A Comparison of Three Metrics for Detecting Cross-Linguistic Variations in Information Volume and Multiword Expressions Between Parallel Bitexts¹

Éric Poirier¹

¹ Université du Québec à Trois-Rivières, Québec, Canada
eric.poirier@uqtr.ca

Abstract. This paper presents the results of a comparison of three metrics for measuring cross-linguistic variations in information volume between parallel segments of a bilingual corpus. The performance of each metric is compared with the results of a human annotation of multiword expressions (MWEs) in each segment. The first metric measures characters in source and target segments and compares the variation, if any, with the expected character count ratio based on averages for the entire source and target texts. The second metric follows the same method except that it measures graphical word count (function and content words combined) in target and source segments. The third metric involves an analysis obtained via the content word precision (CWP) algorithm coded in Python. The purpose of the comparison is to determine which metric is closer to the human annotation and is, therefore, a better indicator of a large spectrum of MWEs.

Keywords: cross-linguistic phraseology, detection of multiword expressions, information volume variation, content word precision algorithm.

1 Introduction

As a contribution to computational studies of multiword expressions (MWEs), this paper presents the results of a comparison, albeit small-scale, of three metrics for detecting MWEs in parallel segments of a bilingual corpus. The untested underlying hypothesis of the comparison is that information volume variation in parallel segments (as established by content word imbalance) correlates with the presence in source or target of a large spectrum of MWEs. Hence, MWEs should occur in parallel segments where there is a cross-linguistic information volume difference, as determined by differences in the number of content words in source and target.

The study of MWEs in parallel segments is of the utmost importance to cross-linguistic phraseological studies as described in Colson (2008), and hence to transla-

¹ I wish to thank the reviewers for their comments and suggestions. I also wish to thank my colleague Paul John from UQTR for his input in reading the final version of the paper. Of course, any remaining omissions or errors are mine.
tion studies and phraseological studies. Determining which metric is most accurate in
detecting MWEs will also contribute to recent works in corpus-based phraseology
such as Granger and Paquot (2008) and will allow the mining of a large spectrum of
MWEs (such as clusters, lexical bundles, n-grams, recurrent sequences) or even new
classes of MWEs. Just like any corpus-based approach, our new bilingual approach is
designed to complement other traditional monolingual approaches such as appear in
volumes 1 and 2 of the Oxford Dictionary of Current Idiomatic English (Cowie &
Mackin, 1975; Cowie, Mackin & McCaig, 1983).

As a preliminary phase of a larger project, this paper presents the results of a small-
scale analysis of the first 25 segments of the bilingual text selected. The text chosen
for the evaluation of metric performance is the parallel English-French Inaugural
Address of J.F. Kennedy (January 20, 1961). The French version used in our corpus is
the official French translation provided on the multilingual pages of the website. The
source text and its translation were automatically aligned in source and target
segments with Logiterm, a proprietary software creation tool of bitexts using HTML
output format.

2 Definitions

The notion ‘volume of information’ is defined as the number of content words in each
parallel language segment. This definition makes it possible to determine a ratio of
information volume in source and target segments as measured by the number of con-
tent words they each contain. For purposes of comparing the three metrics with regard
to detection of MWEs, the information precision of the translation (IPT) is the ratio of
target segment information volume to source segment information volume, as shown
in figure 1.

Figure 1. Formula for calculating IPT

\[
\text{IPT} = \frac{\text{Number of content words in source segment}}{\text{Number of content words in target segment}}
\]

This ratio is an enhancement\(^2\) of the BLEU metric with modified n-gram precision,
as proposed by Papineni et al. (2002), and it is used for the assessment of machine
translation outputs in comparison with a human translation. IPT measures the vari-
tion of information volume in translation.\(^3\) Bilingual segments having the same in-

\(^2\) Instead of taking into account all words of a segment as with BLEU, the IPT ratio focuses exclusively on content words. This approach is more appropriate to translation, which is a meaning-based process. That is, it makes sense in the assessment of translation quality to favor content words over function or grammatical words, which contribute less to the information conveyed in a segment.

