al.*, 2005).
- ⁷ This is an open XML standard that contains flat text with tags and allows for exchanging translation memories between different translation support tools. TMX is developed and maintained by LISA (<http://www.lisa.org>).
- ⁸ This is the technical maximum allowed by the tool.
- ⁹ Instead of annotating the whole parallel corpus, which would be a very time-consuming task, we may opt for annotating a part of it. In that case, precision calculation would be restricted to those phraseological unit candidates which actually occur in that part of the corpus. The annotation is preferably carried out by two persons in order to calculate the interannotator agreement; see also the problems with disagreement in case of manual word alignment (Melamed, 1998).

References

- Bourigault D & Jacquemin C. 1999. Term extraction + term clustering: An integrated platform for computer-aided terminology. *Proceedings of the European Chapter of the Association for Computational Linguistics* (EACL), 8–12 June 1999, University of Bergen, Bergen, pp 15–22.
- Conference of Translation Services of European States (COTSOES). 2003. *Recommendations for terminology work*. Berne: MediaCenter of the Confederation/ SFBL Publications.
- Ingelbeen C. 2008. Terminologische inconsistentie in Nederlandstalige arresten van het Hof van

- Cassatie. Unpublished MA thesis, Lessius/KU Leuven.
- Juridat.** 2008. De portaalsite van de rechterlijke macht in België/Le portail du pouvoir judiciaire de Belgique [The portal of the judiciary in Belgium]. Available at: <http://www.juridat.be> [accessed 18 November 2010].
- Kjær AL.** 1990. Context-conditioned word combinations in legal language. *Journal of the International Institute for Terminology Research (IITF)* 1(1/2): 21–32.
- Kockaert HJ, Vanallemeersch T & Steurs F.** 2007. Term-based context extraction in legal terminology: A case study in Belgium. *Proceedings of Current Trends in Terminology, International Conference on Terminology*, Szombathely, Hungary. 9–10 November 2007. Available at: <http://lirias.lessius.eu/bitstream/123456789/2925/1/Article.pdf> [accessed 15 April 2008].
- Melamed D.** 1998. Annotation style guide for the Blinker Project, IRCS Technical Report #98-06.
- Melamed D.** 2000. Pattern recognition for mapping bitext correspondence. In Véronis J (ed.) *Parallel text processing: Alignment and use of translation corpora*. Amsterdam, The Netherlands: Kluwer Academic Publishers, pp 25–47.
- Och FJ & Ney H.** 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1): 19–51.
- Pazienza MT, Pennacchiotti M & Zanzotto FM.** 2005. Terminology extraction: An analysis of linguistic and statistical approaches. In Markellou P, Rigou M & Sirmakessis S (eds) *Knowledge mining: A quantitative synthesis of research results and findings*. Berlin, Germany: Springer Verlag.
- Rura L, Vandeweghe W & Montero Perez M.** 2008. Designing a parallel corpus as a multifunctional translator's aid. *Proceedings of the XVIII International Federation of Translators (FIT) World Congress*, Shanghai, China 4–7 August 2008.
- Thurmair G.** 2003. Making term extraction tools usable. *Proceedings of the Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, Controlled Translation (EAMT-CLAW-03)*, Dublin City University, Ireland 15–17 May 2003, pp 170–179.
- Tiedemann J.** 2003. Recycling translations: Extraction of lexical data from parallel corpora and their applications in natural language processing. Unpublished PhD thesis, Uppsala University, Sweden.
- Vanallemeersch T.** 2003. TermCalc, experimental terminology extraction tool developed at LNE International, a Belgian translation agency.