[From: *Translating and the Computer 30*, November 2008 (London: Aslib)]

Beyond Terms: Multi-Word Units in MultiTerm Extract

María Fernández Parra & Pius ten Hacken, Swansea University
116435@swansea.ac.uk; p.ten-hacken@swansea.ac.uk

1. Introduction

Multi-word units are lexical units that are written as more than one word. They constitute a rather heterogeneous class, whose only unifying feature is that they represent a mismatch between orthographic representation and lexical units. Included in this class are syntactically governed combinations (e.g. *correspond with)*, complex prepositions (e.g. *in spite of)*, collocations (e.g. *put into practice)*, idioms (e.g. *have a bee in one's bonnet)*, etc.

Whereas multi-word units are linguistically heterogeneous, in translation they raise a very similar set of problems. In order to translate them, they first have to be recognized as belonging together. In the context of Computer-Assisted Translation (CAT), it is interesting to compare multi-word units with terms (cf. ten Hacken & Fernandez Parra 2008). The translation of terms with the help of a CAT tool involves two stages. First, terms have to be identified in a particular field and specified in the termbase. Then, terms have to be recognized in a particular source text (ST) for the termbase to suggest a translation. The same two-stage approach applies to the translation of multi-word units. This suggests that the translation of multi-word units could be supported by the machinery that CAT tools provide for terminology.

However, there are also important differences between terms and multi-word units. One relates to their morpho-syntactic behaviour. Most terms are nouns or complex nominals. Multi-word units include a much larger proportion of verb phrases. Therefore, multi-word units tend to display a larger degree of variability, making it more difficult for a CAT tool to recognize them in a text.

Another difference relates to the ontological status of terms. A term is a link between a form, a concept, and a field. It is therefore possible to collect terms for a particular field and activate them when a text from this field is translated. Multi-word units belong to the general expressive means of a language. Although some of them are marked for register or

text type, many are entirely unmarked. It is therefore not possible to collect a relatively small subset of multi-word units that are most likely to occur in a particular ST. No criteria comparable to the subject field for terminology can be used.

The latter difference has consequences for the identification phase. In terminology it is possible to build a termbase for a field on the basis of a corpus of texts from that field and the study of the subject matter. The translator can be reasonably confident that a new ST from the same field will contain few if any new terms. In the case of multi-word units, individual STs will generally have to be the basis for the identification. There will be too many new multi-word units to rely on an existing database to contain most of them. For this reason, the identification of multi-word units in the ST is much more important than the corresponding process for terms.

It is on the basis of these considerations that we decided to explore how term extraction software could be used for the identification of multi-word units. Here we will concentrate on MultiTerm Extract, the term extraction component of SDL Trados 2007. The questions we investigated were to what extent MultiTerm Extract supports the professional translator in the identification of multi-word units, what the optimal settings are in this context of use, and how the software might be further developed to improve its support.

2. The experiment

2.1. General considerations

In all our experiments we took chapter 10 of UNAIDS (2006) and its official Spanish translation as the material from which multi-word units were extracted. This chapter is entitled "Financing the response to AIDS" and is approximately 10,000 words long. In order to set the target for measuring the success of automatic identification, we started by searching multi-word units manually. We found 72, including *at risk, take into account*, and *close scrutiny*. In the course of our experiments, we found a further 18 multi-word units which had been missed in the manual search. The total of 90 was used as a standard for evaluating the success of automatic identification.

Term candidates identified by extraction software are always uninterrupted strings. As a consequence, it is not possible to get only the string *put to use* as a term candidate if it occurs in a context such as (la).

(1) a. That means streamlining the flow of financial resources to the front lines of the epidemic, putting it to optimal use and providing HIV-related prevention, treatment, care and support as quickly as possible to everyone in need.
b. putting it to optimal use
c. putting it to optimal use and providing HIV-related prevention

In evaluating the performance of extraction software, we counted any proposal in which the three components *of put, to,* and *use* occur and belong together as successful. The minimal string to be extracted from (la) would be (1b). However, we also accept (1c) as a case where *put to use* was recognized correctly.

MultiTerm Extract was designed to be used both for the identification of term candidates and for the verification of termbases. It supports the five types of project listed in (2).

(2) a. Monolingual Extraction
b. Bilingual Extraction
c. QA Project
d. Translation Project
e. Dictionary Compilation Project

For the identification of multi-word units, only (2a-b) are relevant. Whereas these two are meant to be performed for the collection of term candidates on the basis of texts, the projects in (2c-e) are used when a termbase has already been compiled.

2.2. Setting up a Monolingual Extraction project

In a Monolingual Extraction project, a number of settings can be used to influence the selection of term candidates. They are listed in (3). The dialog box in Fig. 1 illustrates the interface for specifying them.

