Chapter 7

What can Euclidean distance do for translation evaluations?

Éric André Poirier
Université du Québec à Trois-Rivières

We describe an empirical method to screen informational translation shifts in parallel segment pairs extracted from a bilingual or multilingual translation corpus using two linguistic features that are independent of the languages matched by the translation. The method applies to most known languages and in one or the other of the two translation directions (direct or inverse). The features measured for each segment in source and target languages are character count and lexical word count (or information volume). Information volume is compiled through an algorithm coded in Python using spaCy v2.1.3 core linguistic models. The values of source and target segment features and the translation precision ratio of each segment pairs are averaged over the text to which they belong and all segment values are standardized in relation to their textual average. The deviation between standardized values for each segment in a pair, as measured by the weighted Euclidean distance, allows for the screening and identification of target segments that are atypical or heteromorphic in comparison with their source segment. Our hypothesis is that those heteromorphic segment pairs, as opposed to isomorphic ones, are more likely to contain informational translation shifts. The objective and reproducible method described herein allows for semi-automatic identification of problematic translations and uncovering of textual and linguistic facts revealing translation processes, contingencies, and determinism.

1 Introduction

We describe below the theoretical framework and the methodological steps of the method that we have applied in a systematic and exploratory way to parallel bilingual corpora in different languages and in different translation directions with

Éric André Poirier. 2021. What can Euclidean distance do for translation evaluations? In Mario Bisiada (ed.), *Empirical studies in translation and discourse*, 165–198. Berlin: Language Science Press. DOI: 10.5281/zenodo.4450089

English, French, and Spanish. The method may be applied manually on small texts and for pedagogical purposes in the analysis, evaluation, and comparison of translations and translation processes, or it can be implemented for manual identification of informational translation shifts in automatically screened segment pairs in large corpora. Automatic POS tagging of all segment pairs was done with spaCy v2.1.3 (a commercial-grade natural language processing environment, (Explosion_AI 2016–2020)) core linguistic models in an algorithm coded in Python version 3.7.3 with language models en_core_web_sm (version 2.1.0) for English and es_core_news_sm for Spanish (version 2.1.0). POS tagging is required to calculate the information volume of each segment. Languages that are covered with our method are determined by the availability of a specific linguistic module in the spaCy environment designed for Python programming.

For illustrative purposes, we present the results obtained with the method applied to the United States President Barack Obama's speech to the Cubans on March 22, 2016, for which an official translation is provided in Spanish. The bitext used for the analysis was compiled with the original English version[1] and its official Spanish translation,[2] both of which are posted on the obamawhitehouse.archive.gov Web site, which includes official speeches delivered by President Obama. The speech has 2,420 words in English, 2,468 in Spanish, and the raw bitext was segmented in 255 segment pairs, as described below in §3.

2 Theoretical framework

Before explaining our method, we describe the typology of informational translation shifts for the manual annotation and analysis that is required to measure the efficiency of the method. This framework also describes key concepts in the evaluation of the efficiency and utility of the method we present regarding the screening of segment pairs which are most likely to contain translation shifts.

2.1 Free and fixed translation shifts

First, let us define what we mean by translation shift and propose a typology of the types of informational shifts found in the segment pairs of parallel translation corpora. The term shift is used in its broad sense to mean "a change in position

[1] https://obamawhitehouse.archives.gov/the-press-office/2016/03/22/remarks-president-obama-people-cuba

[2] https://obamawhitehouse.archives.gov/the-press-office/2016/03/22/discurso-del-presidente-obama-al-pueblo-cubano

or direction".[3] Translation shifts generally refer to specific changes attributable to translation, explained thus: "The transformation which is occasioned by the translation process can be specified in terms of changes with respect to the source texts, changes which are termed 'shifts'." (Bakker et al. 2011: 269)

In this sense, translation shifts do not include systematic or systemic differences between languages. Although no empirical criteria have been provided to differentiate between translation shifts and differences between languages, it is generally accepted that these two transformations in the translation process must be distinguished. To account for these two very different types of shifts, we have adopted the terminology of Wecksteen-Quinio et al. (2015). The authors distinguished fixed shifts that are attributable to differences between languages from free shifts that are attributable to the translation operation itself and result from a choice freely exercised by the translator, from bias on the part of that person, or simply from translation errors. While fixed shifts are mandatory, free shifts are by definition free or the result of a deliberate choice. Strictly speaking, they are members of a group of at least two expressions that adequately translate the expression or the same elements of the source segment. In theory, fixed shifts describe conventional translation processes, while free shifts describe creative, original, or to some extent novel translation processes. Instead of relying exclusively on our own judgment on the acceptability of Spanish translations, we designed a process that supports the empirical definition of free shifts based on the *tertium comparationis* provided by machine translation. For a source expression, if a literal translation in the target text co-occurs with an acceptable literal translation of the same expression in DeepL,[4] the shift in the official translation is fixed. When the target text contains a non-literal translation, if the same source expression is translated literally in DeepL, the shift is considered free. A good example among others (see §4) is the translation of the segment number 187 "that is a measure of our progress as a democracy" that was translated as "Esa es la medida de nuestro progreso", which is not literal and which co-occurs with a literal translation in DeepL "que es una medida de nuestro progreso como democracia". The comparison with DeepL highlights the omission of the content word *democracia* in the official translation. Translation shifts screened with our method are limited to informational translation shifts and can either result in the addition of one or more content words or the omission of one or more content words (see §2.3 below).

[3]Source: Online Cambridge Dictionary at https://dictionary.cambridge.org
[4]https://www.deepl.com/translator

2.2 Informational translation shifts

The term "informational shift" refers to a particular type of translation shift. In the identification of all translation shifts (semantic, lexical, syntactic, stylistic, terminological, socio-linguistic, etc.) that are required for the knowledge and maintenance of a coherent set of translation processes (which constitute the basic elements of translation learning and teaching), informational translation shifts represent a critical group of translation shifts. In fact, they are requisite to the proper identification and definition of all other types of shifts since informational shifts affect the information content of the messages to be translated, which is required to be invariant in the translation of pragmatic texts, and on which the analysis and evaluation of other translation shifts depend.

We hypothesize that informational translation shifts are most likely present when a comparison of source and target segments show an important discrepancy or "distance" in two correlative linguistic features: the string length in characters and the lexical word count. Lexical words are numerous; they carry a lexicalized or stable meaning and form an open class of elements. This is in contrast with grammatical words that are few, do not carry a lexicalized meaning, and form a closed set of elements. By counting lexical words in source and target segments (in two different languages), the method we describe here allows for the quantifying of the translation precision in terms of information volume. This measure is defined in the next section.

2.3 Positive and negative information shifts

As discussed in §2.1, information shifts may result in the addition or the omission of information. The volume of information as measured by the lexical word count is an approximation of the quantity of basic (stable) information present in source and target segments. The translation precision ratio (TPR) is calculated by dividing the information volume of the source segment by the information volume of the target segment and may be "positive", "negative" or "neutral". TPR is a numeric measure of the discrepancy of information volume between target and source segments. When segment pairs contain an equal volume of information in both the source and target segments, the TPR between the two segments is "neutral" with a value of 1.0 and those segment pairs are isomorphic. When segment pairs contain at least one negative information shift, that is, the omission of information in the target segment, the information volume of the target segment is smaller than the information volume of the source segment. The TPR between the two segments is "negative" with a value lower than 1.0 and those

translation segment pairs are negative heteromorphic. When segment pairs contain at least one positive information shift, that is, the addition of information in the target segment, the information volume of the target segment is greater than the information volume of the source segment. The TPR between the two segments is "positive" with a value higher than 1.0 and those translation segment pairs are positive heteromorphic.

Since information shifts mostly occur within the segment level, numerous combinations of positive and negative shifts may exist in isomorphic, negative heteromorphic, and positive heteromorphic segment pairs. For example, an isomorphic segment pair may have one positive shift and one negative shift, each canceling out the value of the other and a heteromorphic segment pair may have multiple negative shifts and positive shifts. In this case, there may be a single positive or negative shift, as the case may be, or there may be multiple negative or positive shifts that combine within a segment pair that is either negative or positive as a whole.

2.4 Antinomic shifts

Antinomic shifts are those whose positive or negative nature is opposite to that of the whole segment to which they belong. For example, a positive heteromorphic segment pair may contain two positive shifts of one lexical word each or a single positive shift of two lexical words, in combination with a negative shift of one lexical word that does not contribute to the positive orientation of the segment pair. The positive or negative orientation of antinomic shifts is opposite to that of the orientation of all the combined shifts of a pair of segments. In neutral isomorphic segments (having a TPR of 1.0), any pair of information shifts that may occur (one positive and one negative) cancel each other out and are therefore both antinomic. For this reason, it should not be concluded that there is no informational translation shift in isomorphic segment pairs. However, as demonstrated in §5, we hypothesize that there are fewer of them in isomorphic segment pairs than in the positive or negative heteromorphic segment pairs.