\(^3\) Interestingly, information volume is associated with adequacy, which is one of two qualities of a good translation based on human judgment, the other being fluency, as defined by
Stylistic effects excluded, information volume has also been used to define basic translation errors, such as addition or omission (Delisle, Lee-Jahnke & Cormier, 1999). Most of the time, a change in information volume (i.e., addition or omission of meaning) implies the addition or omission of a content word, but this is not always the case, particularly in segments containing phraseological units or MWEs. That is, many occurrences of parallel segments unequal to 1 cannot be attributed solely to these two basic translation errors. Other language or textual constraints occurring in the translation process cause variations in IPT. Provided the translation is correct, an information imbalance is an indication of linguistic or textual constraints which are often, but not necessarily, caused by a large spectrum of idiomatic and phraseological units and MWEs. Our untested hypothesis regarding a correlation between information volume imbalance and the presence of MWEs in parallel segments is based on this assumption.

According to our hypothesis, MWEs may be detected in segments that are anisomorphic (having an IPT score other than 1). As Franco Aixelà (2015) suggests, anisomorphisms have multiple origins and are not limited to linguistic structures, such as interpretative, pragmatic and cultural anisomorphisms. In a bilingual parallel corpus, the divergence of a segment’s IPT score from 1 derives not only from linguistic anisomorphisms, but also from various textual anisomorphisms. The latter include idiosyncratic constraints in the translation of MWE in target text, which the metrics may contribute to detecting. MWEs detected in anisomorphic segments may include for example expressions such as {globe; monde entier} and {(the) free; États libres}.

3 Methods

To identify the most effective method of detection of MWEs in parallel segments, the performances of character count metric, of word count metric and of the content word precision algorithm (Poirier, 2014) are compared to a human annotation of the same bilingual segments in a parallel corpus. Inter-annotator agreement on phraseological units in the extract of the parallel bitext was not checked but will be in a subsequent extended research project.

Described below, the human annotation of the bilingual segments served as the gold standard measure of information volume similarity or discrepancy to which the three other metrics were compared.

Snover, Madnani, Dorr and Schwartz (2009). Since adequacy measures whether the translation conveys the correct meaning, measuring adequacy requires the decomposition of sentence meaning into smaller parts (phrase and word level meaning). Measuring correct meaning in a translation therefore involves alignment of information content at the phrase and word level between parallel segments.
3.1 Human annotation protocol

Content word pairs in source and target segments were aligned manually with the meaning-based heuristic described in Poirier (2016). The heuristic requires that each content word in the source segment be manually aligned to at least one content word or to a null token (Ø) in the target segment and reciprocally that each content word in the target segment be manually aligned to at least one content word or to a null token in the source segment. Therefore, anisomorphic segments may contain the null token associated to either 1, 2 or more content words in the opposite segment. These alignments have been coded as 1-to-Ø, 0-to-1, 2-to-Ø, Ø-to-2, and so on.

The first segment of our corpus was the following salutation at the beginning of the address: [English source] Vice President Johnson, Mr. Speaker, Mr. Chief Justice, President Eisenhower, Vice president Nixon, President Truman, Reverend Clergy, fellow citizens; [French target] Monsieur le Vice-président Johnson, Monsieur le Président, Monsieur le Président de la Cour suprême, Monsieur le Président Eisenhower, Monsieur le Vice-président Nixon, Monsieur le Président Truman, révérend clergé, chers concitoyens.

The bilingual segment contains 13 1-to-1 aligned pairs (a) such as {Johnson; Johnson}, {Mr.; Monsieur}, and {Speaker; Président}. There are no 1-to-Ø anisomorphic aligned pairs (b), although there are four occurrences of the same 0-to-1 aligned pair (c) {Ø; Monsieur}. For these alignments, there is no counterpart in English for the formal use of Monsieur in French, which is a recurrent element in French sequences of a phraseological nature. The bilingual segment contains three many-to-many anisomorphic aligned pairs (d): a 2-to-3 pair as in {Chief Justice; Président Cour suprême} and two occurrences of the same 2-to-1 aligned pair: {Vice President; Vice-président}.