(3) a. Minimum term length
b. Maximum term length
c. Maximum number of extracted terms
d. Silence/noise ratio
e. Stopword lists
f. Learning settings

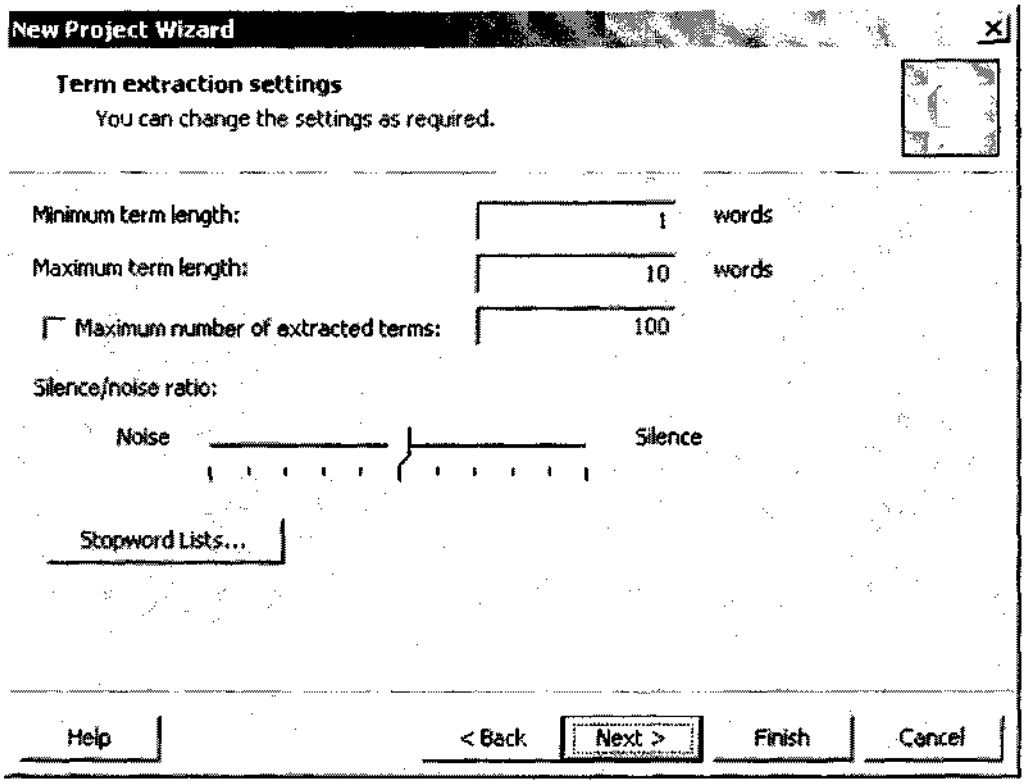

Fig. 1: Default term extraction settings.

The settings (3a-b) together determine the length of the term candidates returned. As we are interested in multi-word units, (3a) should be at least 2. The longest multi-word unit in the target set has a length of 4, so that (3b) should be at least 4. In view of the problem illustrated in the discussion of (1), we also considered a maximum term length of 10. In our experiments we used twelve different settings for (3a-b), three with maximal length 4 and minimal length ranging from 2 to 4 and nine with maximal length 10 and minimal length ranging from 2 to 10.

Setting (3c) can be used to restrict the number of term candidates. Assuming that this would simply remove the tail of the list without affecting the order of confidence for the individual items on the list, we did not use this feature in our experiments. Its optimal value might be an outcome of our experiments.

Setting (3d) is represented in Fig. 1 as a scale from maximal noise to maximal silence with a default setting in the middle. We will refer to them as *noise levels*, ranging from 0 (minimal noise) to 1 (minimal silence). The scale in Fig. 1 suggests nine intermediate points. In practice, however, we verified that there was no difference in output for noise

levels from 0.8 to 1 and from 0.6 to 0.7. As a consequence, there are eight different noise levels that can be chosen.

The option (3e) makes it possible to specify a stopword list. Normally, a stopword list contains high-frequency words that are not part of the target set. Considering that multi-word units such as *as of, all but,* and *by and large* contain typical stopwords, we first tried extraction without specifying a stopword list. We discovered, however, that using the default stopword list reduces noise considerably without excluding such multi-word units. Therefore, we used the default stopword list.

The final option, (3f), is not visible in the screen in Fig. 1, but can be specified elsewhere in the project setup. The underlying idea is that the system can learn from the evaluation by the terminologist of the items returned in extraction. In this way, future extraction projects can be more targeted. The way our experiments were set up makes it impractical to use this option. We compared a large number of settings for (3a-d) independently of each other for the same text. In order to evaluate the contribution to the efficiency by Learning Settings, we would have to consider a single setting over a number of texts. Therefore we did not explore this option.

2.3. Setting up a Bilingual Extraction project

Official texts on the website of the UNO (of which UNAIDS is a part) are published in five languages. We used the parallel English and Spanish version as a basis for Bilingual Extraction projects. The evaluation of the results can be based on the two questions in (4).

(4) a. How does the quality of the output for English compare to the quality of corresponding Monolingual Extraction projects?
b. What is the quality of the Spanish translations?