2.5 False shifts and undetected shifts

Because of the shortcomings of the spaCy v2.1.3 core linguistic models and the erroneous results they sometimes produce as regards POS tagging, we created two other categories of information shifts that could only be detected through manual and meticulous analysis of the segment pairs screened by the weighted Euclidean distance (see §3.3). One difficulty in POS tagging is that most tokens

belong to several lexical or grammatical word classes. Some parts-of-speech are also equivocal regarding their belonging to a lexical or a grammatical class. This is the case, for example, of verbal auxiliaries in English, Spanish or French, or for some particles in phrasal verbs in English – are they adverbs or prepositions? Most POS tagging algorithms struggle to provide a proper analysis of all source and target segment tokens (despite, and with the support of, language-specific rules), and for specific tokens or POS may present original aberrations that need to be corrected. For some older releases of spaCy's POS tagger, Giesbrecht & Evert (2009) report a success rate of less than 93%, and this rate varies (downward) depending on the type of text analyzed. When manual analysis reveals errors or anomalies in POS tagging of tokens, the involved information shifts have been classified as false shifts (in the way that they are false positives) that owe their existence only to POS tagging errors. Another development that would enhance the efficiency of the empirical method described here is the improvement of POS tagging such that every token and every compound or group of tokens would be properly tagged as a lexical or a grammatical item. As we explained in a previous paper (Poirier 2017: 8), converting even a 97% POS tagging accuracy at the segment level makes it less impressive since it can be reasonably argued that most segments (and sentences) generally have at least 10 words or more. For ten segments of 10 words, an accuracy of 97% would imply that as much as three segments out of ten (that is 30% of segments) would contain a POS tagging inaccuracy provided the three words inaccurately tagged out of 100 are distributed in three different segments. Furthermore, considering that parallel corpora involve two different languages (and two different POS tagging sources of errors), this number may skyrocket to 60% of all 10 segment pairs if the two language-specific groups of 30% erroneous segments are each matched to a properly analyzed source or target segments.

When the POS tagging modules produce an erroneous analysis that results in the inexistence of an information shift (and which produces a false negative), these information shifts that go unnoticed have been classified as undetected shifts, i.e., shifts that were not detected because of wrong POS tagging. For example, an undetected shift was found in segment number 63 of our corpus (see §4.1) which contained the expression "a multi-party democracy" matched with the Spanish translation "*una democracia de múltiples partidos*". The source segment was wrongly analyzed as having four lexical words by the English language model of spaCy,[5] giving rise to a false shift and a fourth lexical word. In this case,

[5] In this case, this was due to the the hyphen being wrongly analyzed as an adjective, but this was not the only wrongful POS tagging issue with the hyphen since in parallel segment number 239 (see §4.3) it was analyzed as a proper noun.

the target segment was analyzed correctly with three lexical words. What the module analysis made as a negative heteromorphic segment pair turns out to be a positive (antinomic) heteromorphic segment pair because *multi-party* should be analyzed as a unitary lexical word (compound). Thus, in this segment, our manual analysis found an undetected information shift that both linguistic language models have been unable to bring to light.

3 Corpus data processing methodology

The file used as input is a bitext file in HTML format provided free online by YouAlign[6] (maximum file size for each file is limited to 1MB). The speech file size of our corpus did not exceed this limit but one could use a comma-separated value file format or other proprietary bitext creation software such as Logiterm Pro v5.8.2 for larger files and corpora. It has been verified that the alignment of all segments of the bitext is adequate and that each source segment matches its translation with one or more target segments, if applicable. Manual processing was necessary at this step on the source and target language plain text of the speech. In our corpus, annotations such as "Applause" and "Laughter" that describe the audience's reaction to the speaker's words have not been included and translated in the target text. It seems fair and reasonable to delete those items that were not genuinely communicated by the speaker and not translated in Spanish because they were provided by the context. For reasons that are difficult to explain (and which probably have to do with character encoding or the core and basic language models that were used even if some testings with more complete language models that were available at the time did not demonstrate noticeable improvements), some abbreviated forms with apostrophes in English needed to be modified as the last recourse solution (such as *that's = that is*) because the apostrophes were recognized as lexical words, which is not accurate. In the English source text, the last greeting from the speaker is "*muchas gracias*" in Spanish which obviously does not need to be translated. This last single segment needed to be removed from the bitext since it cannot form a pair of parallel bilingual segments. For the target speech in Spanish, the segmentation results with the dash and colon had to be corrected to match the segmentation results of their corresponding punctuation marks in English.

Once these modifications were made to our corpus, a module written in Python analyses all the pairs of segments of the corpus one by one. In this analysis, two specific linguistic modules are called sequentially for the source segment and the

[6]https://youalign.com

target segment to count their lexical words and measure the total information volume of each segment as well as the TPR of each segment pair. These values are appended to a variable and then exported to a CSV file which can be read in a spreadsheet. The character count of each segment could be quantified afterward in the spreadsheet with the help of a function such as LEN (cell) function in Excel. We also calculated for each source and target language text the average value of information volume and character count by segment for the whole corpus.

Our English-Spanish parallel corpus of Obama's speech to the Cubans contains 255 segments of 9.49 lexical words and 89.68 characters on average in English and 9.68 lexical words and 96.15 characters on average in its translated version in Spanish. These averages were calculated with the values of both linguistic features for the whole corpus as described in the Table 7.1.

Table 7.1: Total values of linguistic features in source and target languages

	Characters	**Lexical words**
English (source)	22,869	2,420
Spanish (target)	24,517	2,468
Difference	+ 7.21%	+ 1.98%

In recent years, we have applied different versions of our methodology and translation precision algorithm to some corpora (see Poirier (2019) for an example of English-French analysis and earlier methodology). We have found that three linguistic features may be measured for each parallel segment pair in bitexts. These are character count, total word count (or token count), and lexical word count. We tested the correlation of each feature in different corpora that were analyzed with our methodology. In order to measure the correlation of these features we simply applied the Pearson correlation coefficient between two variables (values of linguistic features in source and target segments) as defined with the following formula, where *cov* is the covariance, ρ_X and ρ_Y are the standard deviations of X and Y, respectively:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Table 7.2 below presents the correlation which was calculated with different political speeches in English translated in Spanish, such as Abraham Lincoln's Gettysburg Address (1863), Inaugural Address of John F. Kennedy (1961), Martin

Luther King's *I have a dream* (1963), Obama's speech to the Cubans (2016), Donald Trump's State of the Union (2018) and Oval Office Address on Border Wall (2019).

Table 7.2: Correlation of three linguistic features in English-Spanish translations

Speech	Number of words (text)	Number of characters	Total word count	Lexical word count
Gettysburg	268	0.9725	0.963	0.9691
KennedyInauguralAddr	1393	0.9888	0.984	0.9736
DreamMLKing	1673	0.9766	0.9772	0.9684
ObamaCuba	4161	0.9733	0.9663	0.9583
TrumpStateUnion2018	5188	0.9649	0.9408	0.9573
Trump_BorderWall	1119	0.9513	0.9205	0.9501
Averages		0.9712	0.9586	0.9628

Table 7.2 shows that on average, the character count has the strongest correlation (0.9712),[7] followed by the lexical word count (0.9628) and by the total word count (0.9586). Because of this high correlation of these features between the source and target segments, the significance of the lexical word and character differences between the source and target segments is difficult to establish when the length of segment pairs may vary widely. For example, the absence of a lexical word in a target segment that is associated with a source segment of 30 lexical words is not as significant as the absence of a lexical word in a target segment that is associated with a source segment of three lexical words.

3.1 Standardized values of segment pairs

To account for the relative length of each string in segment pairs, and to make each segment pair comparable in terms of their selected features, we standardized the value of the two features for each segment by relating them to their average value for the whole source or target segments in the parallel corpus. To this end, a rule of three was used to determine the standardized values of information volume and character count for each segment in pairs. In the context of Barack

[7] These data support previous works in machine translation, such as the seminal paper of Gale & Church (1993: 89), who found that there exist very high correlations between the length of a paragraph in characters and the length of its translation.

Obama's English-Spanish corpus, let's take for example a source and target segments having respectively 3 and 4 lexical words and 25 and 31 characters. If we relate these numbers to their average value for the corpus (9.49 lexical words and 89.68 characters for the source segments, and 9.68 lexical words and 96.15 characters for the target segment), we get standardized values of 4.07 (3x[9.49+9.68]/7) lexical words, for the source segment, and 5.53 (4x[9.68+9.49]/7) lexical words, for the target segment. The same formula is used for the character count standardized values. The standardized value of the two features measured for each pair of segments is crucial since they will make it possible to detect target segment pairs that are atypical (or unusually distant from their source segment), as measured by the weighted Euclidean distance.

