

Localizing into an unknown dialect: impossible or challenging?

Marta Gómez Palou
University of Ottawa

1. Background

Is localizing a text into an unknown dialect an impossible task, or is it merely an extraordinary challenge? Before pondering this question, we may first want to ask: why would anyone attempt such thing? The current state of the market provides us with several answers. Globalization has become a fixture of our society (Fry and Lommel, 2003:8), and companies sell products to international markets on a daily basis. Multinational companies are often not located in the markets they serve; therefore, they must localize their products. The rise in multinational marketing has given birth to the industry of localization (Esselink, 1998:2). Companies adapt their products linguistically and culturally to their target markets, which have in turn been narrowed down to very specific regions. For example, products are not translated into Spanish, but into Mexican Spanish, or Argentinean Spanish. Translated products reflect their language *locales* (Hall and Hudson, 1997:4). It is important to note, however, that although the demand for translation is increasing, there are not enough translators in the market to meet it (Lange and Bennett, 2000:203).

An alternative solution to this problem could be to have translators work into a language dialect that is not their own, but is this possible? Not with the tools and training currently available. To the best of my knowledge, not much research has been done in this area. My work is an attempt to start meeting the growing need for localization. My research focuses on the evaluation of potential resources for translating into a non-native dialect, specifically electronic corpora. Hence, my question reads:

Would a specially designed monodialectal corpus be a useful resource for translators working into a dialect of which they are not native speakers?

Before beginning our quest to answer the above question, I believe it is necessary to define three terms:

corpus: a “large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria” (Bowker and Pearson, 2002:9).

dialect: a diatopic variety of a language, i.e. a geographic variety such as French Canadian or Argentinean Spanish.

non-native dialect: a dialect that one does not grow up speaking (my definition).

2. Methodology

With the meaning of these terms clarified, we can proceed to examining the research carried out to answer the hypothesis question. In order to test the usefulness of an electronic monodialectal corpus for translators, I designed a translation experiment and

asked two groups of peninsular Spanish) translators to translate two texts for a target audience of Argentinean Spanish speakers. The first text had to be translated using conventional resources, such as dictionaries, parallel texts, reference works and the Internet; the second, using a monodialectal corpus I specially compiled with articles from the weekly supplement on information technology, *Informática 2.0*, published in the Argentinean newspaper *Clarín*. I mixed and matched groups and tasks to limit the effect that differences between the two source texts might have had on the experiment. After completing the translation, translators were asked to answer a short questionnaire about their experience. The resulting translations were then sent to an impartial evaluator to be assessed according to a set of guidelines and a marking grid that I developed for the experiment.

Given that this project was carried out as part of an M.A. thesis, I was required to limit the scope of my research.

- a) Language/dialect: I had to limit my research to one source dialect (peninsular Spanish) and one target dialect (Argentinean Spanish). I chose to work with these dialects for practical reasons. After assessing my chances of finding volunteers to carry out and evaluate the translation, I decided this was my best dialect combination.
- b) Corpora: Since no specialized Argentinean corpus already existed, I had to resign myself to compiling my own specialized corpus, which, due to time restrictions, could not be very large. Therefore, I had to accept that I would be unable to make any statistically relevant conclusions using my corpus. Instead, I focused on developing a robust methodology that could be replicated whenever I had the time and resources to retry the experiment.
- c) Source texts: I used two texts that—ideally—should have contained exactly the same translation problems. However, it was very hard to find two texts with identical characteristics. In response to this complication, I designed the experiment taking into account the different nature of the two source texts.
- d) Participants: All participants were volunteers. This reduced the pool of candidates and required me to pare down the amount of work they would be expected to carry out.

3. Results Analysis

I divided the results obtained from the experiment into quantitative data (i.e. the evaluator's scores) and qualitative data (i.e. the evaluator's comments and examples).

3.1 Quantitative Results

In this section, I will first compare the evaluator's assessment of the texts translated with conventional resources with her assessment of the texts translated using a monodialectal corpus. I will then analyze the scores received by each translator (for their proper/appropriate use of? Clarify how the translators used the characteristics and what they were scored on specifically) proper use of specific linguistic characteristics, such as verb forms, personal pronouns, lexis and phraseology, and describe the results of an overall assessment of each translator's pair of texts.

3.1.1 Translations with Corpus vs. Translations Without

At the overall, “general” level, five out of the eight texts were categorized by the evaluator as being mostly adequate for an Argentinean target audience. Two of the texts were deemed inadequate, while only a single text was classified as adequate. Interestingly, the two texts found to be inadequate were produced using conventional resources, while the only text selected as being wholly acceptable was produced with the help of a corpus.

