TExSIS: Bilingual Terminology Extraction from Parallel Corpora Using Chunk-based Alignment

Lieve Macken
LT^3, Language and Translation Technology Team, University College Ghent, Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 (S9), 9000 Ghent, Belgium
lieve.macken@hogent.be

Els Lefever
LT^3, Language and Translation Technology Team, University College Ghent, Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 (S9), 9000 Ghent, Belgium
els.lefever@hogent.be

Veronique Hoste
LT^3, Language and Translation Technology Team, University College Ghent, Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
Faculty of Linguistics, Ghent University, Blandijnberg 2, 9000 Gent, Belgium
veronique.hoste@hogent.be

We report on TExSIS, a flexible bilingual terminology extraction system that uses a sophisticated chunk-based alignment method for the generation of candidate terms, after which the specificity of the candidate terms is determined by combining several statistical filters. Although the set-up of the architecture is largely language-independent, we present terminology extraction results for four different languages and three language pairs. Gold standard data sets were created for French-Italian, French-English and French-Dutch, which allowed us not only to evaluate precision, which is common practice, but also recall.
We compared the TExSIS approach, which takes a multilingual perspective from the start, with the more commonly used approach of first identifying term candidates monolingually and then aligning the source and target terms. A comparison of our system with the LUIZ approach described by Vintar (2010) reveals that TExSIS outperforms LUIZ both for monolingual and bilingual terminology extraction. Our results also clearly show that the precision of the alignment is crucial for the success of the terminology extraction. Furthermore, based on the observation that the precision scores for bilingual terminology extraction outperform those of the monolingual systems, we conclude that multilingual evidence helps to determine unithood in less related languages.

Keywords: automatic term extraction, bilingual term extraction, parallel corpora, alignment, chunks

1. Introduction

Due to the rapidly evolving technology and the new terms describing those quickly evolving specialized technological fields, terminology management is becoming increasingly important. Although there are several multilingual term banks, such as IATE[1], TermiumPlus[2], EuroTermBank[3], and TermSciences[4], which are freely available, they usually include terms from a wide range of specialized fields and cannot keep pace with the rapidly evolving technological fields. Moreover, customers may prefer different terms to the ones included in the publicly available term banks. Nor do all of the above-mentioned term banks support smaller languages, such as Dutch. For that reason, terminological work will always be part of the translator's workload. It goes without saying that compiling terminology is a labour-intensive process. Therefore, computer-assisted terminology acquisition will definitely lead to increased efficiency.

Terminology extraction can be defined as the study of terms and encompasses diverse activities such as the collection, description and structuring of terms. According to Wright (1997), terms are "the words that are assigned to concepts used in the special languages that occur in subject-field or domain-related texts". Terms consist of single-words or multi-word units that represent discrete conceptual entities, properties, activities or relations in a particular domain (Bowker, 2008). In practice, it is difficult to define precisely what a term is. In particular, the inclusion of complex terms that are fully compositional largely depends on the ultimate use of the term list (e.g. in the case of translation, the inclusion of complex terms might lead to an improved translation consistency). Daille (1996) describes base terms and the mechanisms to create more complex terms on the basis of those base terms for French. Daille admits that "it is difficult to determine whether a modified or overcomposed base term is or is not a term" (Daille, 1996, p. 30) and points out that abbreviations or acronyms can be used as clues that a modified base term is a term. Unfortunately, this approach cannot be generalized as not all terms are abbreviated. The issue of what constitutes a term is even more difficult in a bilingual setting, as the word formation rules differ across languages and terms that are fully compositional in one language might not be compositional in another language, e.g. the French term *vide-poches* is not compositional, whereas the English (*storage compartment*), Italian (*cassettino*

portaoggetti), and Dutch (*opbergvak*) counterparts are. Therefore, in this paper, our reference list does not only contain single word terms and multi-word terms of length two (binary terms) but also multi-word terms involving any number of content words.

In this paper, we introduce the TExSIS bilingual terminology extraction system, a largely language-independent system, and present terminology extraction results for four different languages and three language pairs, viz. French-Italian, French-English and French-Dutch. The TExSIS system uses a sophisticated chunk-based alignment method for the generation of candidate terms. The system is conceived as a two-phase system, which first generates candidate terms directly from linguistically motivated aligned chunks. In a second step, the specificity of the candidate terms is determined by combining several well-known statistical filters. The novelty of the TExSIS approach evidently lies in its first phase. In contrast to most bilingual terminology extraction systems, which first identify term candidates monolingually and then extract translation candidates from parallel corpora using word alignment or co-occurrence information, we take a multilingual perspective from the start. In doing so, we hypothesize that:

(1) chunk alignment based on precise word alignments is beneficial for terminology extraction;
(2) multilingual evidence helps to determine unithood (i.e. the degree of cohesiveness between the elements in multi-word terms).

In order to evaluate the performance of the TExSIS system, we manually created a gold standard corpus, which allowed us not only to evaluate precision (as was done by for example Vintar (2010)), but also recall (Vivaldi & Rodriguez, 2007). A sole focus on precision only gives an indication of whether the terms proposed by the system are relevant. But as precision and recall are closely related (increasing recall generally decreases precision), we firmly believe that it is equally relevant to measure the capability of the system to retrieve *all* relevant terms in a given domain-specific document collection.

We compared the output of the TExSIS system with LUIZ, an academic state-of-the-art bilingual terminology extraction system (Vintar, 2010), which first identifies term candidates monolingually and uses a bilingual lexicon to align the translations and with two commercial systems, viz. Similis and SDL Multiterm Extract. We

clearly show that the TExSIS system outperforms these systems for the three language pairs both in terms of precision and in terms of recall.

The remainder of this paper is structured as follows. Section 2 motivates the present study and gives an overview of related work on automatic term recognition. Section 3 focuses on the corpus and the gold standard. Section 4 introduces the TExSIS architecture, whereas Sections 5 and 6 provide a detailed description of the TExSIS bilingual terminology extraction system. Section 7 discusses the experimental results. Section 8 ends with some concluding remarks and directions for future research.

2. Background and related work

Two different theoretical viewpoints regarding the acquisition of terminology exist. While the early approaches to terminology were onomasiological (starting from the concepts and working towards the terms), the more recent corpus-driven approaches are per definition semasiological (starting from the terms and working towards the concepts). The different theories of terminology are described in Cabré Castellví (2003) and Bowker (2008). Wright (1997) views the process of terminology management as an iterative one in which both the semasiological and onomasiological approach interact.

Terminology extraction can be seen as an important step of a larger process of corpus compilation, terminology extraction and terminology management (Gamper & Stock, 1999). In the terminology extraction phase, terms are identified in a text and - in the case of multilingual terminology extraction - the corresponding translations are retrieved. The extracted terms and their translations can be stored in bilingual glossaries, which are already a valuable aid for technical translators. If the aim is the creation of a term bank, the extracted terms are structured in concept-oriented databases in the terminology management phase. Each database entry represents a concept and contains all extracted term variants (including synonyms and acronyms) in several languages. In most cases, the terminology extraction tool generates lists of candidate terms, which are then verified by human experts.

Basically, there are two methodologically different approaches to terminology extraction, viz. the linguistic and the statistical approach. The linguistic approach is

based on the characteristics of term formation patterns, which are expressed as a part-of-speech code or sequences of part-of-speech codes (e.g. N N, N prep N, Adj N). Term formation patterns for English can be found in Justeson and Katz (1995) and Quin (1997); patterns for French in Daille (1996). In order to determine the morpho-syntactic patterns, linguistically-based systems apply language-specific part-of-speech taggers. As a result, linguistically-based terminology extraction programs are always language-dependent. The statistical approach on the other hand is language-independent and is based on quantifiable characteristics of terms. One such characteristic is that terms tend to occur more frequently in specialized texts than in general domain texts (termhood) Another characteristic is that multi-word terms exhibit a high degree of cohesiveness (unithood). The statistical approaches use several statistical measures such as frequency, association scores, diversity and distance metrics (Daille, 1996). Fulford (2001, p. 261) pointed out that "terms do not tend to possess linguistic features that distinguish them clearly and decisively from non-terms". Hence, we can expect that linguistically-based approaches tend to overgenerate. Also the statistical approaches tend to produce some noise. Moreover, frequency-based systems will not be able to detect newly-coined terms with low frequencies. Hence, most state-of-the-art systems use hybrid approaches that combine linguistic and statistical information. Different methods and systems are described and compared in Kageura and Umino (1996), Cabré Castellví et al. (2001), and Zhang et al. (2008).