The settings for Bilingual Extraction projects include all settings for Monolingual Extraction, so that a point-by-point comparison for the evaluation of (4a) is possible. In addition, the three translation settings in (5) can be specified. The default settings are illustrated in Fig. 2.

(5) a. Search for new translations.
b. Maximum number of translations.
c. Minimum translation frequency.

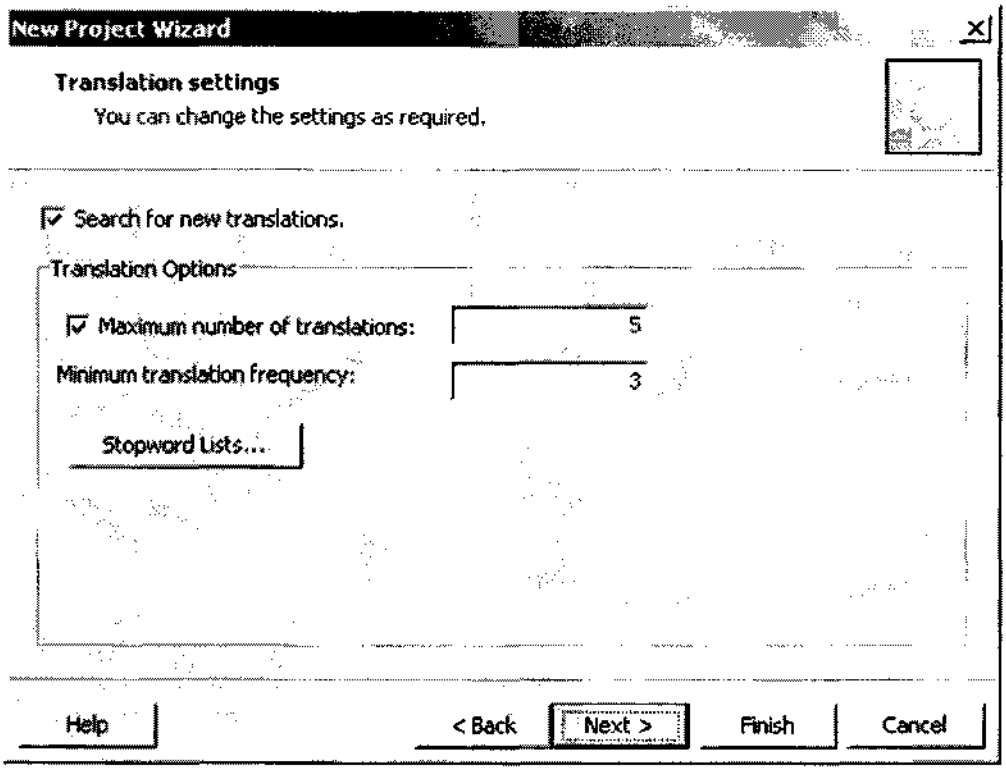

Fig. 2: Default translation settings for Bilingual Extraction project.

Option (5a) is a tickbox. Unticking it restricts the search to pairs of SL-TL expressions already recorded in an existing database. In our experiments, we do not have an existing database to start with, so that there is no sense in unticking it.

Option (5b) limits the number of different translations for a particular SL term candidate. In the case of genuine terms, this makes sense, because a small number of translations is typical. Rogers (2007), for instance, shows how the ideal of one expression for one concept is approximated very closely in the case of a concept in medical technology. If the target is the identification of multi-word units, no such assumption can be made. Therefore, we unticked this option.

Option (5c), finally, limits the number of SL-TL pairs given as output by imposing the condition that each pair has a minimum frequency. Again, this option is more appropriate for terms than for multi-word units. Terms in a particular domain are relatively frequent in a text from that domain and should normally be translated in the same way in each occurrence. Therefore, the default is set at 3. Multi-word units, however, are not necessarily specialized, so that no similar frequency effect can be expected. Therefore we changed the value to 1.

2.4. Processing the list of candidates returned

For each project, a number of term candidates is given as output in a format illustrated in Fig. 3.

