# Learning Español Técnico Simplificado as a new Controlled Language for Machine Translation

**Ilaria Gobbi**
Università "Alma Mater Studiorum" di Bologna – DIT di Forlì (Italy)
*ilaria.gobbi4@unibo.it*

## Abstract

*This study aims to present a new controlled language able to provide an automatic translation of technical documents from Spanish into English and vice versa. Particularly, in order to define a set of linguistic rules helping in technical writing and consequently automatic translating, "Español Técnico Simplificado" (ETS) was developed as a metalinguistic guidebook. The attention was focused on the particular structure of the English technical specification for the avionic documentation, i.e. Simplified Technical English (officially known as ASD-STE100). It was identified how that specification can provide unambiguous technical texts, useful for a reader who is not necessarily a native language speaker.*

*Considering that an ad-hoc corpus was designed to extract linguistic information, the experimental design and method used to develop the ETS guidebook was essentially corpus-based. The results produced by the information extraction and the English specification imitation led to the development of a linguistic (controlled) method tailored to produce a technical document free from any ambiguity. The ETS system relays on the concept that readability and comprehensibility are conditions to be fulfilled by the controlled text. And, through its structure, ETS was created to provide all the necessary tools in order to draft unambiguous texts. Thus, similarly to the English specification, the ETS guidebook is divided into two parts: "Parte 1-Reglas de escritura" concerning syntactical and stylistic rules, and "Parte 2-Diccionario" concerning a dictionary including a limited number of selected lemmas. All rules and lemmas were designed to obey the bijection principle of the Spanish linguistic signs.*

*ETS has profound implications for future studies on technical writing and translation, but above all the ETS method can provide language education by Information Technology system. It is possible to make an automatic translation between two controlled languages, Español Técnico Simplificado <> Simplified Technical English, since the two controlled languages are equivalent to each other. And it is possible to learn and teach the ETS method by an e-learning system that could be specifically designed.*

## 1. Introduction

Most of the existing controlled natural languages (CNL) are created by industry and are addressed to industry. Most of them include several English-based CNLs which are quantitatively higher than all the others non-English-based CNLs. This is due to the fact that English is the most used language in the world as a lingua franca, both for business and technical communication. Nevertheless, excluding only Chinese, which is contained within a wide geographical area, Spanish proves to be the most spoken language worldwide at present (see UNSTAT 2006; Instituto Cervantes 2010). But apart from a few exceptions - e.g. *Francais Rationalisé* (GIFAS 1999), none of the existing CNLs have been designed for being machine-translated one another.

This paper presents a new controlled language able to provide an automatic translation of technical documents from Spanish into English and vice versa. "Español Técnico Simplificado" (ETS) is a CNL resulting from a doctoral project which aimed to define a set of linguistic rules helping in technical writing and consequently in automatic translation procedures. ETS has been developed as a metalinguistic guidebook after focusing the attention on the particular structure of the English technical specification for the avionic documentation, i.e. *Simplified Technical English* (STE), officially known as ASD-STE100 ([ASD 2013). It was firstly identified how the latter specification can help with providing unambiguous technical texts, useful for a reader who is not necessarily a native language speaker. Then, it was considered the possibility to create an equivalent controlled language which could provide unambiguous technical texts in another language. The main idea was to create a Spanish-based CNL able to be matched with an English-based CNL in the machine translation (MT) system, in order to achieve the palindromic relationship "ETS<>STE", in other words *Español Técnico Simplificado* <> *Simplified Technical English*. Since the two controlled languages are equivalent to each other, it is possible to make an automatic translation from the source-controlled language to the target-controlled language. Either as a source or target language, ETS is a Spanish-based CNL that can be trained through an e-learning system specifically designed.

## 2. From Source-CNL to Target-CNL

In order to achieve the palindromic relationship "ETS<>STE", the two CNLs necessarily have to be equivalent. Both ETS and STE should be interchanged in MT as source-CNL and target-CNL. This is the reason why ETS has been designed on the basis of STE. However, to be more precise, ETS is equivalent but not equal to STE. They are equivalent because both CNLs include almost the same linguistic rules. They are not equal because both CNLs were created on the base of their own linguistic systems respectively. Both of them were developed by the corpus-based method. ETS, therefore, is not a direct translation of STE, as the two languages are derived from two different language families.

Writing a controlled text for MT is not only the process through which the writer drafts a text, but also the translating product of this process. Actually, the writer who is supposed to machine-translate the controlled text is a translator himself/herself. For converting an "uncontrolled-text" into a "controlled-text" within the same linguistic system, before pouring it into MT, the writer becomes in effect a translator involved in the pre-editing process. During this process, it is necessary that one of the two "uncontrolled" languages, Spanish or English, is converted into the related controlled language, ETS or STE. In order to make the text converted, the writer needs specific standardized guidelines or, in other words, a specific metalinguistic guidebook.