Bilingual term extraction is faced with the additional problem of finding translation equivalents in parallel texts. There is a long tradition of research into bilingual terminology extraction (Gaussier, 1998; Kupiec, 1993). In most systems, candidate terms are first identified monolingually. In a second step, the translation candidates are extracted from the bilingual corpus on the basis of word alignments or co-occurrence information. In recent work, Itagaki et al. (2007) use the phrase table derived from the GIZA++ alignments to identify the translations, whereas Vintar (2010) uses a bilingual lexicon in her bag-of-equivalents approach.

We present an alternative approach that takes a multilingual perspective from the start and generates candidate terms directly from linguistically motivated aligned chunks. In a second step, we use different statistical measures to determine the term specificity.

Our approach is related to that of Tiedemann (2001) who uses the bilingual word alignment to improve the precision of the terms extracted in the monolingual term extraction phase. The main difference between his work and ours, however, is that Tiedemann starts from monolingual term extraction, while we take as a starting point the aligned linguistically motivated phrases.

3. Corpus construction

The bilingual term extraction module described here has been carried out in the framework of a terminology management project for a major French automotive company. The final goal of the project was a reduction and terminological unification process of the company's database, which contains all text strings that are used for compiling user manuals. French being the source language, all French entries had been translated to some extent into the twenty different languages that are part of the customer's portfolio. Bilingual term extraction was the first step of the more extensive terminology management project. The French database contains about 400,000 entries (i.e. sentences and parts of sentences with an average length of 9 words), which are aligned across all languages by means of a unique ID. We used a part of this large parallel database for the construction of our reference corpus.

3.1 Reference Corpus

For the evaluation of the alignment and terminology extraction module, we created three parallel sub-corpora: French-Italian, French-English, and French-Dutch, each one containing in total 1594 sentence pairs. As we presume that sentence length has an impact on the alignment performance, and thus on term extraction, we created three test sets with varying sentence lengths. We distinguished short sentences containing 2-7 words (911 sentence pairs), medium-length sentences containing 8-19 words (456 sentence pairs) and long sentences containing over 19 words (227 sentence pairs). Each test corpus contains approximately 10,000 words.

In order to measure both precision (number of correct terms) as well as recall (number of retrieved terms) of our automated term extraction module, we created a gold standard corpus for the three test sets. Two linguists divided the annotation work and manually indicated all valid terms in the three language pairs.

We decided to construct a *maximum set* of terms by including all *possible technical terms*. This way both the *base terms* (e.g. "*seat belt*") as well as the *complex terms* (e.g. "*outer front seat belt*") are included in the gold standard. The example below shows the input and gold standard for the following English phrase: "*turbocharging air heating solenoid valve*". The annotators receive the input in the four considered languages (French, Italian, English and Dutch) and list the terms in the target language(s) they master, separated by "#".

Input:

électrovanne réchauffage air de suralimentation .

elettrovalvola di riscaldamento dell'aria di sovralimentazione .

elektroklep verwarming vuldruklucht .

turbocharging air heating solenoid valve .

Terms:

électrovanne réchauffage air de suralimentation#elektroklep verwarming vuldruklucht#turbocharging air heating solenoid valve#elettrovalvola di riscaldamento dell'aria di sovralimentazione

électrovanne#elektroklep#solenoid valve#elettrovalvola

réchauffage#verwarming#heating#riscaldamento

air de suralimentation#vuldruklucht#turbocharging air#aria di sovralimentazione

Next, we ran inter-annotator checks in order to measure the quality of the manual annotation. Table 1 lists the inter-annotator figures for a restricted portion of the French-Dutch, French-English, and French-Italian test corpora that has been labelled by both linguists: we compared 100 sentences for test set 1 and 50 sentences for both test set 2 and 3. We measured precision, recall and the harmonic mean F on the two manually created gold standards, taking one of the two annotations as the reference corpus. The inter-annotator agreement rates indicate that the annotators labelled the same terms in 76% of the cases. No major differences between the language pairs can be observed. As expected most disagreement was found on the longer sentences. However, the inter-annotator scores were sufficiently high to use the reference corpus as gold standard for evaluation purposes.

	Precision	Recall	F-measure
French-English			
Test set 1	85.6	83.4	84.5
Test set 2	77.3	72.4	74.8
Test set 3	76.3	67.9	71.9
All Test sets	79.8	74.2	76.9
French-Dutch			
Test set 1	84.1	81.9	83.0
Test set 2	80.2	75.1	77.6
Test set 3	76.5	68.7	72.4
All test sets	79.6	73.8	76.6
French-Italian			
Test set 1	84.4	82.3	83.4
Test set 2	80.7	75.6	78.0
Test set 3	75.9	67.7	71.6
All test sets	79.8	74.0	76.8

Table 1: Inter-annotator agreement results

3.2. Term characteristics

In order to gain an insight into the composition of the terms in our reference corpus, we performed an analysis of the terms that occur in the three parallel corpora.

As the linguistic approach to automatic term extraction is based on the characteristics of term formation patterns, we were very interested in the distribution of the different part-of-speech patterns in the manually created term base. We PoS-tagged the French, English and Italian corpora with Treetagger (Schmid, 1994) and the Dutch corpus with TADPOLE (van den Bosch et al., 2007) and measured the number of part-of-speech pattern correspondences on the three bilingual gold standard term lists. We sorted the PoS pattern correspondences in descending frequency order of the French PoS patterns. It must be noticed however that our statistics also include wrongly tagged terms (e.g. English terms ending in "-*ing*" are often tagged as verbs when it concerns adjectives or even nouns in reality). Table 2 shows the most frequent corresponding patterns for the six most productive French part-of-speech patterns.

French	Italian	English	Dutch
Noun	Noun (2961) Verb (202) Adjective (163)	Noun (2492) Noun noun (280) Verb –ing form (255)	Noun (3142) Adjective (91) Adjective noun (73)
Noun preposition noun	Noun preposition noun (137) Noun (119) Noun determiner-preposition noun (68)	Noun (151) Noun noun (110) Adjective noun (29)	Noun (289) Noun preposition determiner noun (57) Adjective noun (24)
Noun adjective	Noun (190) Noun adjective (142) Adjective (26)	Noun (204) Adjective noun (145) Noun noun (40)	Noun (277) Adjective noun (154) Noun preposition determiner noun (32)
Adjective	Noun (256) Adjective (217) Verb (18)	Adjective (243) Noun (164) Noun noun (41)	Adjective (240) Noun (237) Adjective noun (79)
Infinitive verb	Infinitive verb (276) Noun adjective (13) Adjective (10)	verb (108) Noun (97) Verb –ing form (19)	Noun (139) Verb present tense (64) Infinitive verb (16)
Noun noun	Noun noun (20) Noun (18) Noun preposition Verb (12)	Noun (36) Noun noun (28) Noun noun noun (11)	Noun (70) Adjective (19) Noun “and” determiner noun (5)

Table 2: Most frequent part-of-speech (PoS) pattern correspondes between French and the other three languages

The frequencies of the different part-of-speech pattern correspondences lead to some interesting observations. The “noun” pattern and the “noun-adjective” (or “adjective-noun”) or “noun-noun” variations are the most frequently occurring part-of-speech patterns and account for more than fifty percent of all terms in all languages, and even for 67 percent of all terms in Dutch. This observation corresponds to the findings of previous research that focused on defining valid grammatical sequences to perform automated term extraction (Daille, 2000; Justeson & Katz, 1995; Quin, 1997). In addition, the difference in compounding strategy for the different languages is reflected in the most frequently occurring part-of-speech patterns. E.g. the English “noun-noun(-noun)” terms often correspond to “noun-preposition-noun” terms in the romance languages (French and Italian), and to “noun” and “noun-preposition-determiner-noun” terms in Dutch. Generally speaking, we notice that the most frequent part-of-speech patterns in the romance languages often correspond to each other.

Apart from the part-of-speech sequences, we also measured some other characteristics of the reference term set, being the percentage of terms containing verbal elements, the proportion of single-word and multi-word terms, and the

proportion of nested terms (terms being part of a bigger term). Table 3 gives an overview of these percentages, measured on the total number of terms in the monolingual reference sets.