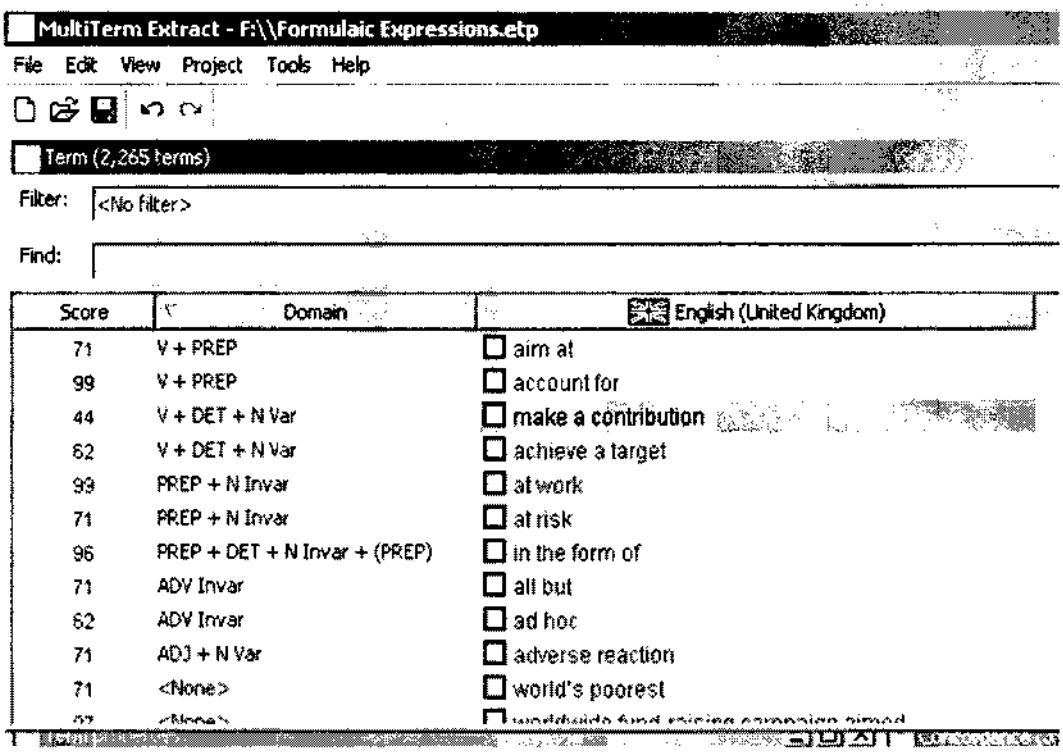

Fig. 3: Output of a term extraction project.

Fig. 3 shows the output of a monolingual project. For a bilingual project, an additional TL column is displayed. Each column can be used as a basis for sorting candidates. The column with the form of the candidates, headed by the language name, has tickboxes for each item. They can be used to verify an item (i.e. normally to confirm that it is a term) and to perform operations such as exporting selected items to a termbase.

The column labelled *Domain* gives project-specific information about the candidates. When the output is produced, all candidates have a value of *<None>*. In each project, the user can specify a list of possible values and then assign one of them to each accepted candidate. In our experiments, we used this column to specify syntactic classes of multi-word units. We can then use the column to sort the multi-word units.

The leftmost column in Fig. 3 shows the *Score.* This is a value between 1 and 99 indicating the system's confidence in proposing the term candidate. It is generated by the system and constitutes an important piece of information to be used in processing the list of results.

3. Overview of results

We carried out and evaluated a total of 192 experiments, systematically varying the term length settings (12 combinations as indicated in section 2.2 above) and the noise levels (8 genuinely different values as explained above) for monolingual and bilingual extraction projects. In presenting the results, we will start by considering these three factors in isolation. Then we turn to the evaluation of scores as a possible way to process the results more efficiently. Finally, we discuss how the type of multi-word unit influences retrievability.

3.1. The influence of noise level

The large number of experiments with multiple variables makes it difficult to present the results in a clear and systematic way. When we consider an individual parameter, such as noise level, there are at least two interesting questions we can ask, formulated in (6).

(6) a. What is the best setting for this parameter?
b. How good are the results?

The most straightforward question is (6a). Here we compare the different settings of a single parameter. Each setting corresponds to a class of experiments with different settings for other parameters. In our case, the 192 experiments are divided into 8 classes of 24 experiments each. In principle, we are only interested in the best result in each class.

The result of each experiment is a list of candidate terms. In (6b) we are interested in criteria to evaluate this list. Classical measures to evaluate such retrieval sets are precision and recall. There is an obvious tendency for precision to go down when recall goes up (and the reverse). Manning & Schütze (1999:269) propose an *F-measure* which combines the two in order to produce a single evaluation measure for the performance of a retrieval system. Fig. 4 gives a table with these three measures for the different noise levels.

