# Automatic Summarization as a Documentation Resource for the Human Translation of Research Articles from the Legal-Technological Field (Spanish-English-French)

María Cristina Toledo Báez
University of Málaga, Spain
toledo(a)uma.es

ABSTRACT
This thesis is concerned with the impact of translation technologies, specifically automatic summarization, on the human translation of specialized texts such as research articles regarding legal-technological field. The main hypothesis aims to prove that term-based summarization constitutes a useful documentation and terminology resource for semi-professional translators. To show this, two main materials are used: on one hand, the automatic summarization system Term-Based Summariser (TBS), developed by the research group in computational linguistics from the University of Wolverhampton (United Kingdom); on the other hand, a trilingual (Spanish, English and French) comparable corpus consisting of a collection of research articles with more than three millions of tokens. Thanks to the use of TBS, this thesis is a pioneer work in translation technologies because automatic summarization has never been part of an empirical study with translators. Besides, automatic summarization is an interesting and innovative alternative that represents a boost in translation studies because it is the merging of computational linguistics, natural language processing and natural language generation.

Regarding methodology, three empirical studies are carried out. First, a comparative analysis of the legal-technological discourse, i.e., the discourse from the information technology law, whose main feature is the combination of law and computing. For this purpose, three representative articles, one in Spanish, one in English and one in French, have been used to compare lexical, terminological, phraseological and morphosyntactical features. The second empirical study focuses on the analysis of the text genre of the corpus, that is, the research article, comparing 60 articles, 20 in each language, in order to test whether the Introduction-Material and Methods-Results-Discussion/Conclusion (IMRD) structure of English scientific articles may be valid to articles, on one hand, on legal sciences and, on the other hand, in the Romance languages of Spanish and French. The third study is a quantitative and qualitative analysis developed to demonstrate whether Term-Based Summariser (TBS) enhances direct and inverse translation of research articles. In order to achieve this, an empirical study with different experiments was carried out by a hundred semi-professional translators, i.e., undergraduate students from the $4^{th}$ year in translation and interpreting. Two types of criteria have been established to demonstrate with quantitative results whether automatic summarization is effective for translators: quality criteria, and number of words translated criteria. In quality criteria both analytic and holistic evaluation are critical since they provide qualitative results with the evaluation of the subcorpus of translated texts. We have settled two evaluation scales for the two methods: analytic evaluation, for which the notion of error is basic, and holistic evaluation, of which translation competence is the key concept. In number of words translated criteria, the key is to test whether the extension of texts translated with TBS is longer or not than the extension of texts translated without TBS. Quantitative results are also presented in this thesis thanks to the electronic survey filled out by the semi-professional translators.

As far as results are concerned, the first empirical study of the legal-technological discourse demonstrates that this type of discourse shares in Spanish, English and French a set of common features: on one hand, lexical, terminological and phraseological features such as latinisms, hellenisms, anglicisms, gallicisms, phraseological collocations and suffixation; on the other hand, morphosyntactical features such as passive voice and the particular use of certain verbs. The second empirical study has proved that the most common structure for research articles is the IMRD structure.We have analyzed the IMRD structure from two perspectives: the domain and the language. As to the domain, we have shown that this structure appears in research articles from social and law sciences, albeit with some modifications. In terms of language, IMRD structure is typical and common in English language but we have shown that it is also used in Romance languages, specifically in Spanish and French, but with some variations. Regarding the third empirical study, quantitative results concerning both quality criteria and number of words translated criteria show two main conclusions. First, all the translations were evaluated with our own evaluation scales and the ones translated with TBS had better results (i.e. fewer errors and more correct answers) than the texts translated without TBS. Second, the data concerning the number of words criteria are quite clear: the extension of texts translated with TBS is longer than the extension of texts translated without TBS. Finally, the qualitative results from the electronic survey demonstrate that most translators considered TBS to be an interesting and useful resource as well as an innovative terminological and documental tool.

In short, the main hypothesis is tested because we have shown by means of qualitative and quantitative results that automatic summarization enhances specialized translation in three languages (Spanish, English and French) and in both direct and inverse combinations, albeit with better results for direct translation.