	French	**Italian**	**English**	**Dutch**
Number of terms in monolingual reference	3516	3543	3452	3565
Terms containing verbal element	20.1%	20.8%	27.1%	10.9%
Terms that are verbs	8.0%	8.3%	9.2%	4.8%
Multi-word terms	69%	70%	72%	51%
Single-word terms	31%	30%	28%	49%
Nested terms	48%	49%	50%	45%

Table 3: Additional statistics on the nature of the reference terms in the four considered languages

The discrepancies between the number of multi-word terms in Dutch and the other three languages can be explained by the different compounding strategy that is adopted in Dutch. In Dutch the various compound parts are glued together in order to form one orthographic unit, whereas in the other languages these parts are separated by spaces (although often linked by prepositions in the romance languages).

4. The TExSIS architecture

Figure 1 gives an overview of the TExSIS architecture.

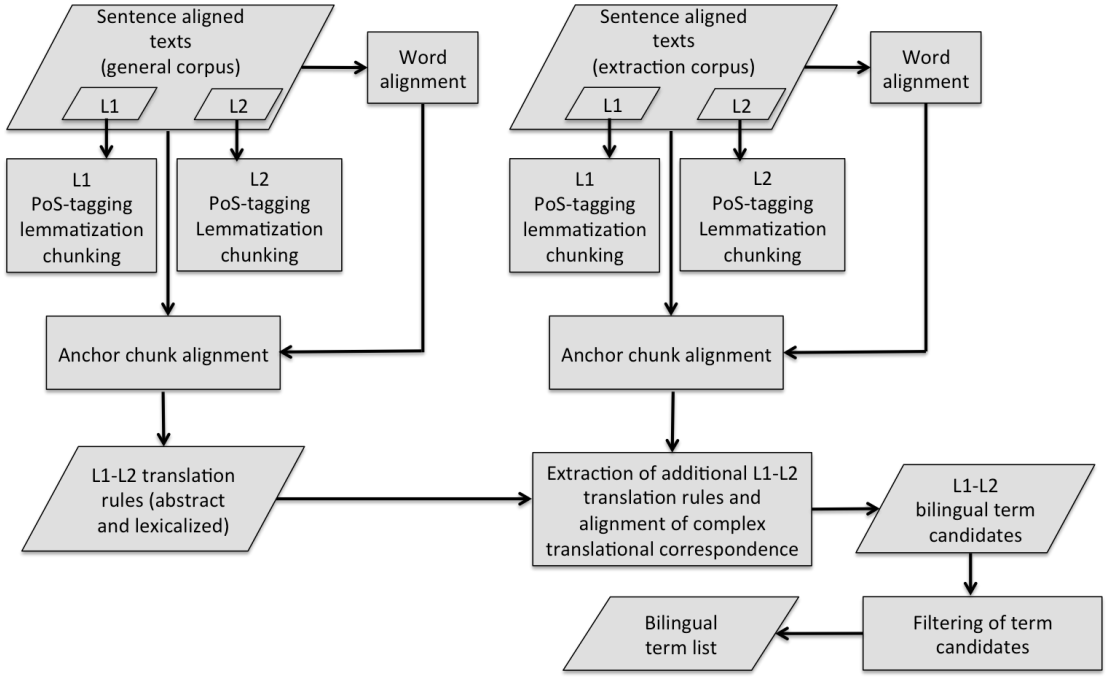


Figure 1: TExSIS architecture

The TExSIS system works directly on sentence-aligned parallel text, on which word alignment is performed. Furthermore, a shallow linguistic preprocessing (part-of-speech tagging, lemmatization and chunking) of both L1 and L2 is carried out.

The core of TExSIS is a two-step sub-sentential alignment system (see Section 5 for a thorough discussion) that links linguistically motivated phrases in parallel texts. The motivation behind this is that, in order to accurately describe specialist knowledge, technical domains abundantly make use of technical terms, which are not only linguistically realized as single-word terms but also as multi-word units. Nakagawa et al. (2002), for example, report in their experiments that 85% of the domain specific terms are compound nouns. The Genia[5] corpus counts 97.8K annotated terms, with more than half of them being composed of more than one word. TExSIS captures this abundance of both single-word and multi-word units by directly exploiting the linguistic correspondences between source and target language at a level higher than the word level. The alignment system is chunk-driven and requires only shallow linguistic processing tools for the source and the target languages, i.e. part-of-speech taggers, lemmatizers and chunkers. The selection of the phrases or chunks is based on lexical correspondences and syntactic similarity. We believe that our chunk-based approach offers a flexible alternative to the identification of candidate terms on the basis of predefined PoS patterns. Although part-of-speech tagging is indeed used as a preprocessing step for the chunker, no language-dependent predefined patterns were made to select candidate multi-word terms, making the approach easily portable to other languages.

After alignment, the terminology extraction module again follows a two-step approach, in which in a first step, candidate terms are generated on the basis of the aligned phrases. In a second step, we combine several statistical filters to determine the specificity of the candidate terms. For a detailed description of this terminology extraction module, we refer to Section 6.

5. Chunk-based alignment

As stated earlier, the central component in the TExSIS system is a chunk-based subsentential aligner. We conceive our subsentential aligner as a cascaded model consisting of two phases:

- In the first phase, anchor chunks, viz. chunks that can be linked with a very high precision on the basis of lexical correspondences and syntactic similarity, are retrieved. The anchor chunks and the word alignments of the first phase are used to limit the search space in the second phase.
- In the second phase, we use a bootstrapping approach to extract language-pair specific translation rules.

In order to find high precision anchor chunks, the chunk-based aligner relies on two building blocks, namely (i) the retrieval of lexical correspondences through word alignment and (ii) the detection of syntactic similarity on the basis of shallow linguistic processing of both source and target language.

The lexical correspondences between both languages are determined by relying on IBM model 4 word alignments (Brown et al., 1993). For our experiments, we used the GIZA++ (Och & Ney, 2003) word alignment toolkit, which implements the IBM models (Brown et al., 1993) to generate the initial source-to-target and target-to-source word alignments. In GIZA++, three symmetrization heuristics are implemented to combine the alignments of both translation directions. Intersecting the two alignments results in an overall alignment with a high precision, while taking the union of the alignments results in an overall alignment with a high recall. The grow-diag-final symmetrization heuristic starts from the intersection points and gradually adds alignment points of the union to link unaligned words that neighbour established alignment points. A reported problem with the union and the grow-diag-final heuristics is that the gain in recall causes a substantial loss in precision, which poses a problem for applications intended for human users. We will investigate whether this is indeed the case for our chunk-based alignment system. As we assume that bilingual terminology extraction methods will benefit from high precision word alignments, we hypothesize that the intersection heuristic will be the most appropriate metric for our experiments.

As for the linguistic preprocessing, we PoS-tagged and lemmatized the French, English and Italian corpora and mapped the part-of-speech tag sets of the different languages and created rule-based chunkers (Macken, 2010b). During text chunking,

syntactically related consecutive words are combined into non-overlapping, non-recursive chunks on the basis of a fairly superficial analysis. The following example shows sentence pairs divided in non-overlapping and non-recursive chunks:

Fr: *Soulever impérativement* | *par les roues avant* | .

It: *Sollevare tassativamente* | *dalla ruote anteriori* | .

En: *It* | *is* | *imperative* | *that* | *the vehicle* | *is raised* | *by the front wheels* | .

Nl: *Til* | *de wagen* | *altijd* | *op* | *bij de voorwielen* | .

In the second phase, a bootstrapping approach is used to extract language-pair specific translation rules. The bootstrapping process is a cyclic process that alternates between extracting candidate translation rules (extraction step) and scoring and filtering the extracted candidate translation rules (validation step). In the extraction step, candidate translation rules are extracted from source and target chunks that have not been linked during the anchor chunk alignment phase. In the validation step we use the Log-Likelihood Ratio (see Section 6 for a discussion on the metric) as statistical association measure to compute an association score between each source and target pattern of all candidate translation rules. Only those translation rules with a Log-Likelihood value higher than a predefined threshold are retained. More details on the bootstrapping approach can be found in Macken and Daelemans (2010).

In order to compile a basic set of language-pair specific rules, we extracted three sub-corpora of short sentences (up to ten words)[6] from Europarl (Koehn, 2005). We selected Europarl as a general corpus as it contains the four languages under consideration. The size of the sub-corpora is given in Table 4. Macken (2010b, p. 92) demonstrated that the largest improvement in alignment could be observed after the first bootstrapping cycle. To reduce the memory requirements of our system, we opted for one bootstrapping cycle (one extraction and one validation step) on the Europarl corpus. The basic set of language-pair specific rules is learned offline. To capture corpus-specific structures or lexical correspondences, an additional bootstrapping cycle is performed at run-time on the extraction corpus prior to term extraction (see Figure 1).