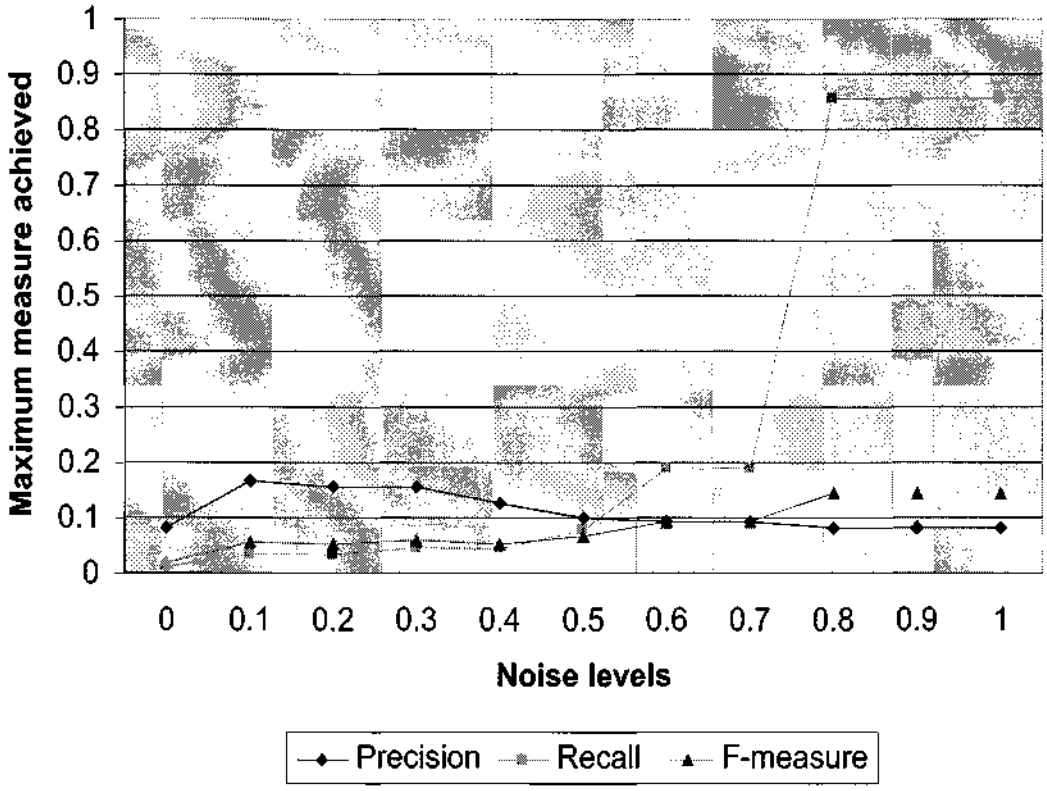

Fig. 4: Precision, recall, and F-measure for different noise levels.

As illustrated in Fig. 4 the noise levels of 0.6-0.7 and of 0.8-1.0 produce identical results. A first observation is that precision is overall very low. The highest precision is 16.7%, achieved with noise level 0.1 in a bilingual project with minimal term length 3 and maximal 4. This results in 2 multi-word units found in a list of 12 term candidates. Recall is also very low, except with high noise levels. The highest recall is achieved with noise levels 0.8-1.0 in monolingual projects with minimal term length 3 and maximal 10. This results in 77 multi-word units found in a list of 1498 term candidates. The F-measure generally follows recall, because there are much bigger differences in recall than in precision, but for high noise levels the value does not nearly go up to the same degree.

It is interesting to consider the practical implications of these results. Precision is particularly important if the entire set of term candidates returned has to be evaluated manually. In such a scenario finding even a small number of multi-word units by a relatively modest effort is not such a bad result. The alternative is to translate these items without special help. Optimal recall only has a practical relevance if additional methods can be found to manipulate the set of candidate terms in such a way that it is possible to find the multi-word units without going through the entire

set manually. In both cases it has to be kept in mind that items such as *put to use* have to be isolated from within a larger term candidate.

3.2. The influence of term length settings

The questions in (6) can also be asked for the term length settings, but this parameter is different from noise level in two respects. First, the minimal and maximal term length can be varied independently. Second, the set of possible values is not finite. The 12 combined settings we investigated are a sample. Each of these corresponds to a class of 16 experiments. Fig. 5 gives the best values for each of these classes.