3.2 Precision deviation factor

To characterize the positive or negative value of the information volume of the whole target segment compared to the whole source segment, we subtracted its TPR from the average value of this ratio for the whole text, a value which is normally close to 1.00 (a target segment normally contains the same volume of information as its corresponding source segment). In Barack Obama's Speech English-Spanish corpus, this figure was 1.02. Any segment pair having a TPR lower than 1.02 would, therefore, have a negative value, and, conversely, any segment pair having a TPR higher than 1.02 would have a positive value. The precision deviation factor (PDF) used in the calculation of the Euclidean distance is simply a multiple (10 times) of this value (positive or negative.[8]) Just like the positive and negative values of information shifts were an indication of a potentially wrong additional or missing information in the translation, the negative or positive value of the Euclidean distance would point to a potentially wrong additional or missing information in the target segment.

3.3 The weighted Euclidean Distance for screening segment pairs

The Euclidean distance is calculated using the standardized lexical word count and the standardized string length in characters that were calculated for the source and target segment of each parallel pair in the corpus. The exact formula of the Euclidean distance (d(p,q)) that we used is defined as follows:

[8] A value of zero is theoretically possible with a TPR of 0.99 but this value did not occur in our corpus. Some adjustments might be required in the following calculations to take this value into account.

$$d(\mathrm{p,q}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

Simply put, the Euclidean distance is measured by the square root of the sum of the squared deviations of the two features (information volume and string length in characters) measured and standardized for each source and target segment of all parallel pairs in our corpus. Since translation is an operation that takes into account meaning, we gave more weight to the difference in information volume than to the difference in the number of characters in the calculation of the weighted Euclidean distance. Multiplying the Euclidean distance by the positive or negative precision deviation factor gives more weight to the information volume and results in a positive or negative value of the distance between the source and target standardized segments of each parallel pair. When the value is negative, the target segment contains fewer lexical words or characters than the source segment and is likely to contain at least one or more negative shifts. Similarly, when the value is positive, it means that the target segment contains more lexical words or characters than the source segment and is therefore likely to contain at least one or more positive shifts.

This method has made it possible to calculate the weighted Euclidean distance separating each pair of segments. Of the 255 pairs of segments in the English-Spanish corpus of Barack Obama's speech in Cuba, the weighted value (by the precision deviation factor) of the Euclidean distance is between -193.81 and 313.26. Segment pairs with extreme negative or positive values of weighted Euclidean distance are highly heteromorphic and their target segment is very likely to contain informational translation shifts. The two most heteromorphic segments and their particular calculations are described in the next table. The segments are preceded by their sequential number in the English-Spanish corpus. Proper and improper content words (leading to false shifts) identified with the spaCy v2.1.3 language models are marked in bold. The volume of information and the number of characters in each source and target segment is in square brackets. In the calculation of the precision deviation factor, TPR is averaged at 1.02 and the weighting of this difference has been multiplied by a constant of 10. The precision deviation factor for segment #12 is therefore $(0.333-1.02)*10 = -6.87$. For calculating standardized values, the average value of source segments features are 9.49 lexical words and 89.68 characters, and for the target segment features these figures are 9.68 lexical words and 96.15 characters. Some slight differences may occur due to the rounding of the decimals and their precision. Table 7.3 presents the detailed

calculations for the most negatively heteromorphic segment pair number 12 and the most positively heteromorphic segment pair number 144 in the corpus. The table shows the translation precision ratio (TPR), the precision deviation factor (PDF), the standardised information volume (SID) and string length (SSL) of the segment, the Euclidean distance (d) and the weighted Euclidean distance (wd).

Table 7.3: Most negative and positive heteromorphic segment pairs and their linguistic feature values (lexical words in bold)

Segment	12. **Thank** you **very much**. = Muchas **gracias**. [3, 20 = 1, 15]	144. **Not everybody agrees** with me on this. = **No** todo el **mundo está** de **acuerdo conmigo sobre** esto. [3, 37 = 6, 52]
TPR	0.33	2.00
PDF	-6.87	9.80
SID	14.38 and 4.79	6.39 and 12.78
SSL	106.19 and 79.64	77.26 and 108.57
d	28.22	31.96
wd	-193.81	313.26

Manual analysis of the shifts in the segment pair number 12 shows that the great negative Euclidean distance is due to a false shift that is attributable to the classification of the Spanish adverb *muchas* as a determinant (a grammatical word), compared to the English adverb *much*, which is classified as an adverb and therefore as a lexical word. In the same segment, there is a second fixed shift with the use of the adverb *very* in English which has no corresponding Spanish equivalent (probably because *muchas* is already used as an adverb). The two shifts taken together explain the shift in the information volume of 2 found between the two segments. The number of characters is in the same negative direction of the information volume shift and reveals that the target segment of the pair is shorter than the source segment.

In segment pair number 144, there are three positive shifts of one lexical word for each in favor of the target segment. First, there is a fixed shift with the correspondence of the verb *agree* (one lexical word) and the Spanish phrase *está de acuerdo* (two lexical words). Then there are two false shifts with a positive value due to the wrong POS tagging of *conmigo* as an auxiliary (lexical word) rather than as a preposition or prepositional phrase, and another wrong POS tagging of

sobre as a verb (lexical word) rather than as a preposition. These two false shifts are due to POS tagging errors in the spaCy v2.1.3 language models. The analysis of *todo* seems to have been well done by the spaCy v2.1.3 language models since it is categorized as a grammatical word even though the type of grammatical word seems to be wrong, i.e. a determiner rather than an indefinite pronoun.

These two examples show the importance of POS tagging in the analysis of translations and the calculation of the information volume. It is to be hoped that significant progress will be made in this area. Despite scientific articles that regularly report success rates of 95% to 98% in POS tagging, it seems that these data are inaccurate, at least with spaCy v2.1.3 POS tagging modules.

4 Results

After having applied the corpus data processing methodology described above, we wanted to validate its efficiency regarding the screening of negative and positive informative translation shifts. For this purpose, we manually analyzed three samples (A, B, and C) of twenty pairs of segments screened automatically with the numeric value of the weighted Euclidean distance. Segment pairs in two of those samples (A and B) were selected for their highest (positive) and lowest (negative) weighted Euclidean distance between the source and target segments (and for being representative of the most negative and positive heteromorphic segment pairs within the analyzed English-Spanish corpus). In a third sample (C), twenty other segment pairs were selected for their very neutral (near zero) weighted Euclidean distance between the source and target segments (and for being representative of the most isomorphic segment pairs within the analyzed English-Spanish corpus). These three groups of 21 segment pairs were analyzed manually as regards the presence or the absence of information shifts described in §2. Detailed data on the manual analysis of each of the three samples of twenty segments is described in the next subsections below.

In each of the three annotation tables in the left column, lexical words in segment pairs are marked in bold to inform the reader of the results of the automatic POS tagging process. For difficult or ambiguous word-forms, parts of speech are indicated in uppercase when needed. The tag set that is used is the same as spaCy v2.1.3 POS tag symbols that are called Universal POS tags and that comes from the Universal Dependencies Scheme.[9] The empty symbol (Ø) is used to describe an item having no semantic match in the target segment. The asterisk symbol (*)

[9]https://universaldependencies.org/u/pos

is used to describe erroneous tagging which explains the false shift POS annotation. Information shifts are described in their order of appearance in the target segment.

4.1 Sample A annotations – most negative heteromorphic pairs

This section contains the manual analysis and annotations of sample A segments for the classification of information shifts observed in the most negative heteromorphic segment pairs. Translation pairs have been sorted from the longest negative Euclidean distance (-67.52)) to the shortest negative Euclidean distance (-15.67).

(12) **Thank**-V you **very**-ADV **much**-ADV. 3, 20 →Muchas-*DET/ADJ **gracias**. 1, 15

-193.81 [-2 lexical words] (2 shifts)

1. False shift POS (-1): much-ADV →muchas-*DET/ADJ
2. Fixed shift (-1): very-ADV →Ø [confirmed with DeepL: *Muchas gracias.*]

(258) And it **will not be easy**, and **there**-*ADV/PRON **will be setbacks**. 8, 52 →Y **no será fácil**, y **habrá reveses**. 5, 33

-165,00 [-3 lexical words] (3 shifts)

1. Fixed shift (-1): will-V →Ø [future tense]
2. False shift POS (-1): there*ADV/PRON be-V →haber-V
3. Fixed shift (-1): will-V →Ø [future tense]

(11) **Thank**-V you **so**-ADV **much**-ADV. 3, 18 →Muchas-*DET/ADJ **gracias**. 1, 15

-133.37 [-2 lexical words] (2 shifts)

1. False shift POS (-1): much-ADV →muchas-*DET/ADJ
2. Free shift (-1): so-ADV →Ø [DeepL: *Muchísimas gracias.*]

(75) **Why**-ADV **now**-ADV **?** **Why**-ADV**now**-ADV**?** 4, 17 →¿por qué **ahora**-ADV**?** 1, 15

-125.86 [-3 lexical words] (2 shifts)