	# words
Fr-It	2,913,573
Fr-En	3,504,576
Fr-Nl	3,351,378

Table 4: Size of the sub-corpora of short sentences extracted from Europarl expressed in number of words

Two types of language-pair specific rules are extracted: abstract rules and lexicalized rules:

- Abstract rules are coded as part-of-speech sequences and capture frequent language-pair specific translation patterns, such as the deletion or insertion of function words, different structures of noun phrases and so on, e.g.
  - (1) Fr-It: PREP+DET+N_1 ⇒ PREP-det+N_1 (e.g. *de la zone* ⇒ *della zona* (En: *of the zone*))
  - (2) Fr-It: N_1+ADJ_2 ⇒ N_1 | PREP-det+N_2 (e.g. *débrayage compresseur* ⇒ *disinnesto del compressore* (En: *disengaging of the compressor*))
  - (3) Fr-En: DET+N_1 ⇒ N_1 (e.g. *une simplification* ⇒ *simplification*)
  - (4) Fr-En: N_1 | PREP+N_2+ADJ_3 ⇒ ADJ_3+N_2+N_1 (e.g. *stratégie de développement durable* ⇒ *sustainable development strategy*)
  - (5) Fr-Nl: N_1+ADJ_2 ⇒ ADJ_2+N_1 (e.g. *tension croissante* ⇒ *stijgende spanning (*En: *increasing tension)*)
  - (6) Fr-Nl: V-fin_1+V-$papa_2$ ⇒ V-fin_1 … V-$papa_2$ (e.g. *est realisé* ⇒ *wordt ... verwezenlijkt* (En: *is performed*)
- Lexicalized rules are coded as token sequences and capture more domain-specific sequences, e.g.
  - (7) Fr-It: *de direction assistée* ⇒ *del servosterzo* (En: *power steering*); *en respectant* ⇒ *rispettando* (En: *by respecting*)
  - (8) Fr-En: *adaptateur d' antenne* ⇒ *aerial adaptor*; *se trouve* ⇒ *is*
  - (9) Fr-Nl: *mise à température* ⇒ *op temperatuur brengen* (En: *bring to the correct temperature*); *déclippe* ⇒ *maak ... los* (En: *unclip*)

The extracted rules can be contiguous or discontiguous (see for example rules 6 and 9).

In order to evaluate the alignment quality, all translational correspondences were manually indicated in the three test corpora (see section 3.1) by one human annotator. We adapted the annotation guidelines of Macken (2010a) to other language pairs, and used three different types of links: regular links for straightforward correspondences, fuzzy links for translation-specific shifts of various kinds, and null links for words for which no correspondence could be indicated. Figure 2 shows an example.

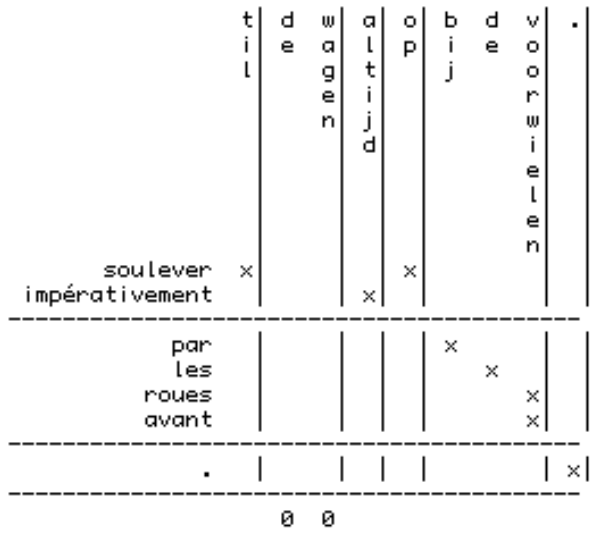


Figure 2: Manual reference for a French-Dutch sentence pair: regular links are indicated by x's, fuzzy links and null links by 0's. The horizontal and vertical lines indicate the chunk boundaries.

To evaluate the system's performance, we used the evaluation methodology of Och and Ney (2003), who introduced the following redefined precision and recall measures,

$$precision = \frac{|A \cap P|}{|A|},\ recall = \frac{|A \cap S|}{|S|}$$

and the alignment error rate:

$$AER(S,P,A) = 1 - \frac{|A \cap P| + |A \cap S|}{|A| + |S|}$$

in which **S** refers to sure alignments, **P** to possible alignments and **A** to the set of alignments generated by the system.

We consider all regular links of the manual reference as **sure** alignments and all fuzzy and null links as **possible** alignments to compare the output of our system with the manual reference. The results are presented in Table 5.

Overall, the results confirm our assumption that shorter sentences are easier to align than longer sentences. The results also show that the alignment quality is closely related to the similarity between languages. Italian and French are syntactically almost identical and hence easier to align. French and Dutch present a very different language structure: in Dutch, the compound parts are not separated by spaces (e.g. *voorwielen* (En*: front wheels*)); separable verbs occur frequently (e.g. *til...op* is the conjugated form of the verb *optillen* (En*: raise*)) and a different word order is adopted.

As expected, the intersection heuristic generates the most precise overall alignment, while the union results in an alignment with the highest recall. Especially for the less related language pairs, the recall gain in the union and grow-diag-final heuristics causes a substantial loss in precision. The chunk-based extension improves the recall of the intersected IBM Model 4 word alignments without sacrificing precision.

		Short			Medium			Long		
		Prec.	Rec.	AER	Prec.	Rec.	AER	Prec.	Rec.	AER
Ft-It	Gdf	97.7	95.6	3.4	95.7	92.7	5.9	90.4	90.0	9.8
	Union	97.3	**95.9**	3.4	95.2	**93.4**	5.7	89.0	**90.4**	10.3
	Intersection	**99.6**	89.3	5.8	**99.2**	84.2	8.9	**96.5**	78.6	13.4
	Chunk-based	99.5	90.1	5.4	99.1	86.5	7.6	96.2	81.4	11.8
Fr-En	Gdf	88.6	88.6	11.4	83.9	83.6	16.3	78.8	79.0	21.1
	Union	87.1	**90.0**	11.5	82.1	**85.0**	16.5	76.3	**80.5**	21.7
	Intersection	**97.8**	76.5	14.2	**96.5**	69.1	19.5	**95.8**	63.0	24.0
	Chunk-based	97.7	77.7	13.5	96.4	70.7	18.4	95.5	65.7	22.1
Fr-Nl	Gdf	85.0	82.1	16.5	78.8	76.9	22.2	64.7	65.1	35.1
	Union	83.8	**83.1**	16.5	76.4	**78.8**	22.4	61.5	**66.8**	36.0
	Intersection	**95.1**	65.7	22.3	**95.4**	57.6	28.1	**91.7**	45.4	39.2
	Chunk-based	94.8	68.4	20.5	94.2	61.5	25.6	91.0	49.2	36.1

Table 5: Precision (Prec.), recall (Rec.) and alignment error rate (AER) for the three symmetrization heuristics on the GIZA++ word alignments and the chunk-based alignment system (highest precision and recall scores are indicated in bold)

6. Terminology extraction

Our terminology extraction module follows a two-step approach:

- In a first step candidate terms are generated on the basis of the aligned phrases.
- In a second step, we combine several statistical filters to determine the **specificity of the candidate terms**.

All aligned phrases other than verb phrases are considered to be term candidates. In order to cover not only complex terms but also base terms, two heuristics are used to generate additional candidate terms. A first heuristic strips off adjectives and a second one considers consecutive NP + PP pairs as additional candidate terms.

Following this approach, six candidate terms are generated for the following sentence pair:

> Fr: *Dégager le mécanisme de lève-vitre de la porte en prenant soin d'extraire sans forcer l'entretoise de réglage de son support.*
> En: *Detach the window mechanism from the door taking care to extract the adjustment spacer from its support without forcing it.*
>
> [1] *mécanisme de lève-vitre de la porte#window mechanism from the door*
> [2] *mécanisme de lève-vitre#window mechanism*
> [3] *porte#door*
> [4] *soin#care*
> [5] *entretoise de réglage# adjustment spacer*
> [6] *support#support*

To filter our candidate terms, we keep the following criteria in mind:

- Each entry in the extracted lexicon should refer to an object or action that is relevant for the domain. This criterion reflects the notion of **termhood** that is used to express "the degree to which a linguistic unit is related to domain-specific context" (Kageura & Umino, 1996, pp. 260-261)
- Multi-word terms should present a high degree of cohesiveness. This criterion reflects the notion of **unithood** that expresses the "degree of strength or

stability of syntagmatic combinations or collocations" (Kageura & Umino, 1996).