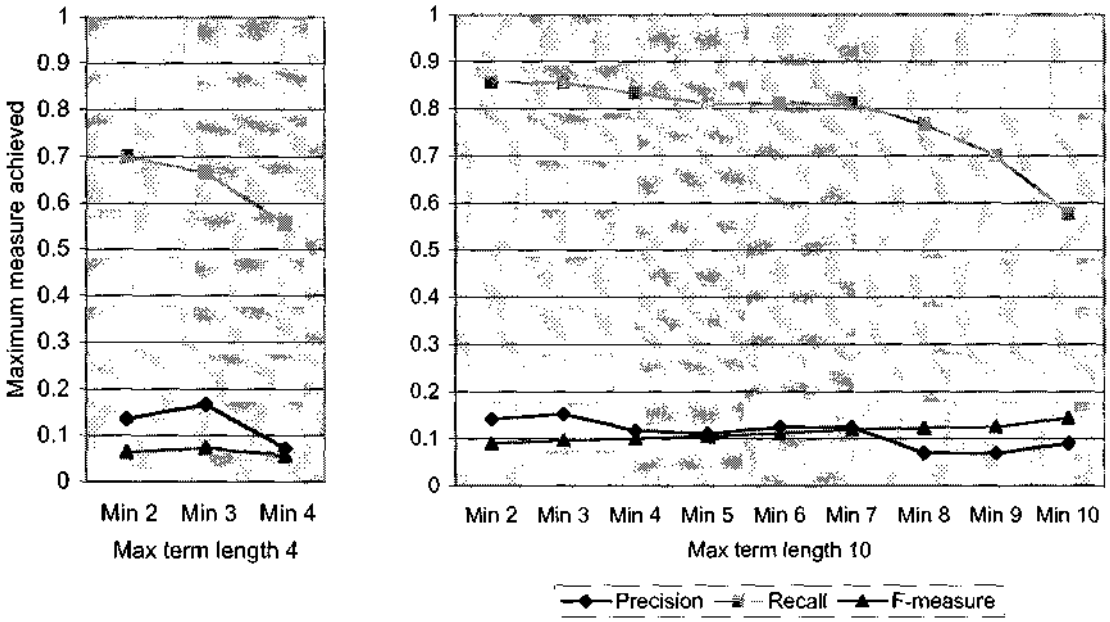

Fig. 5: Precision, recall, and F-measure for different term lengths

It is not surprising that precision and F-measure are low everywhere, because any higher value would have appeared in Fig. 4 as well. It is interesting, however, that with both maximal term lengths considered, the best precision is achieved with a minimal term length of 3. In absolute figures, this corresponds to 2 multi-word units found in 12 for 3-4 and in 13 for 3-10.

Recall measures were on the whole better with a maximum term length of 10. Of course, it is to be expected that increasing the minimum term length results in a lower recall. What is remarkable in Fig. 5 is that this drop in recall is very slow. For term lengths 2-10 to 7-10, recall drops from 85.6% to 81.1%. In absolute terms, this corresponds to 77 multi-word units found in 1,619 for 2-10 and 73 found in 1,125 for 7-10. By

raising the minimal term length from 2 to 7, we can eliminate over 30% of candidates while sacrificing only 5% of genuine multi-word units.

3.3. Monolingual vs. bilingual projects

Monolingual and bilingual projects can be compared as to the results they yield for English. In addition, bilingual projects can be evaluated for the quality of the translation into Spanish. These two modes correspond to the two questions formulated in (4) above.

We did not expect substantial differences in performance for English between monolingual and bilingual projects. The Spanish translation might in some cases help the system, but it might equally confuse it. In fact, when noise level and term length settings are kept constant, we found that results of monolingual and bilingual projects are identical or very similar. This is illustrated for a number of settings in Table 1.

Project type	Term length Min	Max	Noise level	Cand. terms	Expr. identified	Recall	Precision
Mono	10	10	1	613	51	0.5667	0.0832
Bi	10	10	1	646	52	0.5778	0.0805
Mono	3	4	0.3	25	3	0.0333	0.1200
Bi	3	4	0.3	24	3	0.0333	0.1250
Mono	2	10	0.1	21	3	0.0333	0.1429
Bi	2	10	0.1	21	3	0.0333	0.1429
Mono	2	4	0.8	2,482	63	0.7000	0.0254
Bi	2	4	0.8	2,544	62	0.6889	0.0244
Mono	8	10	0.6	26	1	0.0111	0.0385
Bi	8	10	0.6	72	5	0.0556	0.0694
Mono	2	4	0.7	351	12	0.1333	0.0342
Bi	2	4	0.7	444	7	10.1889	0.0383

Table 1: Comparison of monolingual retrieval in monolingual and bilingual projects with selected settings

We observed a general tendency for monolingual projects to score slightly better than bilingual projects with the same settings, but there were some exceptions, as illustrated in the last four lines of Table 1. As it is not possible to identify any trend when bilingual projects might score better, it is generally safer to use monolingual projects.

As indicated in section 2.3, bilingual projects exploit properties of terms that are not shared by multi-word units. It was therefore not a big surprise

to find that translations proposed for term candidates were generally of poor quality. Whereas terms tend to occur relatively frequently in a text in the relevant domain and tend to have a consistent translation, multi-word units belong to general language. Therefore the properties that would have produced better translations for terms could not be used in the context we investigated. A further complication for the automatic search of a translation was that term candidates were often not constituents, but stretches of text that contained the multi-word unit. Non-constituents are often untranslatable. Thus, we counted (7) as an instance of *meet a challenge,* but it is hardly possible to provide a translation of the entire term candidate.

(7) trying to meet that challenge have led to redesign

It is not surprising, then, that the system often does not give a translation at all. Where translations are proposed, they are often completely off the mark, as in (8).

(8) a. include the direct costs incurred in the delivery
b. derivados de las mejoras de los programas y la infraestructura ‘derived from the improvements of the programmes and the infrastructure’

The candidate in (8a) was counted as correctly identifying the multi-word unit *incur costs.* In proposing the translation in (8b), the system clearly selected an incorrect stretch of Spanish as corresponding to (8a). In view of the generally poor quality of translations, no detailed analysis was undertaken.

3.4. The interpretation of confidence scores

As shown in Fig. 3, each term candidate found by MultiTerm Extract is assigned a confidence score. Confidence scores were found to depend on the parameter settings. An example of the influence of different settings on the variation of scores for a single multi-word unit is given in Table 2. The first column in the table is a number added for ease of reference.