1. False shift POS (-1): why-ADV (1) →por qué (0)

2. Free shift (-2): Why now? (2) →Ø (0) [confirmed with DeepL, translated twice]

(259) It **will take time**. 3, 18 →**Tomará tiempo**. 2, 14

-83.19 [-1 lexical word] (1 shift)

1. Fixed shift (-1): will-V →Ø [future tense]

(95) It **is called Miami**. 3, 19 →se **llama Miami**. 2, 15

-78.43 [-1 lexical word] (1 shift)

1. Fixed shift (-1): is called (2) →se llama (1) [confirmed with DeepL]

(132) What **changes come will depend** upon the **Cuban people.** 6, 52 →Lo que **cambie dependerá** del **pueblo cubano.** 4, 42

-71.15 [-2 lexical words] (2 shifts)

1. Free shift (-1): changes-N come-V (2) →lo que cambie-V (1) [DeepL:*Los cambios que se produzcan…*]
2. Fixed shift (-1): will-V →Ø [future tense]

(187) that **is** a **measure** of our **progress** as a **democracy**. 4, 49 →Esa **es** la **medida** de nuestro **progreso**. 3, 37

-70.40 [-1 lexical word] (1 shift)

1. Free shift (-1): democracy-N (1) →Ø (0)

(174) I **am not saying** this **is easy.** 5, 29 →**No digo** que **sea fácil**. 4, 22

-56.31 [-1 lexical words] (1 shifts)

1. Fixed shift (-1): am saying (2) →digo (1)

(173) That **was** because of the **freedoms** that **were afforded** in the **United States** that we **were able** to **bring** about-ADP **change.** 10, 113 →Eso **fue** por las **libertades otorgadas** en los **Estado Unidos** que **pudimos traer** el **cambio**. 8, 86

-55.67 [-2 lexical words] (2 shifts)

1. Fixed shift (-1): were afforded (2) →otorgar (1) [confirmed with DeepL]
2. Fixed shift (-1): were able (2) →pudimos (1) [confirmed with DeepL]

(186) Who **would have believed** that back-ADV in **1959**? 5, 44 →¿Quién **habría apostado** por eso en **1959**? 3, 39

-51.14 [-2 lexical words] (2 shifts)

1. Fixed shift (-1): would have (2) →habría (1) [conditional tense]
2. Fixed shift (-1): back-ADV →Ø

(178) That'**s**-*PROPN/POSS **how**-ADV we **made enormous gains** in **women's rights** and **gay rights**. 10, 68 →**Es como**-*V/ADV **hicimos grandes avances** en los **derechos** de las **mujeres** y de los **homosexuales**. 8, 85

-45.67 [-2 lexical words] (2 shifts)

1. False shift POS (-1): 's-*PROPN/POSS →Ø
2. Free shift (-1): rights-N and …rights-N (2) →derecheos-N y …(1) [DeepL: *…los derechos de las mujeres y los derechos de los gays.*]

(228) that **is why**-ADV their **heartache is so great.** 6, 40 →**Es** por-ADP eso-PRON que la **pena** en sus **corazones es tan grande**. 5, 51

-42.06 [-1 lexical word] (3 shifts)

1. Fixed shift (-1): why-ADV (1) →por-ADP eso-PRON (0)
2. Antinomic fixed shift (+1): heartache-N (1) →pena-N en sus corazones-N (2)
3. Free shift (-1): is-V (1) →Ø (0) [DeepL: *…su dolor de corazón es tan grande.*]

(76) There-ADV **is one**-NUM **simple answer**: 5, 27 →La **respuesta es sencilla**: 3, 25

-36.39 [-2 lexical words] (2 shifts)

1. Free shift (-1): there-ADV →Ø [DeepL: *…Hay una respuesta simple.*]
2. Fixed shift (-1): one-NUM →Ø or una-DET

(188) So **here**-ADV **is** my **message** to the **Cuban government** and the **Cuban people**: 7, 67 →Este es mi **mensaje para***V/ADP el **gobierno** y **pueblo** de **Cuba**: 6, 53

-35.39 [-1 lexical word] (3 shifts)

1. Fixed shift (-1): here-ADV →Ø

2. Antinomic false shift POS (+1): to-ADP →para, parir *V/ADP
3. Free shift (-1): *Cuban government and the Cuban people* →*el gobierno y pueblo de Cuba* [confirmed with DeepL]

(10) **Muchas**-*PROPN/ADJ **gracias.** 2, 15 →Muchas **gracias.** 1, 15

-33.23 [-1 lexical word] (1 shift)

1. False shift POS (-1): muchas-*PROPN/ADJ →muchas *DET/ADJ [target expression used in source text]

(161) We **do have too**-ADV **much**-ADJ **money** in **American politics.** 7, 47 →**Sí**-ADV que **hay demasiado dinero** en la **política estadounidense.** 6, 58

-31.80 [-1 lexical word] (1 shift)

1. Fixed shift (-1): too-ADV much-ADJ →demasiado-ADV [confirmed with DeepL]

(56) For all-DET of our **differences**, the **Cuban** and **American people share common values** in their **own**-ADJ **lives**. 9, 97 →Con todas nuestras **diferencias**, el **pueblo estadounidense** y el **pueblo cubano comparten** los mismos-*DET/ADJ **valores** en sus propias-*DET/ADJ **vidas.** 8, 126

-31.72 [-1 lexical word] (3 shifts)

1. Antinomic free shift (+1): people (1) →pueblo y pueblo (2)
2. False shift POS (-1): common values →mismos-*DET/ADJ valores-N
3. False shift POS (-1): own-ADJ lives →proprias-*DET/ADJ vidas N

(63) the **United States is** a **multi**-ADJ **party democracy**. 7, 45 →**Estados Unidos es** una **democracia** de **múltiples partidos**. 6, 55

-30.36 [-1 lexical word] (2 shifts)

1. False shift POS (-1): –ADJ [-1] →Ø [0]
2. Undetected fixed shifts (+1): multi-party-N (*3/1) →múltiples-ADJ partidos-N (2)

(194) **Many**-ADJ **suggested** that I **come here**-ADV and **ask** the **people** of **Cuba** to **tear something**-N **down**-ADV – but I **am appealing** to the **young people** of **Cuba** who **will lift something**-N up, **build something**-N **new**. 21, 180 →Muchos-PRON **han sugerido** que **vengo aquí**-ADV **para**-*AUX/ADP **pedir**

al **pueblo cubano** que **destruya** algo-PRON; pero yo me **dirijo** a los **jóvenes** de **Cuba** quienes **alzarán** y **construirán** algo-PRON **nuevo.** 15, 163

-29.80 [-6 lexical words] (10 shifts)

1. False shift POS (-1): many *ADJ/PRON →muchos-PRON
2. Antinomic free shift (+1): suggested →han sugerido [DeepL: *Muchos me sugirieron…*]
3. Antinomic false shift POS (+1): and-CONJ →para, parir *AUX/ADP
4. Fixed shift (-1): something-N →algo-PRON [confirmed with DeepL: … *que derribara algo…*]
5. Fixed shift (-1): tear-V down-ADV →destruya-V
6. Fixed shift (-1): am-V →yo-PRON
7. Fixed shift (-1): young people →jóvenes
8. Fixed shift (-1): will →Ø [future tense]
9. Fixed shift (-1): something-N →algo-PRON
10. Fixed shift (-1): something-N →Ø (algo-PRON)

(207) It **gives everyone**-*N/PRON in this **hemisphere hope**. 4, 42 →Le **brinda esperanza** a todos-PRON en este **hemisferio.** 3, 47

-29.14 [-1 lexical word] (1 shift)

1. False shift POS (-1): everyone-*N/PRON →todos-PRON

We found 46 information shifts in sample A, with 12 false shifts POS (due to various POS tagging errors), 24 fixed shifts, and 10 free shifts. For all types of shifts, 5 antinomic shifts were found. We will not go into the details of the analysis but provide to the reader a brief survey of what we can deduct from the data collected. A more detailed review of these results is of high interest for translation studies and training but deserves to be addressed in a separate publication. First, this sample contains mostly fixed information shifts due to source language constraints such as verb compositions (modals, active/passive (mandatory)), transformations for verbal constructions exclusive to one language (*is called* translated by *se llama* in pair 95, or *there be* translated by *haber* in pair 258, for example), some peculiar uses of adverbs in English that may be omitted in Spanish or are translated by a preposition, and some English-Spanish POS tagging difference regarding functional words such as pronouns (in pair 194, the pronoun *something* is analyzed as

a noun and translated with the pronoun *algo*, for example). Second, regarding the free information shifts found in sample A, those are far fewer in number. Some can be explained with the traditional concepts of "concentration" or "concision" used in translation studies. Most of them seem to be due to some sort of Spanish grammatical "flexibility" or "freedom" which affords the translation process highly acceptable syntactic reductions of redundant information in the source language such as non-repetitions of generic nouns in noun phrase coordination like in the pair *Cuban government and Cuban people*, which is reducted to *el gobierno y pueblo de Cuba*. We found one characteristic omission of the notion of democracy in pair 187 that is due to the different political systems of reference between the United States and Cuba, but that illustrates very well one political issue between the two countries.