Therefore, controlled texts that are to be machine-translated are texts that a technical writer has already drafted according to specific guidelines. ETS and STE are linguistic guidelines for Spanish and English, respectively. Particularly, ETS and STE are documents including linguistic rules for a technical writer willing to draft a controlled text that both (human) readers and machine translation do understand correctly and unambiguously. Of course, MT can translate the combination of uncontrolled language and controlled language. Yet, the translation between two CNLs may offer significant time savings in the post-editing process.

### 2.1 The ASD-STE100 specification

The ASD-STE100 technical specification, or *Simplified Technical English*, is a Copyright and a Trademark of ASD AeroSpace and Defence Industries Association of Europe (ASD), formerly AECMA. It is a CNL specifically designed to help users better understand English-language maintenance documentation.

The current issue of STE (Issue 6) is dated January 2013. Based on the *Caterpillar Fundamental English* linguist rules and the *McDonnell-Douglas* dictionary, STE used a corpus-based approach for the development of its metalinguistic content and still resorts to this approach for the progressive and constant revision process. The continuous feedback received by users also provides the specification revision that is managed by the Simplified Technical English Maintenance Group.

The main purpose of STE is language simplification. In fact, STE is meant to:

- avoid ambiguity
- improve clarity and readability of technical writing
- improve comprehension of English technical documentation for users whose first language is not English
- facilitate computer-aided translation and machine translation.

The STE system is established on the principle of "one word-one meaning" (whenever possible). One single meaning is assigned to one single word so as to eliminate any possible misunderstanding on part of the controlled-text user. So, the one-to-one relationship makes sure that the information given in the controlled text turns to be:

- accurate
- complete
- relevant
- concise
- convincing
- meaningful
- unambiguous.

STE is basically a metalinguistic document helping technical writers to draw up texts which prove to be simpler and easier to read. The structure of STE, shown in Fig. 1, is stable and consolidated, and it includes two main parts: *Part 1 – Writing rules* and *Part 2- Dictionary*. Part 1 is specifically dedicated to syntactical and stylistic directives. Part 2 is specifically dedicated to lexical directives.

| | | |
|---|---|---|
| **ASD-STE100<br>Issue 6** | Part 1<br>*Writing Rules* | **65 linguistic rules**<br>Section 1 - Words (17 rules)<br>Section 2 - Noun phrases (3 rules)<br>Section 3 - Verbs (8 rules)<br>Section 4 - Sentences (4 rules)<br>Section 5 - Procedures (5 rules)<br>Section 6 - Descriptive writing (8 rules)<br>Section 7 - Warnings, cautions, and notes (6 rules)<br>Section 8 - Punctuation and word counts (11 rules)<br>Section 9 - Writing practices (3 rules) |
| | Part 2<br>*Dictionary* | **≈ 860 lemmas**<br>Column 1 - *Keyword (part of speech)*<br>Column 2 - *Approved meaning/ALTERNATIVE*<br>Column 3 - *APPROVED EXAMPLE*<br>Column 4 - *Not approved* |

*Fig. 1 Structure of the ASD-STE100 – Issue 6*

## 2.2 The Español Técnico Simplificado guidebook

*Español Técnico Simplificado* is a Copyright of Università di Bologna (see Gobbi 2014). It is a CNL resulting from a doctoral project specifically designed to help users better understand Spanish-language technical documentation.

The current version of ETS is dated June 2014. Based on the *Simplified Technical English* linguistic rules and dictionary, ETS is a result of the English-specification imitating but not translating approach. An ad-hoc corpus of representative written Spanish, made of avionic maintenance manuals, was thus designed with the aim of extracting linguistic information (i.e. rules and lemmas). The corpus is now available on line at the web page: http://docs.sslmit.unibo.it/doku.php?id=sarcophagus:carcass:tutorials:basic_1 (last accessed: 05-09-2014). The experimental design and method used to develop the ETS guidebook was essentially corpus-based. The results produced by both the information extraction and the English specification imitation led to the development of a linguistic (controlled) method tailored to produce a technical document free from any ambiguity.

The ETS system relays on the concept that readability and comprehensibility are conditions to be fulfilled by the controlled text which is produced accordingly. The main purpose of ETS is not only language simplification but also STE-translatability. Translatability from/into STE is a key-role of ETS. This is due to the fact that business and technical speakers already feel quite comfortable with English. The development of a Spanish-based CNL, in turn, appears as an advantage in terms of translation time reduction. Therefore, ETS is meant to:

- avoid ambiguity
- improve clarity and readability of technical writing
- improve comprehension of Spanish technical documentation for users whose first language is not Spanish
- facilitate computer-aided translation, and machine translation especially form/to STE.