ID	Project type	Term length Min	Max	Noise level	Score	Candidate
1	Mono	3	4	0.4	99	Response to AIDS
2	Mono	4	4	0.4	89	expanded response to AIDS
3	Mono	4	4	0.5	99	expanded response to AIDS
4	Bi	4	4	0.5	54	global response to AIDS
5	Bi	4	10	0.5	48	global response to AIDS

Table 2: Some settings and scores for *response to.*

The example *response to,* taken as a basis in Table 2, is not the most prototypical multi-word unit, but it is one that is identified under a large range of settings. The phenomenon of different scores depending on the settings emerges quite clearly, because the settings in each row in the table have a minimal difference compared to the ones in the rows following and preceding it.

When a multi-word unit was found as part of different strings in one project, the score given in the table is the one of the string with the highest score returned. As there were obviously at least three different contexts in which *response to* occurs in our text, it is interesting to note that there were only two cases where *response to* was found in more than one context with the settings listed in Table 2. One concerns project 1. Here, *expanded response to AIDS* has a score of 89, as in project 2. It is obvious why *Response to AIDS* was not found in project 2, because it is excluded by the minimal term length setting. In project 3, *global response to AIDS* was assigned a score of 48, as in project 5. This suggests that much of the difference between monolingual and bilingual scores for *response to* are due to the non-identification of *expanded response to AIDS* in the latter.

The most interesting question about confidence scores is whether they can help in the identification of multi-word units. This question is most relevant when the set of term candidates returned is too large for efficient manual processing. We therefore analysed the set of 1498 term candidates returned for a monolingual project with term length 2-10 and noise level 1, the settings that give a maximal recall. Table 3 gives an overview of the numbers of candidate terms and multi-word units for each score.

Score	Candidates	MWUs	Precision	% of MWUs
99	22	1	4.55	1.30
98	2	0	0.00	0.00
92	1	0	0.00	0.00
89	4	0	0.00	0.00
88	4	0	0.00	0.00
82	1	0	0.00	0.00
78	3	0	0.00	0.00
75	600	49	8.17	63.64
73	281	12	4.27	15.58
71	125	8	6.40	10.39
68	99	4	4.04	5.19
65	80	0	0.00	0.00
61	74	0	0.00	0.00
55	101	1	0.99	1.30
46	101	2	1.98	2.60
	1,498	77		100%

Table 3: Distribution of term candidates and multi-word units over confidence scores.

Table 3 lists all observed confidence scores in the setting with highest recall. The second and third columns list the number of candidate terms for each score and the number of multi-word units among them. The fourth column gives the precision for each individual confidence score. The final column gives the distribution of multi-word units over the confidence scores in percentages.

If we consider using the confidence scores as a way of speeding up the identification of multi-word units, we might group the scores as in Table 4.

Score	Candidates	MWUs	% of Cand.	% of MWUs
78-99	37	1	2.5	1.3
68-75	1105	73	73.8	94.8
46-65	356	3	23.8	0.8

Table 4: Grouping of confidence score data.

It is of course obvious from Table 3 that the scores with the largest numbers of multi-word units are also the scores with the largest numbers of candidates. From Table 4 we can conclude, however, that by concentrating on a range in the middle, we can eliminate more than a

quarter of all candidate terms while missing out only 5% of multi-word units. This is certainly a worthwhile result. The problem is, of course, that the range of 68-75 was only determined after the analysis of the full set. This raises the question whether it is possible to find this range before such an analysis.

The only data available for finding such a range is the first and second columns in Table 3. The sequence in the second column suggests the hypothesis that the upper boundary of the ideal range is marked by a sudden and very large increase in the number of candidate terms. The lower boundary seems much harder to find. If we only consider the score of 75, we miss more than a third of the multi-word units, so that it is worth investigating the next categories. However, the difference in success between 68 and 65 is not predictable on the basis of the sequence in the second column.

Clearly, more research is needed to determine how confidence scores assigned by MultiTerm Extract correlate with the appearance of multi-word units. Although the effect represented in Table 4 is very obvious, we should be careful to extrapolate these findings on the basis of a single text. We found that for other term length constraints, the most productive score in monolingual projects ranged from 69 to 77.

3.5. *Types of multi-word unit*

Multi-word units can be classified in different ways. Fernandez Parra (2007) discusses some criteria found in the literature. For our purposes here, syntactic differences are the most relevant, because they are most likely to influence the performance of MultiTerm Extract. The syntactically most salient properties of different multi-word units are their syntactic categories (part of speech) and the type of variation they allow. In a first instance, we compared the properties of the 77 multi-word units identified by the most successful setting of MultiTerm Extract with the properties of the entire target set of 90 multi-word units. Table 4 gives an overview of the rate of retrieval by pattern.

Class		Example	Types	Identified
1	Adj + N Invar	*solid evidence, close scrutiny*	4	4
2	Adj + N Var	*sharp criticism, adverse reaction*	3	3
3	Adv Invar	*ad hoc, as of, all too*	7	7
4	Prep + Det + N Invar	*along the way, at the outset*	*5*	*2*
5	Prep + N Invar	*at risk, at work*	14	6
6	V + Det + N Var	*play a role, raise a concern*	24	23
7	V + N Invar	*take advantage, raise money*	3	3
8	V + Prep	*depend on, aim at*	9	9
9	V + Prep + N Invar	*take into account, take into consideration*	2	2
10	other	*large and small, incur costs*	19	18

Table 4: Successful identification by syntactic class.