4.2 Sample B annotations – most positive heteromorphic pairs

This section contains the manual analysis and annotations of sample B segments for the classification of information shifts observed in the most positive heteromorphic segment pairs. Translation pairs are presented from the highest positive weighted Euclidean distance (313.26) to the lowest positive weighted Euclidean distance (52.35).

(144) **Not everybody**-N **agrees** with-ADP me on this. 3, 37 →**No** todo-DET el **mundo está**-AUX de **acuerdo**-*V/N **conmigo**-*AUX/ADP **sobre**-*V/ADP esto. 6, 52

313.26 [+3 lexical words] (3 shifts)

1. Fixed shift (+1): agree-V (1) →está-V de acuerdo-*V/N (2)
2. False shift POS (+1): with-ADP →conmigo-*AUX/ADP
3. False shift POS (+1): on-ADP →sobre, sobrar-*V/ADP

(177) that **is how**-ADV we **got health care** for **more**-ADJ of our **people.** 7, 54 →**Es como**-*AUX/CONJ **conseguimos servicios** de **salud para**-*V/ADP una **mayor cantidad** de **personas**-***V/N** del **país**. 10, 84

165.63 [+3 lexical words] (4 shifts)

1. Antinomic false shift POS (-1): how-ADV →como-*AUX/CONJ
2. False shift POS (+1): for-ADP →para, parir-*V/ADP
3. Free shift (+1): more-ADJ (1) →mayor-ADJ cantidad-N (2)

4. Free shift (+1): people-N (1) →personas-*V/N del país-N (2) [DeepL: … *para más de nuestra gente.*]

(191) And we – like-ADP every-DET **country** – **need** the **space** that **democracy gives** us to **change**. 6, 81 →Y nosotros – **al**-*V/ADP+DET **igual**-*ADV/N que todos-DET los **países** – **necesitamos** el **espacio** que la **democracia** nos-*ADV/PRON **da**-AUX **para**-*AUX/ADP **cambiar.** 10, 104

152.58 [+4 lexical words] (4 shifts)

1. Fixed shift (+1): like-ADP (1) →al-*V/ADP+DET igual-*ADV/N (loc adv) (*2/1)
2. False shift POS (+1): Ø →al*V/ADP+DET
3. False shift POS (+1): us-PRON →nos, no-*ADV/PRON
4. False shift POS (+1): to-ADP →para, parir-*AUX/ADP

(46) We **have welcomed** both-DET **immigrants** who **came** a **great distance** to **start new lives** in the **Americas**. 10, 94 →**Ambos**-NUM **hemos abierto** nuestras **puertas** a **inmigrantes** que **recorrieron grandes distancias**-*AUX/N **para**-*AUX/ADP **empezar vidas nuevas** en el **continente americano**. 14, 139

134.12 [+4 lexical words] (4 shifts)

1. False shift POS (+1): both-DET →ambos-NUM
2. Free shift (+1): have welcomed (2) →hemos abierto puertas (3) [DeepL: *hemos acogido a ambos inmigrantes …*]
3. False shift POS (+1): to-ADP →para, parir-*AUX/ADP
4. Free shift (+1): Americas (1) →continente americano (2) [DeepL : *…en las Américas.]*

(157) I **welcome** this **open debate** and **dialogue**. 4, 40 →**estoy dispuesto** a **tener** este **debate** y **diálogo abierto**. 6, 54

134.12 [+2 lexical words] (1 shift)

1. Free shift (+2): welcome-V (1) →estoy-V dispuesto-ADJ a tener-V (3) [DeepL: *Me complace este …*]

(183) You **can see** that in the **election going** on **back**-ADV **home**-*ADV/N. 6, 52 →Lo **podemos apreciar** en las **elecciones** que **están** en **curso ahora**-ADV **mismo**-ADJ en mi **país.** 8, 80

123.81 [+2 lexical words] (2 shifts)

1. Free shift (+3): going-V (1) →están-V en curso-N ahora-ADV mismo-ADJ (4)
2. Antinomic free shift (-1): back-ADV home-*ADV/N (2) →en mi país (1) [DeepL : *... que se están llevando a cabo*-LOC-ADV *en casa.*]

(38) I **want** to **be clear**: 3, 19 →**Quiero dejar** una **cosa clara**: 4, 28

111.83 [+1 lexical word] (1 shift)

1. Free shift (+1): be clear-V (2) →dejar-V una cosa-N clara-ADJ (3) [DeepL : *Quiero ser claro:*]

(145) **Not everybody**-N **agrees** with the **American people** on this. 5, 54 →**No** todo el **mundo está** de **acuerdo** con el **pueblo estadounidense sobre**-*V/ADP esto. 7, 73

106.34 [+2 lexical words] (2 shifts)

1. Fixed shift (+1): agree-V (1) →está-V de acuerdo-N (2)
2. False shift POS (+1): on-ADP →sobre, sobrar-*V/ADP

(113) It **is** an **outdated burden** on the **Cuban people**. 5, 45 →**Es** una **carga anticuada** que **lleva** a **cuestas** el **pueblo cubano**. 7, 60

101.61 [+2 lexical words] (1 shift)

1. Free shift (+2): on-ADP (0) →llevar-V a cuestas-N (2) [DeepL : *... anticuada para el pueblo cubano.*]

(122) It **is** up to you. 1, 16 →Eso **es cosa** suya. 2, 17

83.47 [+1 lexical word] (1 shift)

1. Free shift (+1): is-V up (1) →es-V cosa-N (2) [DeepL : *Depende de usted.*]

(41) But before-ADP I **discuss** those **issues**, we **also**-**ADV need** to **recognize how**-**ADV much**-**ADJ** we **share**. 8, 79 →Pero **antes**-ADV de **hablar sobre**-*V/ADP esos **temas**, **también**-ADV **es** nuestro **deber reconocer cuánto**-ADJ **tenemos** en **común**. 11, 98

71.63 [+3 lexical words] (5 shifts)

1. Fixed shift (+1): before-ADP →antes-ADV
2. False shift POS (+1): Ø →sobre, sobrar-*V/ADP

3. Free shift (+1): we need-V (1) →es-V nuestro deber-N (2) [DeepL: … *también necesitamos reconocer cuánto compartimos.*]
4. Antinomic fixed shift (-1): how-ADV much-ADV (2) →cuánto-ADJ (1)
5. Free shift (+1): share-V (1) →tenemos-V en común-N (2)

(27) The **blue waters** beneath-ADP **Air Force One once**-ADV **carried American battleships** to this **island** – to **liberate**, but **also**-ADV to **exert control** over **Cuba.** 15, 139 →Las **aguas azuladas bajo**-*V/ADJ **Air Force One transportaron** en su **día** los **barcos** de **batalla estadounidenses** hasta esta **isla**, **para**-*AUX/ADP **liberar** pero-CONJ **también**-ADV **para**-*AUX/ADP **ejercer control sobre**-*V/ADP **Cuba**. 20, 175

67.31 [+5 lexical words] (5 shifts)

1. Fixed shift (+1): beneath-ADP →baja, bajar-*V/ADJ [confirmed with DeepL: … *bajo* …]
2. False shift POS (+1): to-ADP →para, parir-*AUX/ADP
3. False shift POS (+1): to-ADP →para, parir-*AUX/ADP
4. False shift POS (+1): over-ADP →sobre, sobrar-*V/ADP
5. Free shift (+1): battleships-N (1) →barcos-N de batalla-N (2) [DeepL: … *acorazados* …]

(190) **Not** because **American**-ADJ **democracy is perfect**, but **precisely** because we **are not**. 8, 76 →**No** porque **pienso** que la **democracia** en **Estados**-PROPN **Unidos**-PROPN **sea perfecta**, sino **precisamente** porque **no** lo **somos**. 10, 104

66.67 [+2 lexical words] (2 shifts)

1. Free shift (+1): Ø →pienso-V
2. Free shift (+1): American-ADJ (1) →Estados-PROPN Unidos-PROPN (2) [DeepL: *No porque la democracia americana sea perfecta* …]

(155) He **has** a **much**-ADV **longer**-ADJ **list-N**. 4, 26 →Él **tiene** una **mucho**-ADV **más**-ADV **lista-N larga**-ADJ. 5, 35

63.25 [+1 lexical word] (1 shift)

1. Fixed shift (+1): much-ADV (1) →mucho-ADV más-ADV (2)

(128) Before **1959**, some **Americans saw Cuba** as-ADP **something**-N to **exploit**, **ignored poverty**, **enabled corruption**. 10, 98 →Y desde **1959**, algunos-PRON **estadounidenses veían Cuba como**-V un **lugar** del que se **podían aprovechar**, **ignoraron** la **pobreza** y **permitieron** la **corrupción**. 12, 142