- All term pairs should be valid translation pairs. So translation quality is also taken into consideration.

Different statistical measures were used to determine termhood, unithood and translation quality: to measure the termhood criterion and filter out general vocabulary words, we applied Log-Likelihood filters on all single-word terms; to measure unithood we calculated C-value for all multi-word terms; to measure translation validity we used a new, yet very straightforward measure FreqRatio, which compares the frequencies of the extracted source and target term.

The **Log-Likelihood Ratio** (LLR) allows us to detect single-word terms that are *distinctive* enough to be kept in our bilingual lexicon (Daille, 1996). This metric considers word frequencies weighted over two different corpora (in our case a technical automotive corpus and Europarl (Koehn, 2005)) and assigns high Log-Likelihood values to words having much higher or lower frequencies than expected.

In the formula below, **N** corresponds to the number of words in the corpus, whereas the observed values **O** correspond to the real frequencies of a word in the corpus and the index **i** refers to the corpus used (extraction corpus and general reference corpus).

The formula for calculating both the expected values (E) and the Log-Likelihood have been described in detail by Rayson and Garside (2000):

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

The resulting Expected values are then used to calculate the Log-Likelihood:

$$-2\log\lambda = 2\sum_i O_i \log\left(\frac{O_i}{E_i}\right)$$

Generally speaking, the higher the Log-Likelihood value, the more significant the difference between the two frequency scores. According to Manning and Schütze (2003) $-2\log(\lambda)$ has a distribution similar to that of chi-square and can therefore be used for hypothesis testing using the statistical tables for the distribution of chi-square.

For a contingency table with two rows and two columns the critical value is 3.84 for the significance level of 0.05 (McEnery, Richard, & Tono, 2006). Therefore, during filtering, we only retain translation rules with a Log-Likelihood value higher than 3.84 if the candidate term is overused in the domain-specific corpus. The following example shows some candidate terms that are filtered out by applying the log-likelihood threshold:

- Nl: *aandacht, aantal, gebied*
- Fr: *attention, nombre, exemple*
- It: *settore, problemi, processo*
- En: *attention, country, progress*

C-value (Frantzi & Ananiadou, 1999) aims at handling the extraction of nested terms by examining the frequencies of a term used as part of a longer term. Although the measure is applicable to single word units, we only apply it to multi-word term candidates. The reasoning behind our decision is that the Log-Likelihood Ratio is already an appropriate filter for single word terms and that C-Value was specifically designed for multi-word term recognition.

The C-Value is computed as follows:

$$CValue(a) = \log_2|a| f(a) \qquad \textit{if a is not nested,}$$

$$CValue(a) = \log_2|a|\left(f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b) \right) \qquad \textit{otherwise}$$

where

a is the candidate term,

f(a) is the frequency of *a* in the corpus,

$\boldsymbol{T_a}$ is the set of candidate terms that contain *a,*

$\boldsymbol{P(T_a)}$ is the number of candidate terms containing *a,*

b is a candidate term that contains *a,*

f(b) is the frequency of *b* in the corpus.

We discarded all multi-word candidate terms with a C-value of 0, which means that the multi-word term only occurs in the corpus as part of a larger term. The following candidate terms are examples of terms that are discarded on the basis of their zero C-value:

- Fr: *piston huilés* is discarded as term because it only occurs in the domain corpus as part of *axes de piston huilés* (En: *lubricated gudgeon pins*)
- Nl: *zijdelingse steun* is discarded as term because it only occurs in the domain corpus as part of *beugels voor de zijdelingse steun* (En: *side remaining wires*)

Our terminology extraction system is intended for bilingual terminology extraction. As a consequence, only term pairs are considered. If a source or target term is discarded on the basis of its LLR or C-Value, the whole term pair is discarded. To assess translation validity, we designed a new metric, **FreqRatio**, which is the ratio of the frequency of the source term and the frequency of the target term. It is intended to filter out partial translations or less frequent translations. In our experiments, the FreqRatio threshold was heuristically set to 0.2, which means that the candidate source-target pair only accounts for 20% of the translations in the extraction corpus of either the source or the target term.

The following candidate term pairs were discarded on the basis of a FreqRatio value below the threshold:

- Fr-En: *cylindre#faulty cylinder* (partial translation; correct term pair is *cylindre défaillant#faulty cylinder*)
- Fr-It: *toit escamotable#tetto* (partial translation; correct term pair is *toit escamotable#tetto a scomparsa*)
- Fr-Nl: *insert structural#structureel inzetstuk* (infrequent translation, more frequent term pairs are *insert structural#verstevingingsplaat* or *insert structural#inzetstuk*).

In order to rank both single and multiword terms, we applied the term weighting measure of Vintar (2010), which is computed as:

$$W(a) = \frac{f_a^2}{n} \sum_{1}^{n} \left(\log \frac{f_{n,D}}{N_D} - \log \frac{f_{n,R}}{N_R} \right)$$

in which $\mathbf{f_a}$ is the absolute frequency of the candidate term in the (technical) extraction corpus, **n** is the number of words constituting the term, $\mathbf{f_{n,D}}$ and $\mathbf{f_{n,R}}$ are the

frequencies of each constituent word in the extraction and in the general reference corpus respectively and $\mathbf{N_D}$ and $\mathbf{N_R}$ are the sizes of these two corpora expressed in number of tokens.

7. Experimental results

The TExSIS terminology extraction module was tested on all sentences from the three test corpora. The output was compared to a monolingual and multilingual reference term list that was derived from the manually created gold standard. In addition, we made an exhaustive comparison between the output of our TExSIS terminology extraction system and the output of the LUIZ system (Vintar 2010), a state-of-the-art bilingual term recognition system that was developed for Slovene-English. The LUIZ system first uses morphosyntactic patterns and statistical ranking to extract domain-specific terms monolingually, and aligns in a second step translation equivalents in the two monolingual term lists using a bag-of-equivalents approach. This approach uses the Twente word aligner (Hiemstra, 1996) to obtain a translation lexicon with accompanying probability scores. The lexicon is used to select the best translation equivalent for the source term from the list of target candidate terms.

We implemented the LUIZ system for the three considered language pairs. The French morphosyntactic patterns were based on (Daille 1996), the English ones were inspired on the patterns listed in (Justeson and Katz 1995). For Italian, we made an adaption of the French patterns, whereas for Dutch we constructed a new list of valid term patterns.

To complete the evaluation, we also compared the output of the TExSIS system with the output of two commercial bilingual term extractors, being Similis[7] and SDL Multiterm Extract[8].

7.1 Monolingual terminology extraction

In order to set an upper bound for our bilingual terminology extraction system, we ran experiments for monolingual terminology extraction where word alignment errors were ruled out. For these experiments, we deployed our bilingual terminology extraction framework but used one particular language both as source and target language, and as a consequence no word alignment errors were percolated to the term

extraction step. As a result, the evaluation results give an insight into possible term extraction performance in the case of optimal alignment. We present the results of the TExSIS system, which starts from chunks, and of LUIZ, which starts from predefined part-of-speech patterns. Figure 3 presents the evaluation results for precision, while Figure 4 shows the recall results for both systems for all four considered languages. In order to make a sound comparison of both systems, we ranked all terms based on the term weighting measure of Vintar (2010), as explained in Section 6. This way, the "Top500" label in Figure 3 and 4 refers to the set of the 500 highest ranked terms, etc.