Moreover, a look at the scores assigned for specific characteristics in Table 1 shows that the non-corpus users earned 4 inadequate scores, 9 mostly adequate scores and 3 adequate scores. In contrast, the corpus users had only a single inadequate score, 9 mostly adequate scores, and 6 adequate scores.

	AGUA 1	AGUA 2	FUEGO 1	FUEGO 2
	WITHOUT CORPUS		WITHOUT CORPUS	
General Aspects	Mostly Adequate	Mostly Adequate	Inadequate	Inadequate
Verb Forms	Mostly Adequate	Adequate	Inadequate	Mostly Adequate
Pers. Pronouns	Mostly Adequate	Adequate	Inadequate	Adequate
Lexis	Mostly Adequate	Inadequate	Mostly Adequate	Mostly Adequate
Phraseology	Mostly Adequate	Mostly Adequate	Mostly Adequate	Inadequate
	WITH CORPUS		WITH CORPUS	
General Aspects	Mostly Adequate	Adequate	Mostly Adequate	Mostly Adequate
Verb Forms	Mostly Adequate	Mostly Adequate	Inadequate	Adequate
Pers. Pronouns	Mostly Adequate	Mostly Adequate	Adequate	Adequate
Lexis	Mostly Adequate	Mostly Adequate	Mostly Adequate	Adequate
Phraseology	Adequate	Adequate	Mostly Adequate	Mostly Adequate

Table 1

If we look at the scores given to specific characteristics, we can see whether different resources result in different performances:

- Verb Forms: It appears that both resources provide equally good support for translators. Non-corpus users received 1 score of inadequate, 2 of mostly adequate, and 1 of adequate. The very same scores were given to corpus users.
- Personal Pronouns: The corpus seems to offer a slight advantage. Non-corpus users received 2 adequate scores, 1 mostly adequate and 1 inadequate score, while corpus users received 2 adequate scores and 2 mostly adequate scores.
- Lexis: The corpus led to a better performance. 3 out of the 4 scores earned by non-corpus users were mostly adequate, while the final was inadequate. In contrast, corpus users earned 3 mostly adequate scores and 1 adequate score.

- d) Phraseology: It is at this level that the most significant difference appears. The non-corpus users received 3 scores of mostly adequate and 1 of inadequate, while the corpus users received 2 scores of mostly adequate and 2 of adequate.

To sum up, it appears that the corpus does in fact offer a slight advantage over the conventional resources. The most improvement can be seen in the phraseology category; but the corpus also yielded slight improvements in the lexis and personal pronouns categories. However no change whatsoever was noted in the verb forms category. These results contributed to a slightly better overall performance by the corpus users.

3.1.2 Translators' performances

We can also analyze the results in Table 1 according to the scores given to each translator. Firstly, Agua 1 did show a slight improvement at the level of specific categories. The same score was retained in 3 of the 4 categories; however, in the fourth category (i.e. phraseology), Agua 1's score improved and changed from mostly adequate to adequate.

Secondly, Fuego 1 showed definite improvement in the personal pronoun category when using the corpus; the score reversed from inadequate to adequate.

Thirdly, Fuego 2 showed the most improvement when using the corpus, going from 1 inadequate score, 2 mostly adequate scores and 1 adequate score to 1 mostly adequate and 3 adequate scores. While Fuego 2 kept the same score in the personal pronoun category, improvement was seen in the three other categories when the corpus was used.

Finally, in spite of showing an overall improvement, Agua 2 revealed mixed results at the level of individual categories. In two of the categories—verb forms and personal pronouns—this translator actually obtained poorer results when working with the corpus, dropping from a score of adequate in both categories when working with conventional resources to a score of mostly adequate in both categories when working with the corpus. However, for the categories of lexis and phraseology, this translator showed improvement when working with the corpus, the scores for these categories rose from inadequate and mostly adequate to mostly adequate and adequate, respectively.

3.1.3 Overall Translation Pair Assessment

To end my analysis of the quantitative results, I will comment on a final evaluation performed on the texts. The evaluator, who did not know which texts were translated using the corpus and which were translated without it, was specifically asked to look at the pair of texts produced by each translator and to judge which of the two was the most acceptable for an Argentinean target audience.

With the exception of Agua 1, for whom the evaluator found that there was no difference between the two texts in terms of their general acceptability, the evaluator determined that the translations produced with the help of the corpus were more acceptable than

those translated using only conventional resources. This seems to further indicate that a corpus is a useful resource for helping translators work into a non-native dialect.