61.40 [+2 lexical words] (2 shifts)

1. False shift POS (+1): as-ADP →como, comer-*V/ADP
2. Free shift (+1): something-N to exploit-V (2) →lugar-N del que se podían-V aprovechar-V (3) [DeepL: *... como algo para explotar ...*]

(129) And since-ADP **1959**, we **have been shadow-boxers** in this **battle** of **geopolitics** and **personalities**. 8, 91 →Desde-ADP **1959, hemos sido como**-*V/ADP **boxeadores** con un **contrincante**-*ADV/N **imaginario** en esta **batalla** de **geopolítica** y **personalidades**. 10, 118

55.43 [+2 lexical words] (2 shifts)

1. False shift POS (+1): Ø →como, comer-*V/ADP
2. Free shift (+1): shadow-boxers-N (2) →boxeadores-N con un contrincante-*ADV/N imaginario-ADJ (3) [DeepL: *... hemos sido boxeadores en la sombra en esta batalla ...*]

(45) Like-ADP the **United States**, the **Cuban people can trace** their **heritage** to both **slaves** and **slave-owners**. 10, 98 →Al **igual**-*ADJ/N que en **Estados Unidos**, el **pueblo cubano puede encontrar** sus **orígenes tanto**-*ADV/PRON en los **esclavos**-*PRON/N **como**-*V/CCONJ en los **dueños** de los **esclavos**-ADJ. 12, 135

53.21 [+2 lexical words] (5 shifts)

1. Fixed shift (+1): like-ADP (0) →Al igual-*ADJ/N (1)
2. False shift POS (+1): both-DET →tanto-*ADV/PRON
3. Antinomic false shift POS (-1): slaves-N →esclavos-*PRON/N
4. False shift POS (+1): and-CCONJ →como, comer-*V/CCONJ
5. Undetected fixed shift [+1]: slave-owners-N (*2/1) →dueños-N de los esclavos-N (2)

(91) **Hope** that **is rooted** in the **future** that you-PRON **can choose** and that you-PRON **can shape**, and that you-PRON **can build** for your **country**. 11, 118 →**Esperanza** que **tiene** una **base** en el **futuro** que **ustedes**-*N/PRON **pueden**

elegir; que **ustedes**-*N/PRON **pueden moldear**; que **ustedes**-*N/PRON **pueden construir para**-*V/ADP su **país.** 15, 139

53.15 [+4 lexical words] (4 shifts)

1. False shift POS (+1): you-PRON →ustedes, vosotros-*N/PRON
2. False shift POS (+1): you-PRON →ustedes, vosotros-*N/PRON
3. False shift POS (+1): you-PRON →ustedes, vosotros-*N/PRON
4. False shift POS (+1): for-ADP →para, parir-*V/ADP

(70) We **have begun initiatives** to **cooperate** on **health** and **agriculture**, **education** and **law enforcement**. 9, 96 →**Hemos lanzado iniciativas para**-*V/ADP **cooperar** en **temas**-*V/N de **salud** y **agricultura**, **educación** y **autoridades** del **orden público**. 12, 115

53.13 [+3 lexical words] (3 shifts)

1. False shift POS (+1): to-ADP →para, parir-*V/ADP
2. False shift POS (+1): Ø →temas, temer-*V/N
3. Free shift (+1): law enforcement (2) →autoridades del orden público (3) [DeepL: *… la aplicación de la ley.*]

(25) **Havana is only 90 miles** from **Florida**, but to **get here**-ADV we **had** to **travel** a **great distance** – over **barriers** of **history** and **ideology**; 15, 129 →La **Habana** se **encuentra tan**-*N/ADV **solo**-*N/ADV a **90 millas** de **Florida**, pero **para**-*AUX/ADP **llegar** hasta **aquí tuvimos** que **recorrer** una **gran distancia**: **derribar**-V las **barreras** de la **historia** y la **ideología**; 18, 177

52.56 [+3 lexical words] (3 shifts)

1. Free shift (+1): only-ADV →tan-*N/ADV solo-*N/ADV [DeepL: *La Habana está a sólo 90 millas …*]
2. False shift POS (+1): to-ADP →para, parir-*AUX/ADP
3. Free shift (+1): over-ADP →derribar-V) [DeepL: *… por encima de las barreras de la …*]

(135) But **having removed** the **shadow** of **history** from our **relationship**, I **must speak honestly** about the **things** that I **believe** – the **things** that we, as **Americans**, **believe**. 13, 63 →Pero **ahora**-ADV que **hemos quitado** la **sombra** de la **historia** de nuestra **relación**, **debo hablar honestamente sobre**-*V/ADP las **cosas** en las que yo **creo** –-*AUX/PUNCT las cosas en las que nosotros, **como**-*V/ADP **estadounidenses**, **creemos**. 17, 198

52.35 [+4 lexical words] (4 shifts)

1. Free shift (+1): Ø ⟶ahora-ADV [DeepL: *Pero habiendo eliminado la sombra …*]
2. False shift POS (+1): about-ADP ⟶sobre, sobrar-*V/ADP
3. False shift POS (+1): – ⟶–-*AUX/PUNCT
4. False shift POS (+1): as-ADP ⟶como, comer-*V/ADP

Sample B contains 59 information shifts in total with 30 false shifts POS, 8 fixed shifts, and 21 free shifts. For all types of shifts, 4 antinomic shifts and 1 undetected shift were found. By comparison with sample A, we can see that free shifts are more than two times the number of fixed shifts, which were two times more numerous than the former in sample A. Here is a brief overview of the trend we can observe from the annotations. A lot of the numerous positive free shifts seem to be associated with some form of mandatory and translation-inherent explicitations (see Blum-Kulka (1986) who proposed the explicitation hypothesis, as well as the work of Becher (2010; 2011) who rejected the hypothesis and the more recent synthesis article by Murtisari (2016) on the concept of explicitation in translation studies). One example is the periphrastic translation of *shadow-boxers* with *boxeadores con un contrincante imaginario* in pair 129 or the creative translation of *battleship* with *barcos de batallas* in pair 27. A good example of "political" explicitness that tends to reduce a statement is found in pair 190 when President Obama state that American democracy could be seen as "perfect" (even though he clearly states that this is not the case). The translation makes explicit that the statement is its own way of thinking by adding the verb phrase *pienso que*. As regards the eight positive fixed shifts, these are mostly the opposite operations that were described in the analysis of sample A negative fixed shifts, such as the addition of an adverb or the translation of a preposition by an adverb as in pair 41. These last data validate the existence of mandatory explicitations and implicitations processes that are symmetrical and dependent from syntactic and lexical structures of languages (see Klaudy 2011).

4.3 Sample C annotations – most isomorphic and less heteromorphic pairs

This section contains the manual analysis and annotations of sample C segments for the classification of information shifts observed in the most isomorphic and least heteromorphic segment pairs. Translation pairs all have near-zero weighted

Euclidean distance and are presented from the highest negative weighted Euclidean distance (-0.42) to the highest positive weighted Euclidean distance (1.29).

(176) But **democracy is** the **way** that we **solve** them. 4, 44 →Pero la **democracia es** la **forma** de **cambiarlos**. 4, 45

−0.42 [0 difference in lexical words] (0 shift)

(104) **Look** at **Papito Valladeres**, a **barber**, whose-DET **success allowed** him to **improve conditions** in his **neighborhood**. 9, 105 →Miren a Papito Valladeres, un barbero, cuyo-PRON **éxito** le **permitió mejorar** las **condiciones** en su **vecindario**. 9, 103

−0.36 [0 difference in lexical words] (0 shifts)

(133) We **will not impose** our **political** or **economic system** on you. 6, 59 →**No vamos** a **imponerles** nuestro **sistema político** ni económico. 6, 60

−0.31 [0 difference in lexical words] (0 shifts)

(18) We **will**-V **do** whatever **is necessary** to **support** our **friend** and **ally**, **Belgium**, in **bringing** to **justice** those who **are responsible**. 12, 123 →**Haremos** lo que **sea necesario para**-*AUX/ADP **apoyar** a nuestra **amiga** y **aliada**, **Bélgica, para**-*AUX/ADP **ajusticiar**-V a aquellos que **sean responsables**. 12, 125

−0.30 [0 difference in lexical words] (4 shifts)

1. Fixed shift (-1): will-V do-V (2) →haremos-V (1)
2. False shift POS (+1): to-ADP →para-*AUX/ADP
3. False shift POS (+1): in-ADP →para-*AUX/ADP
4. Free shift (-1): bringing-V to justice-N (2) →ajusticiar-V (1) [DeepL: …, *para llevar a la justicia a los responsables.*]

