

# Multilingual Terminological Resources: Comparing Machine Translation and Corpus-Based Translation

Melania Cabezas-García<sup>1</sup> and Pilar León-Araúz<sup>1</sup>

<sup>1</sup> University of Granada, C/ Buensuceso 11, Granada, 18002, Spain

## Abstract

Terminological resources increasingly use machine translation as a method to speed up time and reduce costs. With a view to enhancing the multilingual representation of multiword terms (e.g. *passive stall-regulated wind turbine*) in terminological resources, we describe an analysis of English-Spanish multiword term translation in various machine translation systems, paying special attention to the errors encountered. A comparison of machine translation output with the equivalents found in a comparable corpus is also presented. Even though machine translation often shows errors, it can serve as a basis for human post-editing, thus saving time and costs in terminological work. Comparable corpora, on the other hand, offer better results, but searches are more time-consuming.

## Keywords

Multiword term, machine translation, corpus, specialized translation, terminology

## 1. Introduction

With a view to expanding markets and disseminating knowledge, specialized texts generate a large volume of translations. Terminological resources should assist in this respect by means of the inclusion of multilingual information. In this sense, machine translation is increasingly being used as a method to speed up time and reduce costs [1].

This paper focuses on the translation of distinctive units of scientific texts, i.e. multiword terms (e.g. *passive stall-regulated wind turbine*), which pose problems both to human translators and natural language processing systems. However, multiword term machine translation has not been the focus of attention with some exceptions, such as [2]. This is especially true of more complex multiword terms that have three or more constituents.

In order to enhance the multilingual representation of multiword terms in terminological resources, we carried out the following tasks: (i) we analyzed English-Spanish multiword term translation in various machine translation systems; (ii) developed a proposal of the causes that may generate errors in multiword term machine translation; and (iii) compared machine translation output with the equivalents that may be manually found in corpora. For this purpose, a set of three-, four-, and five-term English multiword terms related to environmental science were extracted from a specialized corpus on this field (10,228,919 words, [3]). Environmental science was chosen due to the large volume of translations generated as a result of the increasing environmental awareness.

## 2. Machine Translation versus Corpus-Based Translation

---

*1st International Conference on "Multilingual digital terminology today. Design, representation formats and management systems", June 16 – 17, Padova, Italy*

EMAIL: melaniacabezas@ugr.es (M. Cabezas-García); pleon@ugr.es (P. León-Araúz)

ORCID: 0000-0002-8622-1036 es (M. Cabezas-García); 0000-0002-8520-2749 (P. León-Araúz)



© 2022 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

Far from the classic challenging view of machine translation, according to which it would replace human translators, machine translation also presents opportunities not only to human translators, as evidenced in the great demand for machine translation post-editing (i.e. reviewing and enhancing a machine translation), but also to terminologists. Evidently, including post-editing in the workflow brings added value to machine translation, minimizing possible mistakes and providing quality equivalents to be included in terminological resources.

Even though training a neural machine translation system by means of carefully selected corpora from the specialized subject field could provide better results than using generic machine translation engines, the truth is that translators usually do not have user friendly tools to train their own domain-specific engines. For this reason, the selected English multiword terms were provided without context to different generic machine translation engines: Google Translate and DeepL (neural systems), and Apertium (rule-based system).

To compare machine translations with equivalents found in corpora, parallel or comparable corpora can be used. Parallel corpora are sets of original texts aligned with their translations, thus facilitating the identification of equivalents. However, such corpora are scarce, especially in languages other than English, and generally show a marked influence of the source text on the translation. In contrast, comparable corpora are more useful. Since they are two sets of original texts of the same type and subject, they can be used to analyze native expressions in each language [4].

Therefore, a Spanish comparable corpus was used, which includes environmental texts originally written in this language (10,667,434 words). Techniques for identifying multiword term equivalents in corpora [5] were employed since translation identification in comparable corpora is not as direct as in machine translation.