Translator	Resources used	Text translated with conventional resources	Text translated with corpus
	Agua 1	Equally acceptable	Equally acceptable
	Agua 2	Less acceptable	More acceptable
	Fuego 1	Less acceptable	More acceptable
	Fuego 2	Less acceptable	More acceptable

Table 2

4. Qualitative Results

In this section, I will present the comments and examples provided by the evaluator, as well as the translators' responses to a questionnaire on the experiment.

4.1 Evaluator Comments and Examples

Here I will examine the comments and examples provided by the evaluator with regard to the translators' performance relating to each individual's use of linguistic characteristics (i.e. verb forms, personal pronouns, lexis and phraseology).

Firstly, given that verb forms and personal pronouns must agree in person and number, I will analyze the comments regarding these two characteristics in the same section. The four translations reveal two different strategies used by the translators. Two translators decided to use the formal second person *Usted*, while the other two opted for the informal second person *vos* or *vosotros*. In the context of this analysis, I will disregard the translations using *Usted* since, as the evaluator points out, the use of this person is correct but formal and not specifically Argentinean.

With respect to the two translators who opted to use the informal second person, their use of verb forms was as follows. When translating using only the conventional resources, Agua 1 opted for *vos* but used the future simple form (considered formal), which created a clash in language register. Meanwhile, it is hard to determine whether Fuego 1 used the second person of the plural *vosotros* or the archaic formal *vos* used in Spain. As *vosotros* is never present in Spanish from Argentina and neither is the formal *vos*, the evaluator considered this choice absolutely inadequate. Interestingly, when Fuego 1 moved on to translate the next text with the help of the corpus, this translator decided to abandon the choice of *vosotros*/formal *vos* and use instead the *vos* characteristic of Argentinean Spanish. Agua 1 used *vos* in both translations (though with an erroneous verb tense when not using the corpus) and was praised by the evaluator for using the imperative with the second person *vos* when translating with the help of the corpus. The evaluator found errors in the use of *voseo* verb forms for both translators.

With regard to lexis, the evaluator identified four pairs of terms that have a different frequency of use in Argentinean Spanish as compared to peninsular Spanish. These pairs

are represented in the corpus as illustrated in Table 3. Of these four pairs, *computadora/ordenador* and *elegir/escoger* appear in both source texts, while *video/vídeo* appears only in Text 1, and *oficina/despacho* appears only in Text 2.

Preferred term in <i>Argentinean Spanish</i> (no. of occurrences in <i>ClarínInformática</i>)	Preferred term in <i>peninsular Spanish</i> (no. of occurrences in <i>ClarínInformática</i>)
<i>computadora</i> (102)	<i>ordenador</i> (0)
<i>video</i> (90)	<i>video</i> (2)
<i>elegir</i> (25)	<i>escoger</i> (2)
<i>oficina</i> (23)	<i>despacho</i> (0)

Table 3

When not using the corpus, only two translators opted for *computadora*, while the other two used *ordenador*. However, when using the corpus, all four translators opted for the term *computadora* or the borrowed acronym *PC* with a feminine article (which also appeared in the corpus). As for the pair of verbs *elegir/escoger*, when not using the corpus, two translators chose *elegir*, while two selected *escoger*. However, when using the corpus, one of the translators who had previously used *escoger* switched to *elegir*, which is more appropriate for the target audience.

With regard to the pair *oficina/despacho*, which appeared only in Text 2, the opposite trend occurred. When using conventional resources, both translators correctly employed *oficina*, but when using the corpus, one translator selected *oficina* and the other chose *despacho*.

Finally, in the case of the spelling variant *video/vídeo*, which appeared in Text 1, the evidence is inconclusive. When the translation was carried out using conventional resources, one translator opted for *video* and the other for *vídeo*. When the text was translated using the corpus, one translator used *video* and the other used a construction that avoided the term altogether.

In summary, one could say that in the course of translating the two texts, there were a total of six times where each of the four translators were able to choose between a term that was more typical of Argentinean Spanish and one that was more typical of peninsular Spanish. This means that in the whole experiment, there were a total 24 opportunities—12 when using conventional resources and 12 when using the corpus—for me to see what lexical solution would be chosen. The chosen solutions are summarized in Table 4.