KEYWORDS: automatic summarization, english, french, legal-technological discourse, spanish, specialized translation, research article.

Completion of Thesis
Place: University of Málaga, Spain
Year: 2009
Supervisor: Dr Gloria Corpas Pastor
Original language: Spanish

# El resumen automático como recurso documental para la traducción de artículos de investigación del ámbito jurídico-tecnológico (español-inglés-francés)

María Cristina Toledo Báez
Universidad of Málaga, España
toledo(a)uma.es

RESUMEN
La presente tesis doctoral propone un nuevo enfoque en el ámbito de las tecnologías de la traducción al emplear el resumen automático como recurso documental para la traducción directa e inversa en tres lenguas, a saber, español, inglés y francés, de artículos de investigación del ámbito jurídico-tecnológico. Para ello, se emplean, por un lado, el programa de resumen automático Term-Based Summariser, desarrollado por el Research Group in Computational Linguistics de la Universidad de Wolverhampton (Reino Unido), y, por otro, un corpus comparable trilingüe de artículos de investigación de más de tres millones de *tokens*.

En lo que concierne a la metodología, se han desarrollado tres estudios empíricos de enjundia. En primer lugar, se ha llevado a cabo un análisis comparativo del discurso jurídico-tecnológico de los artículos de investigación, cotejando los rasgos terminológicos, fraseológicos y morfosintácticos en las tres lenguas de estudio. En segundo lugar, se ha realizado un análisis del género textual del corpus, esto es, el artículo de investigación, mediante la comparación de 60 artículos, 20 en cada lengua de trabajo, en aras de comprobar si la estructura *Introduction-Method-Results-Discussion* (IMRD), la más común en ciencias puras y en ámbitos anglosajones, es válida y extrapolable a, por un lado, las Ciencias Jurídicas y, por otro, a las lenguas romances español y francés. Por último, se ha llevado a cabo un análisis cuantitativo y cualitativo en torno a la viabilidad del resumen automático como recurso documental y terminológico para la traducción directa e inversa de artículos de investigación del ámbito jurídico-tecnológico mediante experimentos con un centenar de traductores semiprofesionales, esto es, de último año de carrera. El análisis cuantitativo se centró en dos criterios concretos: de una parte, el criterio de calidad, cuya implantación práctica se desarrolló en plantillas de evaluación de traducciones tanto analíticas, basadas en el error, como holísticas, con la competenca traductora como base, empleadas ambas como herramientas para evaluar las traducciones de los experimentos realizadas con o sin resumen automático; de otra parte, el criterio de número de palabras, mediante el cual se contabilizó el número total de palabras traducidas con o sin el resumen automático. Por su parte, el análisis cualitativo se representó mediante una encuesta electrónica en la que se reflejan las opiniones e impresiones de los traductores semiprofesionales.

Los tres estudios empíricos arrojan resultados significativos. El primero de ellos, el análisis contrastivo del discurso jurídico-tecnológico en español, inglés y francés muestra que aparecen rasgos comunes en las tres lenguas citadas, si bien en las dos lenguas romances se aprecian mayores coincidencias. En lo que respecta al segundo, el análisis contrastivo del género de los artículos de investigación confirma que la estructura anglosajona IMRD sí es extrapolable, por un lado, a las Ciencias Sociales y Jurídicas, aunque con ciertas modificaciones, ya que desaparece con frecuencia la sección de métodos, y, por otro, al

español y al francés, aunque consideramos que la estructura de un artículo está más determinada por la disciplina en sí que por la lengua. El resultado del tercer análisis, centrado en la viabilidad del resumen automático para la traducción, es el más trascendente para la presente tesis doctoral; de hecho, es el que confirma la hipótesis inicial: el resumen automático basado en términos, concebido como novedosa aplicación de la Lingüística Computacional a los Estudios de Traducción, constituye un recurso documental innovador y fiable que podría formar parte de futura una estación de trabajo del traductor.

PALABRAS CLAVE: artículos de investigación, discurso jurídico-tecnológico, español, francés, inglés, resumen automático, traducción especializada.