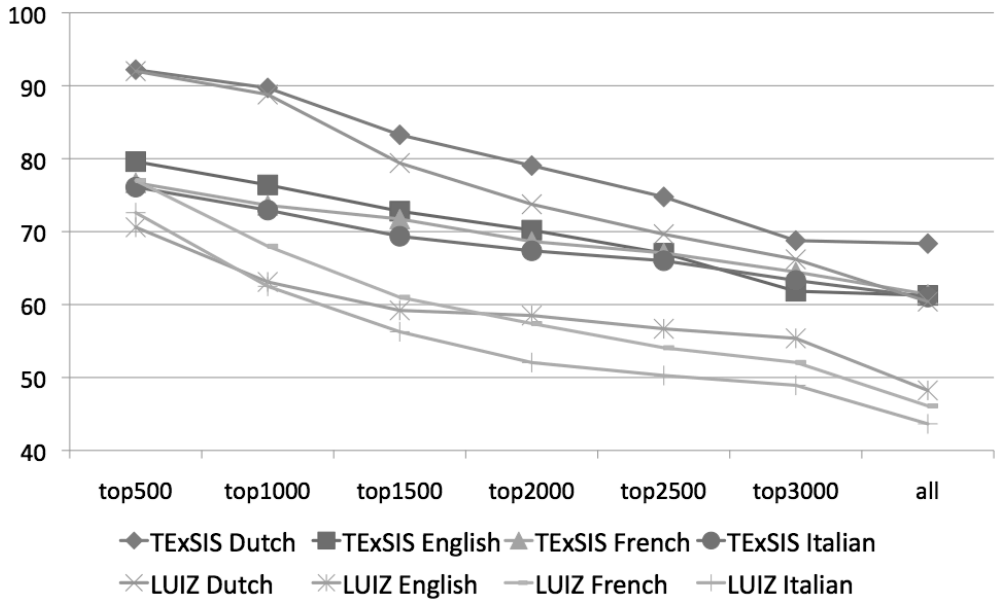


Figure 3: Precision results for the TExSIS and LUIZ systems for all four considered languages

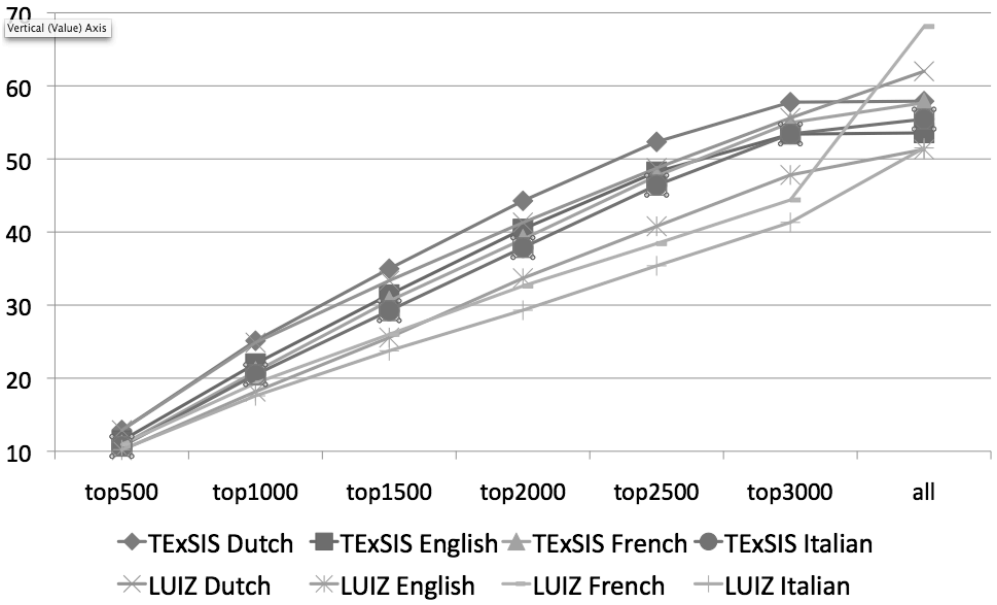


Figure 4: Recall results for the TExSIS and LUIZ systems for all four considered languages

The term weighting measure of Vintar ranks the terms as expected: precision decreases and recall increases when more (lower ranked) terms are considered. The figures also illustrate the importance of reporting both precision and recall scores when assessing terminology extraction systems. For example for Dutch, a precision score of 92% can be obtained for the 500 highest ranked terms, but this only accounts for 13% of the terms in the reference set.

A second observation is that the results reveal interesting differences between the languages. The best results are obtained for Dutch both in the TExSIS and in the LUIZ system. This can be attributed to the large proportion of single-word terms in the extraction corpus, which is due to the Dutch compounding strategy (see also Table 3). A shallow error analysis revealed that the lower recall figures for English and Italian can partly be ascribed to part-of-speech and chunking errors, e.g. *locking clips* is not extracted as a term because *locking* is tagged as a verb.

It can furthermore be noticed that the TExSIS system outperforms the LUIZ system for all considered languages, both for precision and recall. This can be explained by the fact that the LUIZ system does not perform any filtering on the ranked terms. As the TExSIS system successfully filters erroneous (parts of) terms, it obtains much higher precision scores for the monolingual term extraction task. Especially for English and Italian, the C-value filter seems to have a very positive

effect on the accuracy of the term extraction, e.g. the C-value filter correctly discards the English partial term *roof ECU* (should be *retractable roof ECU*).

7.2. Bilingual terminology extraction

7.2.1. Different flavours of TExSIS

In order to assess the effect of the underlying alignments on the quality of our bilingual terminology extraction system, we built four systems, viz. three TExSIS systems using the three main word alignment heuristics (grow-diag-final, union and intersection) and one TExSIS system using the chunk-based extension to the intersected GIZA++ word alignments.

TExSIS	**French-Italian**			**French-English**			**French-Dutch**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Gdf	61.10	38.85	47.50	54.10	20.87	30.12	45.96	17.61	25.46
Union	61.17	38.85	47.52	52.96	20.99	30.07	45.47	17.49	25.28
Intersection	**63.49**	37.57	47.21	66.40	17.19	27.32	60.75	14.87	23.89
Chunk-alignment	61.95	**42.12**	**50.15**	**66.55**	**25.23**	**36.59**	**62.60**	**24.57**	**35.29**

Table 7: Precision (P), recall (R) and F-measure (F) scores representing the effect of four different alignment procedures on bilingual terminology extraction for the three language pairs on all extracted term pairs.

Table 7 lists the precision, recall and F-measure scores for the four different flavours of the TExSIS system. As can be observed in the results, the chunk-based TExSIS system outperforms the other systems for the three language pairs. The only exception is French-Italian, where the intersection alignment metric is more precise. This alignment metric, however, suffers from lower recall figures, which finally results in a lower overall F-measure score. Another observation is that there are huge differences in terms of F-measure among the different language pairs and that these differences can be attributed to alignment quality. So, the more precise the word alignment, the better the terminology extraction results. The best overall results are obtained for French-Italian, two closely related languages, which are easier to align (see Table 5). French-Dutch, on the other hand, is the most difficult language pair to align. This might be attributed to the fact that GIZA++ is not well suited to model n:1 word alignments which frequently occur due to the different compounding strategy in

both languages. French-English and French-Dutch profit most from the chunk-based alignment extension to GIZA++ with large gains in recall.

An interesting observation is that the precision scores for French-English and French-Italian are higher than the precision scores obtained for the monolingual systems for French, Italian and English, which were 61.4, 61.2 and 61.3 respectively. This suggests that multilingual evidence can help to determine unithood. The highest precision scores are obtained for French-English, two languages that use different term formation patterns.

7.2.2. Impact of filtering

In order to assess the impact of the different filtering techniques, Table 8 presents an overview of the number of term pairs in the gold standard, and the number of terms before and after filtering in the TExSIS system for each of the three language pairs. Table 9 presents the impact of each statistical filter (Log-Likelihood Ratio, C-Value and FreqRatio) on the precision, recall and f-measure scores .

A first observation is that the number of French-Italian candidate terms before filtering is much higher compared to the number of French-English and French-Dutch candidate terms, which eventually translates into higher recall figures.

We see two reasons for this. Firstly, as the TExSIS terminology extraction system is built upon aligned chunks, the higher French-Italian alignment quality results in a higher number of term candidates. Secondly, as “noun + prep + noun” is a frequent term formation pattern in both French and Italian, different “NP + PP (+ PP)” combinations are considered as term candidates. While this strategy might generate too many term candidates, the high percentage of French-Italian term candidates filtered by C-Value seems to suggest that C-Value is a suitable filter for multi-word terms.

No huge differences in the impact of Log-Likelihood Ratio and FreqRatio can be observed in Table 9 for the different language pairs. This seems to suggest that the statistical filters are suited for different language pairs.