### 3. Translating Multiword Terms using Machine Translation and Corpora

Multiword terms pose problems both to human translators and natural language processing systems since their adequate translation must consider aspects such as their internal dependencies, the semantic relation between constituents, the specialization of elements, etc. [5]. Many of these issues involve human intelligence, which machine translation lacks. General multiword expressions (e.g. *take a seat, by and large, let's go, as soon as*) have been widely explored in machine translation [6-10]. However, specialized multiword terms have received considerably less attention.

Not surprisingly, our results revealed that machine translation systems' output varies in the different engines. They often show errors of different nature and magnitude, which were used to establish the different causes behind them and could be used to enhance machine translation systems. These errors include: (i) the wrong identification of internal dependencies (*[doubly fed] [induction generator] > inducción alimentada doblemente generador*, lit. *\*generator doubly fed induction*); (ii) the wrong translation of constituents (*wave turbulence interaction parameterization > interacción de turbulencia ondulatoria parameterization*); and (iii) the wrong identification of the internal semantic relation (*wind-generated electricity > viento-electricidad generada*, lit. *\*generated wind-electricity*). However, machine translation can serve as a basis for human post-editing, thus saving time and costs in terminological work.

Comparable corpora, on the other hand, offer better results, but searches are more time-consuming. Ideally, these different techniques should be integrated into translators' and terminologists workflow, something that language service providers in the 2020s are bound to do. Furthermore, these results can be integrated into training for future translators and terminologists, who will have to work in this ever-changing reality.

### 4. Acknowledgements

This research was carried out as part of projects PID2020-118369GB-I00, Transversal integration of culture into an environmental terminological knowledge base (TRANSCULTURE), funded by the Spanish Ministry of Science and Innovation; and project A-HUM-600-UGR20, Culture as a transversal module in an environmental terminological knowledge base (CULTURAMA), funded by the ERDF Operational Programme for Andalucía 2014-2020.

## 5. References

- [1] M. Arcan, E. Montiel-Ponsoda, J. P. McCrae, P. Buitelaar. Automatic Enrichment of Terminological Resources: the IATE RDF Example, Proceedings of LREC 2018, 930-937, 2018.
- [2] Valavani, Christina, Christina Alexandris, and George K. Mikros. "Improving machine translation output of German compound and multiword financial terms: a comparison with cross-linguistic data." *Human-Intelligent Systems Integration 2* (2020): 29-34.
- [3] P. León-Araúz, A. San Martín, A. Reimerink. The EcoLexicon English Corpus as an open corpus in Sketch Engine, Proceedings of the 18th EURALEX International Congress, edited by Čibej, J., Gorjanc, V., Kosem, I., Krek, S., 893-901, Ljubljana, Euralex, 2018.
- [4] L. Bowker. Terminology and translation, in: H. Kockaert and F. Steurs (Eds.), *Handbook of Terminology*, John Benjamins, Amsterdam, Philadelphia, 2015, pp. 304-323.
- [5] M. Cabezas-García, P. León-Araúz. Procedimiento para la traducción de términos poliléxicos con la ayuda de corpus, in: G. Corpas Pastor, M. R. Bautista Zambrana, C. M. Hidalgo Ternero (Eds.), *Sistemas fraseológicos en contraste: Enfoques computacionales y de corpus*, Comares, Granada, 2021, pp. 203-230.
- [6] Hurskainen, Arvi. "Multiword Expressions and Machine Translation." *Technical Reports in Language Technology*, Report No 1 (2008): 1-18.
- [7] A. Barreiro, J. Monti, B. Orliac, F. Batista. When multiwords go bad in machine translation, MT Summit Workshop Proceedings on Multi-word Units in Machine Translation and Translation Technology, 26-33, 2013.
- [8] Constant, Mathieu, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner and Amalia Todirascu. "Multiword Expression Processing: A Survey." *Computational Linguistics* 43, 4 (2017): 837-892.
- [9] Ebrahim, Sara, Doaa Hegazy, Mostafa Gadal-Haqq M. Mostafa and Samhaa R. El-Beltagy. "Detecting and Integrating Multiword Expression into English-Arabic Statistical Machine Translation." *Procedia Computer Science* 117 (2017): 111-118.
- [10] A. Zaninello, A. Birch. Multiword Expression aware Neural Machine Translation, Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), 3816–3825, 2020.