(234) And I **have come here**-ADV – I **have traveled** this **distance** – on a **bridge** that **was built** by **Cubans** on both-DET **sides** of the **Florida Straits**. 13, 131 →Y **he venido aquí**-ADV – he-*ADP **viajado** esta **distancia** – **sobre**-*V/ADP un **puente construido**-ADJ por los **cubanos** a **ambos**-NUM **lados** del **Estrecho** de la **Florida**. 13, 129

−0.29 [0 difference in lexical words] (4 shifts)

1. False shift POS (-1): have-AUX →-he-*ADP/AUX
2. False shift POS s(+1): on-ADP →sobre, sobrars-*V/ADP

3. Free shift (-1): was-AUX built-V (2) →construido-ADJ (1) [DeepL: *… en un puente que fue construido por …*]
4. Fixed shift (+1): both-DET →ambos-NUM [DeepL: *… a ambos lados del Estrecho de Florida.*]

(216) I **know** that **many**-ADJ of the **issues** that I **have talked** about **lack** the **drama** of the **past**. 8, 83 →**Sé** que **muchos**-ADV de los **problemas** de los que **he hablado carecen** del **drama** del **pasado**. 8, 82

−0.23 [0 difference in lexical words] (0 shift)

(208) We **took different journeys** to our **support** for the **people** of **South Africa** in **ending apartheid**. 9, 93 →**Tomamos diferentes pasos** en nuestro **apoyo** al **pueblo** de **Sudáfrica para**-*AUX/ADP **acabar** con el **apartheid**. 9, 94

−0.20 [0 difference in lexical words] (2 shifts)

1. Fixed shift (-1): South Africa (2) →Sudáfrica (1)
2. False shift POS (+1): in ADP →para-*AUX/ADP

(201) But-CONJ no-DET one **should deny** the **service** that **thousands** of **Cuban doctors have delivered** for the **poor** and **suffering**. 10, 109 →Pero-CONJ nadie-PRON **debe negar** el **servicio** que **miles** de **médicos cubanos han prestado** a los **pobres** y a los que **sufren**. 10, 108

−0.17 [0 difference in lexical words] (0 shift)

(84) **Creo** en el **pueblo Cubano**. 3, 25 →**Creo** en el **pueblo cubano**. 3, 25

0,00 [0 difference in lexical words] (0 shift)

(86) This **is not just**-ADV a **policy** of **normalizing relations** with the **Cuban government**. 8, 77 →Esto **no es solo**-ADJ una **política** de **normalizar relaciones** con el **gobierno Cubano**; 8, 77

0.00 [0 difference in lexical words] (0 shift)

(102) **Look** at **Sandra Lidice Aldama**, who **chose** to **start** a **small business**. 8, 66 →**Miren** a **Sandra Lidice Aldama**, que **eligió abrir** un **pequeño negocio**. 8, 66

0.00 [0 difference in lexical words] (0 shift)

1. Note: This is the central pair of isomorphic segments, being exactly in the middle of two other isomorphic segment pairs.

(151) **the death penalty**; 2, 18 →la **pena** de **muerte**; 2, 18

0.00 [0 difference in lexical words] (0 shift)

(66) **Cuba has emphasized** the **role** and **rights** of the **state**; 6, 53 →**Cuba ha reforzado** el **papel** y los **derechos** del **estado**; 6, 53

0.00 [0 difference in lexical words] (0 shift)

(197) And **given**-V your **commitment** to **Cuba's**-*PROPN/POSS **sovereignty** and **self-determination**, I **am also**-ADV **confident** that you **need not fear** the **different voices** of the **Cuban people** – and their **capacity** to **speak**, and **assemble**, and **vote** for their **leaders**. 22, 229 →**Teniendo** en **cuenta**-*V/N su **compromiso** con la **soberanía** y la **autodeterminación** de **Cuba**, **también**-ADV **estoy seguro** de que **no tiene** que **temer** las **diferentes voces** del **pueblo cubano** – y-*ADJ/CCONJ su **capacidad** par-*V/ADP **hablar**, y **reunirse**, y **votar** por sus **líderes**. 23, 233

0.42 [+1 lexical word] (5 shifts)

1. Free shift (+1): given-V (1) →teniendo-V en cuenta-*V/N (2) [DeepL: Y dado su compromiso con la soberanía …]
2. False shift POS (-1): 's-*PROPN/POSS →de-ADP
3. False shift POS (-1): self-determination (2) →autoderminación (1)
4. False shift POS (+1): – and-ADP →– y-*ADJ/CCONJ
5. False shift POS (+1): to-ADP →para-*V/ADP

(236) And I **know how**-ADV they **have suffered more**-ADJ than the **pain** of **exile** – they also **know** what it **is** like-ADP to **be** an **outsider**, and to **struggle**, and to **work harder** to **make sure** their **children can reach higher** in **America**. 22, 207 →Y **sé** que **han sufrido más**-ADV que el **dolor** del **exilio**: **saben** lo que se **siente** al **ser** un **extraño**, al **luchar**, al **trabajar** más-ADV duro-*AUX/ADJ para-*V/ADP **asegurarse** de que sus **hijos puedan llegar más**-ADV **lejos**-ADV en los **Estados Unidos**. 23, 203

0.47 [1 lexical word] (5 shifts)

1. Fixed shift (-1): how-ADV (1) →que-SCONJ (0)
2. Fixed shift (+1): harder-ADJ (1) →más-ADV duro-*AUX/ADJ (2)
3. False shift (+1): to-ADP →para-*V/ADP
4. Fixed shift (+1): higher-ADJ (1) →más-ADV lejos-ADV (2)

5. Fixed shift (-1): make-V sure-ADJ ⟶asegurarse-V
6. Free shift (+1): America-PROPN ⟶Estados-PROPN Unidos-PROPN [DeepL: *…más alto en América.*]

(162) But, in **America**, it **is** still-ADV possible-ADJ for **somebody**-*N/PRON like-ADP me – a **child** who **was raised** by a **single mom**, a **child** of **mixed race** who **did not have** a **lot** of **money** – to **pursue** and **achieve** the **highest**-ADJ **office** in the **land**. 23, 212 ⟶Pero en **EEUU**, **todavía**-ADV **es posible** que alguien-PRON **como**-*V yo, un **niño** que **fue criado** por una **madre soltera**, un **niño** de **raza mixta** que **no tenía** mucho-DET **dinero**, **pueda-V ir**-V **atrás**-ADV de y **conseguir** el **cargo más alto** del **país**. 24, 206

0.63 [1 lexical word] (6 shifts)

1. False shift POS (-1): somebody-*N/PRON ⟶alguien-PRON
2. False shift POS (+1): like-ADP ⟶como-*V/ADP
3. Fixed shift (-1): did not have (3) ⟶no tenía (2)
4. False shift POS (-1): a lot-N ⟶mucho-*DET/ADJ
5. Free shift (+2): pursue-V (1) ⟶pueda-AUX ir-V atrás-ADV de (3) [DeepL: *…persiga y logre el cargo más alto de la tierra.*]
6. Fixed shift (+1): highest-ADJ ⟶más-ADV alto-ADJ

(209) But **President Castro** and I **could** both-DET **be there-ADV** in **Johannesburg** to-ADP **pay tribute** to the **legacy** of the **great Nelson Mandela**. 12, 120 ⟶Pero el **presidente Castro** y yo **pudimos estar allí**-ADV en **Johannesburgo para**-*AUX/ADP **rendir homenaje** al **legado** de **gran Nelson Mandela**. 13, 121

0.69 [1 lexical word] (1 shift)

1. False shift POS (+1): to-ADP ⟶para-*AUX/ADP

(136) As **Marti said**, "-*PROPN/PUNCT **Liberty is** the **right** of every-DET **man** to **be honest**, to **think** and to **speak** without **hypocrisy**."-PUNCT 12, 105 ⟶Como **dijo Martí**: "-*PROPN/PUNCT La **libertad es** el **derecho** de todo-DET **hombre** a **ser honesto**, **pensar** y **hablar** sin **hipocresía**"-*N/PUNCT. 13, 106

0.74 [1 lexical word] (1 shift)

1. False shift POS (+1): hypcocrisy.-PUNCT ⟶hipocresía-*N/PUNCT

(214) From the **beginning** of my **time**-N in **office**, I **have urged** the **people** of the **Americas** to **leave** behind-ADP the **ideological battles** of the **past**. 11, 133 →Desde el **inicio** de mi **mandato**, **he instado** a los **pueblos** del **continente americano a dejar atrás**-ADV las **batallas ideológicas** del **pasado**. 12, 131

1.16 [1 lexical word] (3 shifts)

1. Free shift (-1): time-N in office-N →mandato-N [DeepL: *Desde el principio de mi tiempo en la oficina …*]
2. Free shift (+1): Americas (1) →continente americano (2) [DeepL: *He instado a los pueblos de América …*]
3. Fixed shift (+1): leave-V behind-ADP (1) →dejar-V atrás-ADV (2)