The manual count of content word balance between source and target in the first segment can be schematized as a+b+c+d, that is, 13+0+0+6 = 19 in English and 13+0+4+5 = 22 in French. This count shows that the segment is anisomorphic in terms of information volume. Assuming the proposed correlation of anisomorphic segment to MWE is correct, this segment should contain an MWE, which can be confirmed with the aligned pairs above, since these contain some MWEs in a broad sense. In comparing the performance of the three metrics, success constitutes the detection of an anisomorphic segment in which an MWE has been detected with the human annotation.

---

4 The aligned pair is cited without its function words because the items aligned are content words.

5 This anisomorphism is simply due to a morphological spelling difference which might be perceived as arbitrary on a theoretical/linguistic level, but which still counts, on a practical level in translation, as an incorrect usage. Although minor, this error is typical of misuse of MWE that translators are expected and required to recognize and use properly.
### 3.2 Character count metric

This metric compares the number of characters in source and target segments. Since character uses are generally not correlated with meaning (as opposed to information volume or content word counts), a global ratio of characters in source to target text is first compiled to normalize, if any, the expected discrepancy of character counts by segment. The value of the ratio could vary from one text and one translator’s style to another. In the corpus, the character count (in Word 2013, but any other software or version of Word applied consistently to all the segments could be used) of the whole text in source and target languages shows a ratio of 1.23 between English source text and its French translation. That is, there are 6199 characters (without spaces) in the source text and 7664 characters (without spaces) in the target text (with the calculation being very similar when spaces are taken into account). The ratio used to qualify segments as isomorphic was based on a character count difference of 1.23 between English and French segments. To allow for a standard deviation from this ratio, arbitrary but adjustable isomorphic minimum and maximum thresholds were calculated as a 10% difference from the expected character count ratio of 1.23. Minimum thresholds were rounded down to the nearest whole number, and maximum thresholds were rounded up. For the first segment of our corpus (as cited in the previous section), these thresholds are described in table 3 below.

<table>
<thead>
<tr>
<th>Characters in source</th>
<th>Characters in target</th>
<th>Target expected count (1.23 ratio)</th>
<th>Isomorphic threshold minimum</th>
<th>Isomorphic threshold maximum</th>
<th>Character count metric prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>151</td>
<td>233</td>
<td>185.73</td>
<td>167</td>
<td>204</td>
<td>anisomorphism</td>
</tr>
</tbody>
</table>

For the first segment, the comparison of the actual character counts in the target segment with the expected character count, as calculated with the normalized ratio for the entire text, predicts that the target segment is anisomorphic with respect to the source segment. In other words, the actual character count in target segment is much higher (233) than the expected character count (204), as determined by the normalized ratio.

---

6 There is empirical evidence in professional translation that French translations are approximately 15% longer than their English source text in terms of words, but we have not found scientific works confirming this principle in terms of character length. A recent study on POS-Editing (Béchara, Ma & van Genabith, 2011) found that in a set of 52,383-word English-French segment pairs, the average segment length was 13 words in English and 15 words in French, which confirms the 15% variation in words.
3.3 Word count metric

This metric compares the number of all content and function words present in source and target segments. For this metric, words are defined as any chain of characters in a segment that are separated by a space. As with the character count metric, a ratio of words in source to target entire text was first calculated to normalize the expected discrepancy of word counts between source and target segments. A ratio of 1.09 was calculated with 1363 words (in Word 2013) for the entire source text and 1485 words (in Word 2013) for the whole target text. To allow for a standard deviation from this ratio, arbitrary but adjustable isomorphic minimum and maximum ratio thresholds were calculated as a 5% difference from the expected word count ratio of 1.09. The deviation interval of 5% is an approximation of tolerance, taking into account the fact that word counts are partly based on function words (not correlated with lexical meaning and information content) and partly based on content words (correlated to meaning and information content). Depending on the type of texts and translations, this deviation interval could be modified as needed. The 5% standard deviation defines a minimum tolerance of 1.09-0.0545=1.0355 word count ratio and a maximum tolerance of 1.09+0.0545=1.1445 word count ratio for isomorphic segments. As for the character count metric, thresholds were rounded up and down to the nearest whole number. For the first segment of our corpus, these thresholds are described in table 4 below.