	French-Italian	French-English	French-Dutch
Term pairs in gold standard	3896	3931	4242
Candidate term pairs before filtering	3356	1681	1849
Term pairs after filtering	2644	1486	1663

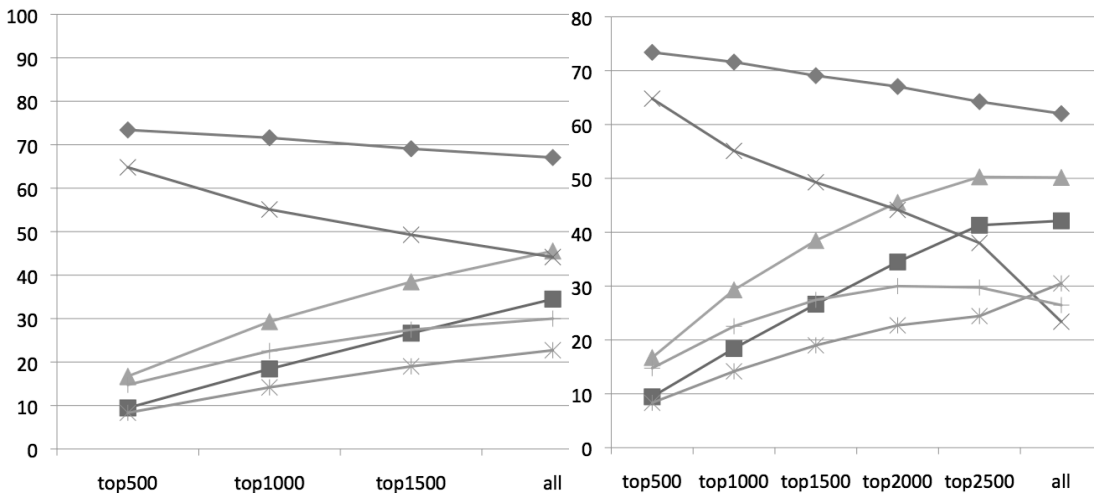
Table 8: Number of candidate term pairs filtered in comparison with the gold standard

	French-Italian			French-English			French-Dutch		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
No filtering	54.93	47.26	50.81	63.46	26.89	37.77	60.13	26.13	36.43
FreqRatio	55.27	46.90	50.74	63.94	26.73	37.70	60.76	25.91	36.33
C-Value	60.45	42.48	49.89	64.47	25.51	36.56	60.99	24.83	35.29
LLR	55.87	47.11	51.12	64.99	26.71	37.86	61.60	25.99	36.46
All filters	61.95	42.12	50.15	66.55	25.23	36.59	62.60	24.57	35.29

Table 9: Impact of the different filtering techniques on precision (P), recall (R) and F-measure scores (F) for the three language pairs on all extracted term pairs.

7.2.3. Comparison with other bilingual term extraction systems

We first present a comparison of the chunk-based TExSIS system with the LUIZ bilingual terminology extraction system (Vintar, 2010). In doing so, we compare two different approaches to bilingual terminology extraction, viz. the two-step approach that first identifies term candidates monolingually and in a second step aligns the source and target terms and the TExSIS approach that takes a multilingual perspective from the start.



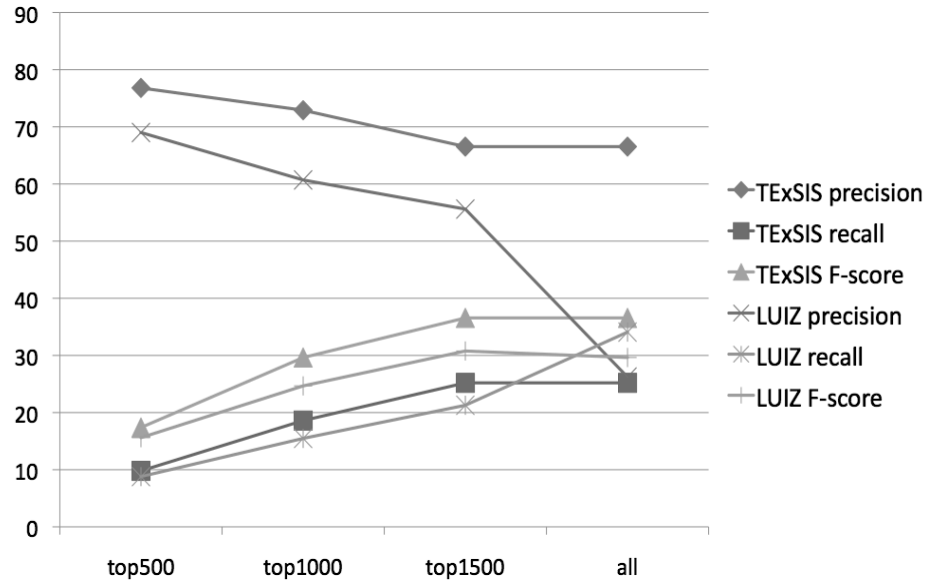


Figure 5: Precision, recall and F-scores for the TExSIS and LUIZ systems for French-Dutch (top left), French-Italian (top right) and French-English (bottom).

The results show that TExSIS outperforms the LUIZ system for the three language pairs. Moreover, the difference in precision between the TExSIS and the LUIZ approach are much larger for bilingual than for monolingual term extraction. A plausible explanation is that the LUIZ system, which first constructs monolingual term lists, cannot recover in the alignment step from errors generated during monolingual term extraction.

We also include a comparison with two commercial bilingual term extractors: Similis and SDL Multiterm Extract.

Similis is a sub-sentential translation memory system with an integrated bilingual terminology extraction module. It is a linguistically enhanced translation memory in that it contains monolingual lexicons and chunkers to group words into phrases (Planas, 2005). No detailed description of the terminology extraction module is available, but as the underlying technology aligns source and target chunks, we assume that the approach of Similis is very similar to ours.

SDL Multiterm Extract is a statistically based system that first generates a list of candidate terms in the source language (being French in our case) and then looks for translations of these terms in the target language. We ran SDL Multiterm Extract with its default settings (default noise-silence threshold and stopword list) on all test sentences.

Table 10 lists the precision, recall and F-measure scores for these three systems. As can be observed in the results, our terminology extraction module outperforms the two other systems for the three language pairs. SDL Multiterm Extract does not contain any linguistic information apart from a stop word list, which can explain the low precision scores. The precision scores of Similis are competitive to the TExSIS scores. A deeper insight into the SIMILIS systems is required to better interpret the difference in recall scores between both systems.

	French-Italian			**French-English**			**French-Dutch**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
TExSIS	61.95	**42.12**	**50.15**	**66.55**	**25.23**	**36.59**	**62.60**	**24.57**	**35.29**
Similis	60.08	22.22	32.44	65.60	19.31	29.84	57.77	11.23	18.81
SDL	39.69	9.10	14.81	30.65	8.44	13.24	30.88	6.14	10.24

Table 10: Precision (P), recall (R) and F-measure (F) scores for bilingual terminology extraction for the three language pairs on all extracted term pairs.

As we also wanted to evaluate the impact of term frequency on recall, we grouped the terms in the gold standard in different frequency classes. Other authors already stressed the usefulness of low frequent terms. See for instance (Lardilleux, Lepage, & F., 2011), where rare words are used as a foundation in the design of a multilingual sub-sentential alignment method.

The results show that for all systems, term or term pair frequency has a clear impact on recall: the higher the term (pair) frequency, the higher the recall. This is illustrated by Figure 6, which shows the impact of term frequency on recall for French-Italian. In general, the TExSIS recall scores are higher than the recall scores of SDL Multiterm Extract and Similis, both on high-frequency and low-frequency terms. The added value of using linguistic information is clearly demonstrated in the graph. While SDL Multiterm Extract is able to retrieve around 50% of the high frequency terms, the performance drops drastically for low-frequency terms.

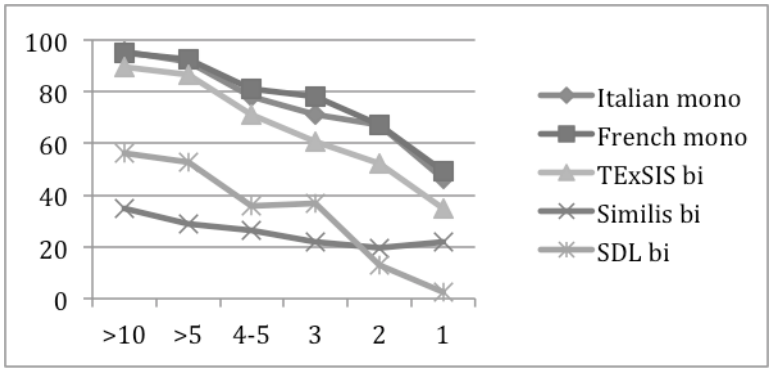


Figure 6: Impact of term frequency on recall for French-Italian

8. Summary and prospects for future research

In this paper, we reported on TExSIS, a bilingual terminology extraction system that extracts term pairs on the basis of aligned chunks. After alignment, candidate terms are generated from the aligned chunks after which the specificity of the candidate terms is determined by combining several statistical filters.