(210) And in **examining** his **life** and his **words**, I **am sure** we both-DET **realize** we **have more**-ADJ **work** to **do** to **promote equality** in our **own**-ADJ **countries** – to-ADP **reduce discrimination based** on **race** in our **own**-ADJ **countries**. 20, 195 →Y al **examinar** su **vida** y sus **palabras**, **estoy seguro** de que **ambos**-NUM **nos**-*ADV/PRON **damos**-V **cuenta**-N de que **tenemos mucho trabajo** por **hacer** –-*PROPN/PUNCT **para**-*AUX/ADP **reducir** la **discriminación basada** en la **raza** en **ambos países**. 21, 187

1.18 [1 lexical word] (6 shifts)

1. Fixed shift (+1): both-DET →ambos-NUM
2. False shift POS (+1): we-PRON →nos-*ADV/PRON
3. Free shift (-4): to promote equality in our own countries →Ø [DeepL: *… para promover la igualdad en nuestros propios países.*]
4. Fixed shift (+1): realize-V (1) →damos-V cuenta-N (2)
5. False shift POS (+1): –-PUNCT →–-*PROPN/PUNCT
6. False shift POS (+1): to-ADP →para-*AUX/ADP

(239) "-*PROPN/PUNCT You **recognized** me, but I **did**-V **not recognize** you," -PUNCT **Gloria said** after-ADP she **embraced** her **sibling**. 9, 93 →"--*PROPN/PUNCT Tú me **reconociste**, pero yo-PRON **no** te **reconocí**"-*PROPN/PUNCT, le **dijo Gloria** a su **hermana después**-ADV de **abrazarla.** 10, 94

1.29 [+1 lexical words] (3 shifts)

1. Fixed shift (-1): did V →yo-PRON
2. Fixed shift (+1): after-ADP (0) →después-ADV (1)
3. False shift (+1): "-PUNCT (0) →*PROPN/PUNCT (1)

In this sample, antinomic shifts may be positive or negative since the segment pairs all have zero or near zero weighted Euclidean distance while almost half of them are negatively close to zero or positively close to zero. Sample C has 40 information shifts in total, with 18 false shifts POS, 14 fixed shifts, 8 free shifts, and 19 antinomic shifts. This sample contains the highest number of antinomic information shifts among the three samples annotated. Because the TPR of most segments is neutral (=1.0), the number of antinomic shifts is doubled as it is the case for segments 18, 234, and 208 which account for 10 antinomic shifts. The other 9 antinomic shifts appear in lengthy segment pairs having a small positive TPR value. Their number could be reduced if the segmentation of the text could have a finer or smaller granularity to the level at least of propositions. This is a development that would enhance the efficiency of the empirical screening method of information shifts described here.

We can observe for sample C annotations that segments having zero weighted Euclidean distance contain no information shift at all. The number of these segments is small (5), but it's worth noting the efficiency of the method for screening pairs having no information shift at all. We can also note that some particular lengthy segment pairs have a lot of information shifts while all the other short segment pairs have zero information shifts. The average length of the 11 segment pairs having at least one information shift is 144 characters while the average length of the 10 segment pairs having zero information shift is less than half of this amount with 63.8 characters. In the case of mostly isomorphic segment pairs, the short length in characters seems to be predictive of the absence of information shifts. This correlation hypothesis needs to be further tested and set for different corpora.

5 Conclusion

We described in detail an empirical method for screening segment pairs in parallel corpora for informational translation shifts. Our manual analysis of the three samples A B and C of parallel pairs screened with our method confirm our hypothesis that heteromorphic segment pairs, as opposed to isomorphic ones, contain higher numbers of informational translation shifts. These tendencies can be observed with the number of information shifts that were detected in the most negative (46) and positive (59) heteromorphic segment pairs, compared to the number of information shifts present (40) in more isomorphic pairs (among which 5 pairs having a weighted Euclidean distance of exactly zero contained no information shift). If we discard the false information shifts which are erroneous,

the observation is perhaps strengthened with a lower volume of information shifts (22) in sample C by comparison with 34 in sample A and 29 in sample B. Regarding our hypothesis for information shift screening, another criterion that need to be taken into account is that we found that the length of mostly isomorphic pairs seems to be predictive of the presence or absence of information shift. The discovery of this correlation for sample C pairs needs to be further tested with other corpora and against heteromorphic segment pairs.

What was also considered surprising in the annotations of the two most heteromorphic samples is that negative heteromorphic segment pairs tend to contain much more (mandatory) fixed shifts than free shifts while the exact opposite holds for positive heteromorphic segment pairs. This could be in line with the explicitation hypothesis in translation which can be viewed as a tendency to add content in the target segments in translation (thus creating an information asymmetry) by giving more details and explanations than what is given in the source text (to make sure for instance that the content is well understood or clear for the intended audience). Further studies and progress on the empirical methods developed herein are needed to shed light on this result.

A better knowledge of the origin, the cause and the impact of fixed information shifts are essential for a better knowledge of language constraints in translation (in contrastive phraseology, translation difficulties, and their idiosyncratic solutions) while the study of free information shifts should shed light on cognitive issues in translation operations (errors, individual and cultural biases). The manual examination and categorization of 145 informational shifts have shown that fixed and free shifts are relevant categories for the study of these phenomena. In order to reduce false shifts (false positives) and undetected shifts (false negatives), new POS tagging models and methods for English and for other major languages would need to be developed. The situation was found to be worse for the Spanish language, where many significant errors in POS tagging were found, especially for many simple tokens such as *sober* and *para* used as prepositions that were wrongly tagged as verbs.

In the methodology we propose, we also demonstrated the usefulness of machine translation in the comparison of translation solutions by leveraging the standardization of style and expressions that seem to be favored because of their consumption of the enormous amount of corpus data. In fact, we have shown that machine translation may be used to distinguish automatically most instances of fixed shifts, which are confirmed when machine translation also produces the same information shift, from free shifts, which are confirmed when the information shift in human translation is not present in an otherwise grammatically and semantically correct machine translation.

Finally, in the context of increased interest towards more formal and objective methods in human and machine translation assessment and evaluation, we hope that the methodology described in this paper could lay the foundation for language-independent translation assessment procedures and models. For example, the weighted Euclidean distance could be used in association with other automatic translation quality control methods that rely on reviewing translations of specific lexical items in a source segment against conventional translations found in bilingual dictionaries or other reference material or documentations.

Abbreviations

TPR	Translation precision ratio	PDF	Precision deviation factor

References

Bakker, Matthijs, Cees Koster & Kitty Van Leuven-Zwart. 2011. Shifts. In Mona Baker & Gabriela Saldanha (eds.), *Routledge encyclopedia of translation studies*, 2nd edn., 269–274. London: Routledge.

Becher, Viktor. 2010. Abandoning the notion of "translation-inherent" explicitation: Against a dogma of translation studies. *Across Languages and Cultures* 11(1). 1–28.

Becher, Viktor. 2011. *Explicitation and implicitation in translation: A corpus-based study of English-German and German-English translations of business texts*. Universität Hamburg. (Doctoral dissertation).

Blum-Kulka, Shoshana. 1986. Shifts of cohesion and coherence in translation. In Juliane House & Shoshana Blum-Kulka (eds.), *Intercultural communication: Discourse and cognition in translation and second language acquisition*, 17–35. Tübingen: Narr.

Explosion_AI. 2016–2020. *spaCy v2.1.3, Industrial Strength Natural Language Processing in Python*. https://spacy.io (5 August, 2020).

Gale, William A. & Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics* 19(1). 75–102.

Giesbrecht, Eugenie & Stefan Evert. 2009. Is part-of-speech tagging a solved task? An evaluation of POS taggers for the German web as corpus. In Iñaki Alegria, Igor Leturia & Serge Sharoff (eds.), *Proceedings of the Fifth Web as Corpus Workshop (WAC5)*, 27–35. San Sebastian: Elhuyar Fundazioa.

Klaudy, Kinga. 2011. Explicitation. In Mona Baker & Gabriela Saldanhaa (eds.), *Routledge encyclopedia of translation studies*, 2nd edn., 104–108. London: Routledge.

Murtisari, Elisabet Titik. 2016. Explicitation in translation studies: The journey of an elusive concept. *Translation & Interpreting* 8(2). 64–81.

Poirier, Éric André. 2017. A comparison of three metrics for detecting crosslinguistic variations in information volume and multiword expressions between parallel bitexts. In Ruslan Mitkov (ed.), *Proceedings of EUROPHRAS 2017*, 1–10. Geneva: Editions Tradulex.

Poirier, Éric André. 2019. Repérage des décalages informationnels de traduction au moyen du criblage automatique des segments hétéromorphes d'un corpus parallèle. *TTR: Traduction, terminologie, rédaction* 32(2). 279–308.

Wecksteen-Quinio, Corinne, Mickaël Mariaule Corinne & Cindy Lefebvre-Scodeller. 2015. *La traduction anglais-français: Manuel de traductologie pratique*. Louvain-la-Neuve: De Boeck.