Table 4. Word count result for the first segment

<table>
<thead>
<tr>
<th>Words in source</th>
<th>Words in target</th>
<th>Target expected count (1.09 ratio)</th>
<th>Isomorphic threshold minimum</th>
<th>Isomorphic threshold maximum</th>
<th>Word count metric prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>30</td>
<td>20.71</td>
<td>19</td>
<td>23</td>
<td>anisomorphism</td>
</tr>
</tbody>
</table>

The comparison of the actual word count in target segment with the expected word count predicts that the first target segment is anisomorphic compared with the source segment. In other words, the actual word count in the target segment is much higher (30) than the maximum word count threshold (23), as compiled with the normalized ratio for the whole text.

3.4 CWP algorithm

The CWP algorithm is coded in Python, as described in detail in Poirier (2014). This method is an automatized version of the human annotation method and is based on the deletion in parallel segments of function words, which, in English and French (as in most languages), form a closed set of items. A comparison is made between the

---

7 The algorithm works with any language which uses distinct and separate function and content word sets. It is thus operational for English, French and Spanish corpora. For other lan-
number of content words in the source and target segments. As for the manual annotation, no specific threshold was used as an expected difference in content word counts for the target segments. Tolerance margins are not required for short parallel segments containing few content words, but might be useful when segments are longer. There is no known standard as to when to use maximum and minimum thresholds to consider that source and target segments are isomorphic in terms of content words. This uncertainty, and the added complexity needed for the algorithm to take into account segment length in the comparison of content words, made us defer the implementation of these options to a future version of the algorithm.

For the first segment of our corpus, the content word counts in source and target segments, as determined by the CWP algorithm, are described in Table 5 below.

Table 5. CWP result applied to the first segment

<table>
<thead>
<tr>
<th>Content words in source</th>
<th>Content words in target</th>
<th>CWP metric prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>22</td>
<td>anisomorphism</td>
</tr>
</tbody>
</table>

The comparison of the content word count in target segment with the content word count in source segment gives the same result as the human annotation (19, 22 and an anisomorphic parallel segment) because both used a method based on the distinction between function and content words. However, the performance of the algorithm over the 25 segments is not identical to the human annotation, with more details on these differences provided below in the next section.

4 Results

Among the first 25 segments manually annotated, only five were of isomorphic nature, according to the human annotation. This suggests that volume information imbalance is very common in translation and, correlative, if our untested assumption is correct, that the text analyzed contains numerous occurrences of MWEs (19 segments out of 25 would include one or more MWEs).

The table below shows the success rate of the three methods in the detection of anisomorphic segments and potential MWEs.

Table 6. Success rates compared with the human annotation (25 segments)

<table>
<thead>
<tr>
<th>Character</th>
<th>Word</th>
<th>CWP</th>
</tr>
</thead>
</table>

languages, the distinction between function and content words has not been tested for the application of the CWP algorithm. Most typological classifications of languages (such as isolating, agglutinative and inflecting) are not concerned with the possibility or even the feasibility of separating function and content words. In the current initial state of development of the algorithm, the concepts of function and content words seem to be universal, and the limitations of these concepts presently seem to be more related to the use of similar or equivalent words classified differently (see discussion section) than to the existence of languages not having function and content words.
As we can see from Table 6, the CWP algorithm, at 88%, has the highest success rate in the detection of anisomorphic segments. Although interesting, the CWP results are not always identical to the human annotation. In three occurrences, the CWP algorithm did not properly detect an isomorphic or an anisomorphic parallel segment. One occurrence is an anisomorphic segment analyzed as isomorphic (false negative), and in the two other occurrences, an isomorphic segment is wrongly analyzed as anisomorphic (false positive). Two of these failures are due to a difference between French and English in word categorization and to a function word such as “all”, used as quantifier (function word), being translated by “toutes”, used as an indefinite adjective (content word). The third wrong detection is due to the use of “dare” as an auxiliary (specifically, modal) verb in English. The algorithm processed it as a full lexical verb, while the human annotation aligned it with a null token (not translated) in French, as in segment 25 [We dare not tempt them with weakness = Nous ne les tenterons pas par notre faiblesses].