We developed a chunk-based alignment method that extends the statistical word alignments of the IBM Model 4 models with automatically extracted language-pair specific rules. The alignment is conceived as a two-phase approach in which the detection of high precision anchor chunks is followed by a cyclic bootstrapping process for the extraction and validation of candidate translation rules. We showed that the chunk-based extension improves the recall of the word alignments without sacrificing precision.

In order to validate the resulting terminology extraction module, we created three parallel data sets for French-Italian, French-English and French-Dutch, each containing ca. 1600 sentences and manually indicated all single-word and multi-word terms (of any length). We compared the TExSIS approach to bilingual terminology extraction with the more commonly used approach of first identifying term candidates monolingually and then aligning the source and target terms. Therefore, we implemented the approach described in (Vintar 2010) for the three language pairs considered. A comparison of the terminology extraction output with the monolingual and bilingual reference term lists derived from the manually created gold standard revealed that our system outperforms the state-of-the-art LUIZ system. We showed

that the precision of the alignment is crucial for the success of the terminology extraction. Based on the observation that the precision scores for the bilingual terminology extraction outperformed those of the monolingual systems, we concluded that multilingual evidence helps to determine unithood in less related languages.

Although the TExSIS terminology extractor compares favorably to other systems for the creation of monolingual and bilingual term lists, there are still numerous open issues. One major issue is the distinction between terms and non-terms. For our experiments, we chose to include both the base terms and complex terms in the gold standard term lists. But as far as we know there is no standardized evaluation framework for terminology extraction. As a first step towards such a framework, we will develop several benchmark data sets for different languages and domains. A major objective for future research is also to move beyond the flat term lists extracted by TExSIS and to mine the semantic relations between the extracted terms. In order to do so, we will investigate different statistical and linguistic semantic models and we will use cross-lingual evidence both from parallel and comparable data.

Acknowledgements

We would like to thank PSA Peugeot Citroën for funding this project.

Bibliography

Bowker, L. 2008. "Terminology". In Baker, M. & G. Saldanha (eds.), *Routledge Encyclopedia of Translation Studies*. 286-290. London, New York: Routledge.

Brown, P. F., V. J. Della Pietra, S. A. Della Pietra, & R. L. Mercer. 1993. "The Mathematics of Statistical Machine Translation: Parameter Estimation". *Computational Linguistics, 19*(2), 263-311.

Cabré Castellví, T. 2003. "Theories of terminology. Their description, prescription and explanation". *Terminology, 9*(2), 163-199.

Cabré Castellví, T., R. E. Bagot, & J. V. Palatresi. 2001. "Automatic term detection. A review of current systems.". In Bourigault, D., C. Jacquemin & M.-C. L'Homme (eds.), *Recent advances in computational terminology*. 149-166. Amsterdam: John Benjamins.

Daille, B. 1996. "Study and implementation of combined techniques for automatic extraction of terminology". In Klavans, J. L. & P. Resnik (eds.), *The balancing act: combining symbolic and statistical aproaches to language*. Massachusetts: MIT Press.

Daille, B. 2000. "Morphological rule induction for terminology acquisition.". In *Proceedings of the 18th International Conference on Computational Linguistics*. 215–221. San Francisco.

Frantzi, K., & S. Ananiadou. 1999. "The C-value / NC-value domain independent method for multi-word term extraction". *Journal of Natural Language Processing, 6*(3), 145-179.

Fulford, H. 2001. "Exploring terms and their linguistic environment. A domain-independent approach to automated term extraction". *Terminology, 7*(2), 259-279.

Gamper, J., & O. Stock. 1999. "Corpus-based terminology". *Terminology, 5*(2), 147-159.

Gaussier, E. 1998. "Flow Network Models for Word Alignment and Terminology Extraction from Bilingual Corpora". In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (Coling-ACL)*. 444-450. Université de Montréal, Montreal, Quebec, Canada.

Hiemstra, D. 1996. *Using statistical methods to create a bilingual dictionary.* University of Twente.

Itagaki, M., T. Aikawa, & X. He. 2007. "Automatic Validation of Terminology Consistency with Statistical Method". In *Proceedings of the Machine Translation Summit XI*. 269-274. Copenhagen, Denmark.

Justeson, K., & S. Katz. 1995. "Technical terminology: some linguistic properties and an algorithm for identification in text". *Natural Language Engineering, 1*(1), 9-27.

Kageura, K., & B. Umino. 1996. "Methods of automatic term recognition. A review". *Terminology, 3*(2), 259-289.

Koehn, P. 2005. "Europarl: a parallel corpus for statistical machine translation". In *Proceedings of the Tenth Machine Translation Summit*. 79-86. Phuket, Thailand.

Kupiec, J. 1993. "An algorithm for finding noun phrase correspondences in bilingual corpora". In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL)*. Columbus, Ohia, United States.

Lardilleux, A., Y. Lepage, & Y. F. 2011. "The Contribution of Low Frequencies to Multilingual Sub-sentential Alignment". *International Journal of Advanced Intelligence, 3*(2), 189-217.

Macken, L. 2010a. "An annotation scheme and Gold Standard for Dutch-English word alignment". In *Proceedings of the Seventh Language Resources and Evaluation Conference (LREC)*. Valletta, Malta.

Macken, L. 2010b. *Sub-sentential alignment of translational correspondences*. Brussels, Belgium: UPA University Press Antwerp.

Macken, L., & W. Daelemans. 2010. "A Chunk-Driven Bootstrapping Approach to Extracting Translation Patterns". In Gelbukh, A. (ed.), *Lecture Notes in Computer Science, 6008: Proceedings of the 11th International Conference on Intelligent Text Processing and Computational Linguistics*. 394-405. Berlin Heidelberg: Springer-Verlag.

Manning, C. D., & H. Schütze. 2003. *Foundations of Statistical Natural Language Processing*: Massachusetts Institute of Technology.

McEnery, T., X. Richard, & Y. Tono. 2006. *Corpus-based Language Studies. An advanced resource book*. London: Routledge.

Nakagawa, H., & M. Tatsunori. 2002. "A simple but powerful automatic term extraction method". *COLING-02 on COMPUTERM 2002: second international workshop on computational terminology - Volume 14*. 1-7. Stroudsburg, PA, USA: Association for Computational Linguistics.

Och, F. J., & H. Ney. 2003. "A systematic comparison of various statistical alignment models". *Computational Linguistics, 29*(1), 19-51.

Planas, E. 2005. "SIMILIS. Second-generation translation memory software". In *Proceedings of the 27th International Conference on Translating and the Computer (TC27)*. London, United Kingdom.

Quin, D. 1997. "Terminology for Machine Translation: a Study". *Machine Translation Review, 6*, 9-21.

Rayson, P., & R. Garside. 2000. "Comparing corpora using frequency profiling". In *Proceedings of the Workshop on Comparing corpora, 38th annual meeting of the Association for Computational Linguistics*. 1-6. Hong Kong, China.

Schmid, H. 1994. "Probabilistic part-of-speech tagging using decision trees". In *Proceedings of the International Conference on new methods in Language Processing*. Manchester, UK.

Tiedemann, J. 2001. "Can bilingual word alignment improve monolingual phrasal term extraction?". *Terminology, 7*(2), 199-215.

van den Bosch, A., B. Busser, W. Daelemans, & S. Canisius. 2007. "An efficient memory-based morphosyntactic tagger and parser for Dutch". In *Proceedings of the Computational Linguistics in the Netherlands 2006*. 191-206. Leuven, Belgium.

Vintar, S. 2010. "Bilingual term recognition revisited. The bag-of-equivalents term alignment approach". *Terminology, 16*(2), 141-158.

Vivaldi, J., & H. Rodriguez. 2007. "Evaluation of terms and term extraction systems. A practical approach". *Terminology, 13*(2), 225-248.

Wright, S. E. 1997. "Term selection: the initial phase of terminology management". In Wright, S. E. & G. Budin (eds.), *Handbook of terminology management*. 13-23. Amsterdam: John Bejamins.

Zhang, Z., J. Iria, C. Brewster, & F. Ciravegna. 2008. "A comparative evaluation of term recognition algorithms". In *Proceedings of the Sixth international conference of Language Resources and Evaluation (LREC)*. Marrakech, Morocco.

Notes

[1] http://iate.europa.eu
[2] http://www.btb.termiumplus.gc.ca
[3] http://www.eurotermbank.com
[4] http://www.termsciences.fr
[5] http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/
[6] We opted for short sentences, as they are more likely to contain only a few unlinked chunks.
[7] http://similis.org/linguaetmachina.www/index.php
[8] http://www.translationzone.com/en/translator-products/sdlmultitermextract/