Similarly to the English specification, the ETS guidebook (as shown in Fig. 2) is divided into two parts: *Parte 1-Reglas de escritura* concerning syntactical and stylistic rules, and *Parte 2-Diccionario* concerning a dictionary including a limited number of selected lemmas. All rules and lemmas were designed to obey the bijection principle of the Spanish linguistic signs. The main principle *un significante-un significado* (signifier-signified) is devoted to the unambiguous comprehension of the controlled-text user.

| | | |
|---|---|---|
| **Guía de ETS** | Parte 1<br>*Reglas de escritura* | **65 reglas de escritura**<br>Sección 1 - Palabras (17 reglas)<br>Sección 2 - Sintagmas (2 reglas)<br>Sección 3 - Verbos (6 reglas)<br>Sección 4 - Oraciones (5 reglas)<br>Sección 5 - Procedimientos (5 reglas)<br>Sección 6 - Escritura descriptiva (8 reglas)<br>Sección 7 - Avisos (6 reglas)<br>Sección 8 - Signos de puntuación y número de palabras (10 reglas)<br>Sección 9 - Practicas de escritura (A definir posteriormente) |
| | Parte 2<br>*Diccionario* | **≈ 1008 lemas**<br>Columna 1 - *Entrada (parte de la oración)*<br>Columna 2 - *Significado admitido/ALTERNATIVAS*<br>Columna 3 - *EJEMPLO ADMITIDO*<br>Columna 4 - *No admitido* |

*Fig. 2 Structure of the ETS guidebook*

## 3. ETS for MT through the IT system

The ETS metalinguistic guidelines provide a model, a standard of writing which is meant for several technical domains such as information technology, software, transport, automotive and medical equipment manufacturing, power generation, high-tech sectors, translation and language services, etc. Writers and translators who are interested in writing and/or translating correctly, whereas unambiguously Spanish texts, could use the ETS guidebook as a benchmark.

The guidebook can be distributed for a self-training, but ETS is also a method that can be trained. In fact, if conveniently designed, ETS teaching and learning can be offered by information technology. Teaching would be done by a specific distance-learning course. Learning would even be done by a double loop: e-learning course and the practical use of the ETS method in MT. ETS education could be conceived as a distance-learning course delivered worldwide through a specific educational portal, developed and implemented by both academic and professional communities. ETS pre-editing, MT, and post-editing procedures could all be part of a specific program including virtual lessons, exercises, classrooms and exams, given and supervised by online professors, and tutors. And all the e-learning process would be equivalent to specific or required hours of classroom activities. But the most important element would be that users could be reached everywhere in the world.

## 4. Conclusions

The STE equivalent structure makes the ETS method very accurate and precise in providing a machine (or even human) translation between those two CNLs in a very restricted time. Of course, post-editing procedures are required in any case, but they would be drastically reduced in number. Consequently, for this perspective, ETS has profound implications for future studies on technical writing and translation, and for potential future attractiveness on the industrial part. This is due mainly to the specificity of the method adopted by ETS, consisting in providing language education by Information Technology system. In fact, it is possible to make an automatic translation between two CNLs, since the two controlled languages are equivalent to each other. And it is possible to learn and teach the ETS method by an e-learning system that could be specifically designed. Thus, geographical barriers for education will be eliminated and the international business environment interest in the method will be grown accordingly (and considerably).

## References

[1] ASD 2013**.** ASD-STE100 Technical Specification – Issue VI. Available on line at: http://www.asd-ste100.org/request.html (last accessed: 03-09-2014)

[2] AAVV 2010. El español, una lengua viva. Madrid: Instituto Cervantes. 37-43. Available on line at: http://www.cervantes.es/imagenes/File/prensa/El%20espaol%20una%20lengua%20viva.pdf (last accessed: 03-09-2014)

[3] GIFAS 1999. Guide du Français Rationalisé, Edition N. 2. Paris.

[4] Gobbi Ilaria 2014. Español Técnico Simplificado. PhD dissertation defended on 12-09-2014. University of Bologna: Forlì.

[5] UNIBO. Carcass, Corpus Spagnolo aeronautico. Available on-line at: http://docs.sslmit.unibo.it/doku.php?id=sarcophagus:carcass:tutorials:basic_1 (last accessed: 03-09-2014)

[6] United Nations Statistics Division. Demographic Yearbook Special Census Topics. Volume 2 - Social characteristics. Table 5. Availabe on line at: http://unstats.un.org/unsd/demographic/sconcerns/popchar/popchar2.htm (last accessed: 03-09-2014)