	Using conventional resources	Using the corpus
No. of times Argentinean solution chosen	7 (58%)	7 (58%)
No. of times peninsular Spanish solution chosen	5 (42%)	2 (17%)
No. of times a “neutral” strategy was used.	0 (0%)	3 (25%)
Total	12 (100%)	12 (100%)

Table 4

As the data reveals, regardless of whether or not the corpus was used, the correct Argentinean solution was selected 58% of the time. However, it is interesting to note that,

when conventional resources were used, the incorrect peninsular Spanish term was selected in the remaining 42% of cases. However, when the corpus was used, the number of times an incorrect peninsular Spanish term was selected dropped to 17%, while the remaining 25% of the time, a neutral solution was chosen (i.e. a solution that would be acceptable in either dialect). This data may suggest that the corpus offers translators a wider range of options for finding acceptable translation solutions than conventional resources.

4.2 Translator satisfaction

I will proceed to consider the insights shared by the translators their degree of satisfaction with the translations produced, the resources available, and the helpfulness of both the conventional resources and the corpus.

With regard to translators' satisfaction with their work, two of the participants were partially satisfied with both of their texts. The third translator was uncertain about how well the dialect had been reproduced in the translation completed using conventional resources, though this same translator was more satisfied with the translation produced using the corpus. Similarly, the fourth translator was unsatisfied with the translation carried out using conventional resources and was convinced that the translation done with the corpus did a better job of reproducing Argentinean Spanish.

In terms of the conventional resources used during the experiment, the translators expressed dissatisfaction with these kinds of reference materials, which they felt did not meet their specific needs in the context of this experiment. For example, they complained that the peninsular-Argentinean Spanish dictionaries were unidirectional, which limited the type of searching that could be carried out. Another complaint was that these dictionaries did not include information on verb conjugation. However, the translators did acknowledge that bilingual Spanish-English dictionaries were useful for non-dialect specific issues, such as finding equivalents for some specialized terms, including *snapshot*, *photo printer* or *double-sided printing*.

As for translators' impression of the corpus, they generally viewed it as a useful resource. They found the corpus helpful for finding appropriate collocations, terminology, gender, personal pronouns and verb conjugations. In addition, two of the translators highlight the fact that they made some serendipitous finds. For example, while searching for a solution to a particular problem, they also found a solution to a different problem (e.g. in the context surrounding the initial search). No such serendipitous finds were reported when translators used the conventional resources. That said, there was a general concern among the translators about the corpus being too time-consuming, since they were not very familiar with it.

5. Conclusion

The analysis of the data collected in this experiment seems to support the finding that a specially designed monodialectal corpus can indeed help translators translate texts into a

non-native dialect. The quantitative data generated during this experiment show that, in the majority of cases, translations produced with the help of the corpus were deemed more adequate for the needs of the target audience than those produced solely with the assistance of conventional resources. These findings were supported by the qualitative data collected from the evaluator and the translators. For instance, the evaluator notes that, in a number of cases, a translator who made poor choices when working with the conventional resources sometimes went on to correct these choices when working with the corpus. This study seems to indicate that corpora are a promising resource for translators working into a non-native dialect. The study results may not be definitive due to the scope of the project, but they definitively suggest that more research is needed in this direction.

6. Further Work

Obviously, to continue research in this direction it would be necessary to conduct the study on a larger scale, in terms of participants, languages, texts and evaluators. However, I am especially interested in observing the influence of another variable: translators' familiarity with the dialect. I am curious to measure the variation, if any, of the results in light of translators' degree of familiarity with the non-native dialect. It seems logical to assume that a certain knowledge of the dialect would enhance the translator's use of the corpus by offering general guidelines on what to look for. If such an assumption could be proven, it would be very useful to determine the minimum level of familiarity required for an efficient exploitation of the corpus. This could later be designated as a potential skill to be taught to translation trainees in order to prepare them for translating into non-native dialects.

References

- Bowker, L.; Pearson, J. (2002). *Working with Specialized Language: A Practical Guide to Using Corpora*. London; New York: Routledge.
- Esselink, B. (1998). *A practical guide to software localization: for translators, engineers and project managers*. Amsterdam, Philadelphia: J. Benjamins Pub. Co.
- Fry, D.; Lommel, A. (2003). *The Localization Industry Primer (2nd. Ed)* LISA, The Localization Industry Standards Association.
- Hall, P.V.A.; Hudson, R. (1997). *Software without frontiers: a multi-platform, multi-cultural, multi-nation approach*. New York: Wiley.
- Lange, C.A.; Bennett, W.S. (2000). *Combining Machine Translation with Translation Memory at Baan*. In Sprung, R.C. (Ed). *Translating Into Success*. (pp. 203–218). Amsterdam, Philadelphia: Benjamins.