5 Discussion

Two of the failures of the CWP algorithm may be due to arbitrary uses and categorizations of function and content words in English and French. Resolving this issue seems impossible, since each language classifies semantically equivalent words in different categories, so it is probably best viewed as a demonstration of language idiomaticity in grammatical uses and analysis of specific words. Adopting this view just pushes the limits of MWEs a bit further towards the recognition of irreducible, incompatible, and hence phraseological, uses of lexical items between languages, even if on their meaning side, these expressions are transparent and easily translatable.

The third failure of the CWP algorithm seems to be different in that it relies on a correct analysis of a particular verb acting as a function word (modal auxiliary) in some contexts and as a content word (main verb) in other contexts. In this case, it seems that adding an automatic POS tagging step in the CWP algorithm may help to resolve this specific issue. However, automatic POS tagging is not 100% accurate, and the figures are even lower when compiled on a segment level.

---

8 As in I dared not move [modal] vs I dared him to move [main verb].

9 Giménez and Marquez (2004) report an accuracy of close to 97% on a word count basis. Converting this level of accuracy at the segment level makes it less impressive since it can be reasonably argued that most segments (and sentences) are often comprised of 10 words or more. For example, for ten segments of 10 words, the 97% accuracy implies that as much as three segments out of ten (30% of segments) could contain a POS tagging inaccuracy pro-
Although the CWP algorithm permits detection of a large spectrum of MWEs, a significant limitation of the automatic comparison of information volume is that literally translated MWEs (such as “tomber comme des mouches” = “to drop like flies”, “donner le feu vert” = “give the green light”) cannot be detected by a discrepancy in information volume, even though the examples given are both well-known MWEs from a monolingual and traditional approach in phraseological studies. However, this limitation of the CWP algorithm may well be viewed positively as a supplemental approach to traditional monolingual mining of MWEs. Another limitation of the CWP algorithm is that MWEs entirely constituted of function words and having no content words, such as “this much”

9

however, this limitation of the CWP algorithm may well be viewed positively as a supplemental approach to traditional monolingual mining of MWEs. Another limitation of the CWP algorithm is that MWEs entirely constituted of function words and having no content words, such as “this much”

Conclusion and further studies

Comparison of the three metrics in detecting information volume variation shows that the results of the CWP algorithm are closer to the manual annotation of the same parallel segments. The CWP algorithm can thus help to detect a large spectrum of corpus-based MWEs, albeit with the limitations detailed in the discussion section. Analysis of failures has shown that implementing a POS tagging module to enhance the content word analysis with the CWP could marginally increase its success rate on a segment-level basis.

The implementation of POS tagging and a contextual analysis of ambiguous words are left to a future study on a larger corpus. In such an experimental design it would be advisable to use human annotation only for segments in which MWEs have been detected to confirm or invalidate the presence of MWEs after the application of the CWP algorithm. Inter-rater agreement could also be measured at the same time. Further studies are also required in the assessment of two types of anisomorphic segments in relation with MWEs: those in which source segments contain more content words than do target segments and those in which target segments contain more content words than do source segments.

Automatic or semi-automatic detection of MWEs in parallel segments constitutes a significant contribution to translation studies for teaching and translation quality assessment purposes, as well as to cross-linguistic phraseology as part of bilingual lexicography and usage-based linguistics. Extending MWE categories to textual constraints could help translation studies specialists distinguish between free translation shifts due to circumstantial and creative reasons and language-constraint translation shifts due to idiomatic textual uses in languages. The possibility of separating and studying both categories of translation shifts represents an advance in translation studies.

vided the three words inaccurately tagged out of 100 are distributed in three different segments.

This MWE was found in our corpus and detected as such with the manual annotation (but not the CWP). According to the Oxford English dictionary (online), it means “The fact about to be stated”.

9

10
The detection of MWEs in parallel segments also contributes to monolingual and traditional studies in phraseology and linguistics by extending the limits of language idiomaticity and uncovering new items in established categories of MWE or suggesting new classes of MWE.

References