In Table 4, all patterns for which only one multi-word unit appears in the target set have been grouped together as class 10. The conclusion that can be drawn from this table is that the main source of problems are the patterns with a preposition at the start, classes 4 and 5. Whereas less than half of multi-word units in these classes were identified (42%), all but two (97%) of the other classes were successfully identified. Even in the face of small numbers, this result is significant.

An unexpected result shown in Table 4 is that the variability of a multi-word unit does not have a large impact on its retrieval. The two problem classes do not have variable components and most of the non-identified multi-word units do not allow their components to be separated, e.g. *by contrast, in particular.*

A third criterion we considered, besides syntactic categories and variability, was the Mutual Information Index (I). This is a measure of how strongly the components of a multi-word unit collocate. Expressions such as *ad hoc* have a very high index (almost 33,000 in the British National Corpus), *adverse reaction* has I = 592, *sharp criticism* has I = 36, *red car* has I = 8.5. We calculated I for all two-word units in our set and found that all expressions with I > 50 were found. The item with the highest score that was not found is *by contrast* (I = 43). In the range 0 < I < 50, about a quarter of the expressions was not identified by MultiTerm Extract. Within this range, no further correlation between I and the probability of identification could be established.

4. Conclusion

In this paper we reported on experiments with the automatic identification of multi-word units by means of MultiTerm Extract. Before drawing any conclusions, we would like to emphasize two limitations in the scope of our work. First, MultiTerm Extract was not designed for the purpose we used it for, so that our conclusions should not be taken as an evaluation of the software. Second, we deliberately explored a large range of different settings. In order to keep the work manageable, we restricted the scope of our experiments to a single text of moderate length. Therefore, our experiments should be taken as a pilot study to be followed up by a study of the most promising settings with a larger amount of data.

In general, our experiments show a low degree of precision in the results returned by MultiTerm Extract. In principle, one could pursue two approaches. One is to maximize precision, even if at a low level. In this case, our experiments suggest that a noise level of 0.1 and a term length of 3-4 or 3-10 are optimal. They result in a short list of term candidates and although most multi-word units are not identified, a number are found with little subsequent effort. The more interesting approach, however, is to ignore precision in the original list of term candidates, maximize recall, and try to find methods for retaining relatively large numbers of multi-word units without going through the full list.

We found that with high noise settings (0.8 to 1) and term length 2-10 or 3-10, maximal recall was achieved. Monolingual projects gave slightly better results than their bilingual counterparts. Three conclusions can be drawn from our experiments.

- There is a range of confidence scores, in our case 68-75, which, for a term length of 3-10, excluded a substantial amount of candidates (26%) while only excluding a small part of multi-word units (5%).

- Raising the minimum term length also reduces the number of term candidates much more than the number of multi-word units. In our case, with term length 7-10 we could eliminate over 30% of term candidates compared to a term length of 2-10, while only excluding 5% of multi-word units.

- The multi-word units of the type Prep + (Det +) N were particularly difficult for the system to identify. More than half of them were never found, against only 3% of all other syntactic types in our text. This

effect overrides any differences in variability of form or Mutual Information Index.

The experiments we did prepare the ground for the following questions to be explored on the basis of a larger amount of text.

- How can the range of confidence scores with a relatively large number of multi-word units be determined *a priori]*

- What are the optimal term length settings? Are they text-independent and if not, how can they be determined for a particular text?

Given the results we obtained, it is not necessary to consider bilingual projects and noise settings below the maximum. By reducing the number of parameters, it will be feasible to carry out relevant experiments on a larger amount of text.

References

Fernández Parra, María (2007), 'Towards a definition and classification of formulaic language for its translation in specialized texts', in Nenonen, Marja & Niemi, Sinikka (eds.), *Collocations and Idioms 1: Papers from the First Nordic Conference on Syntactic Freezes, Joensuu, May 19-20,* 2006, Joensuu: University of Joensuu, pp. 113-127.

ten Hacken, Pius & Fernández Parra, María (2008), 'Terminology and Formulaic Language in Computer-Assisted Translation', *SKASE Journal of Translation and Interpretation* 3:1-16.

Manning, Christopher & Schütze, Hinrich (1999), *Foundations of Statistical Natural Language Processing,* Cambridge, MA: MIT Press.

Rogers, Margaret (2007), 'Lexical chains in technical translation: A case study in indeterminacy', in Antia, Bassey Edem (ed.), *Indeterminacy in Terminology and LSP: Studies in Honour of Heribert Picht,* Amsterdam: Benjamins, pp. 15-35.

UNAIDS (2006), *2006 Report on the Global AIDS Epidemic,* http://www.unaids.Org/en/KnowledgeCentre/HIVData/GlobalReport/2006/ (24 June 2008).