A Specification and Validating Parser for Simplified Technical Spanish

Remedios Ruiz Cascales
Dept. of Computer Science and Information Systems
University of Limerick
remeruiz@hotmail.com

Richard F. E. Sutcliffe
Dept. of Computer Science and Information Systems
University of Limerick
Richard.Sutcliffe@ul.ie

Abstract

AECMA Simplified English for aircraft maintenance manuals is one of the best-known examples of a controlled language. In this paper we describe the development of its equivalent for Spanish. We also present a prototype validating parser and outline its evaluation.

1 Introduction

A Controlled Language is a subset of a natural language which has a restricted lexicon and controlled grammatical structures. The objective of a Controlled Language is to improve readability, standardisation, accessibility and translatability of documentation. The best-known example is European Association of Aerospace Industries (AECMA) Simplified English (SE) (AECMA, 1998). Following the success of SE, specifications for languages other than English have been developed. Examples include FREM (Français Rationalisé Entendu Modulaire) for French aircraft maintenance manuals (Barthe, 1998), ScaniaSwedish for truck maintenance documentation (Almqvist and Hein, 1996), and Controlled Siemens Documentary German for software documentation (Schachtl, 1998). However, no research has been undertaken on the development of a Controlled Spanish.

The objective of the work reported here was to design a Controlled Spanish for aircraft maintenance manuals which is similar to AECMA Simplified English and FREM. The stages involved in the work included the development of a number of writing rules in conjunction with a dictionary, and the design and evaluation of a Controlled Spanish checker to verify conformance with some of the rules. The result comprises the Simplified Technical Spanish (STS) Specification, which includes 36 rules, the STS General Vocabulary, which consists of a list of 875 words, and the STS Parser which has been designed to identify five different types of error which relate to six rules in the STS Specification.

In this article we start by outlining some of the previous work in the areas of controlled languages and validating parsing. We then describe the method used to develop STS and summarise the resulting specification. Next, we explain the capabilities of the STS validating parser and report the results of an initial evaluation of it. Finally we draw conclusions and make suggestions for further work.

2 Previous Work

2.1 AECMA Simplified English

AECMA SE comprises a restricted vocabulary of 1,565 words with an additional set of 57 rules for using that vocabulary (AECMA, 1998). SE originated when in 1979 the Association of European Airlines (AEA) asked the European Association of Aerospace Industries (AECMA) to investigate the readability criteria of its aircraft maintenance documentation. Through its Documentation Working Group (DWG), AECMA set up a project group called the Simplified English Working Group (SEWG), to research the problem and provide a solution. SE was the result of this initiative.

The author of SE can use only three sources of words. Firstly there are Approved Words from the SE Guide. These constitute the base vocabulary which contains 1,565 words. There are 196 verbs in the base vocabulary and these are approved in four forms: the infinitive, the third person singular, the past simple and the past participle. Manufacturers then add Technical Names and Manufacturing Processes to the base vocabulary. For the Boeing Simplified English Checker these comprise 7,000 extra terms (Wojcik, 1998).

There are 57 Writing Rules in the SE Guide (AECMA, 1998). The following are some of the better known ones:

- A sentence length limit of 20 words (25 for descriptive text),
- A paragraph limit of 6 sentences,
- A compound noun length limit of 3 words,
- A prohibition on progressive be and perfective have,
- A prohibition on the passive in procedures (discouraged in descriptions),
- A prohibition on the -ing form of the verb,
- A requirement that sequential steps be in separate sentences,
- A requirement that words only be used in their approved sense,
- A recommendation that articles be used where possible.

2.2 Other Simplified Languages

The influence of AEMCA SE has led to many initiatives to develop SEs for applications other than aircraft maintenance documentation. Some examples include Agilent Technologies English (Smartny, 2002), Attempto Controlled English (ACE) (Fuchs and Schwitter, 1996), Boeing Technical English (Wojcik, Holmback, and Hoard, 1998), Caterpillar Technical English (Kamprath, Adolphson, Mitamura, and Nyberg, 1998), Diebold Controlled English (Moore, 2000), Ericsson English (Ericsson, 2000), General Motors CASL (Means and Godden, 1996), Global English (Means, Chapman and Liu, 2000), Kodak English (Kodak, 2000), Nortel Standard English (Smartny, 2002), Océ Technologies English (Smartny, 2002), Perkins PACE (Douglas and Hurst, 1996) and Xerox MCE (Xerox, 2001).

While most research has been concerned with English there has been some work in other languages, including the following: Controlled Chinese (Zhang and Shiwen, 1998), Controlled Siemens Documentary German (Schachtl, 1998), GIFAS FR for French (Barthe, 1998), ScaniaSwedish (Almqvist and Hein, 1996). However, as far as the authors are aware, no specification for a CL in Spanish has yet been developed. This provided the motivation for the current project.

2.3 Validating Parsing

Controlled languages usually restrict the syntactic constructions which a technical author may use in a conforming document. A validating parser or controlled language checker is a program which takes as input a text and specifies for each constituent sentence whether or not it conforms syntactically to the rules of the language. Ideally it should explain violations in easily understandable terms and perhaps also suggest changes which would lead to conformance.

The best-known validating parser for SE is the Boeing Simplified English Checker (BSEC) (Wojcik, Harrison and Bremer, 1993) which has been in production since 1990. A set of over 350 rules is used to achieve a broad coverage of English technical writing. Some of the more important requirements of AECMA SE that the BSEC can detect are: sentence length (20 or 25 words), paragraph length (6 sentences), noun cluster length (3 words or less), missing articles (based on count and mass distinctions), unapproved verbal auxiliaries (passive, progressive, perfect, modals), unapproved -ing participles, multiple commands in a single sentence, and warning, caution and note errors. No other Simplified English checker is as complete or accurate in support of SE requirements as BSEC. The Boeing checker also catches some grammatical and stylistic errors that are not explicitly addressed in the SE standard. Among other things, it detects subject-verb agreement errors, double word errors, misspelled words and punctuation problems.

Boeing has also developed an experimental Meaning-Based Checker (BMBC) to generate more accurate analyses. The BMBC builds on the syntax-based BSEC by adding the capability 1) to determine when an approved word is used in an unapproved meaning and 2) to select only those alternatives for an ambiguous unapproved word that are appropriate for the meaning in which it is used (Holmback, Duncan and Harrison, 2000).

CASLChecker (Godden, 2000) was developed on 1993 using the METAL parser and grammar as a starting point (Lamiroy and Gebruers, 1989). Lexically, it recognises words which do not belong to the CASL vocabulary, it recommends approved CASL synonyms for some non-CASL-approved words (e.g., use ‘generator’ instead of ‘alternator’), and it provides examples to show right and wrong usage. Syntactically, the CASLChecker identifies structures which violate CASL restrictions (e.g. gapped constructions), and informs users of syntactic errors (e.g. passives). It provides right and wrong usage examples for specific error types and gives a recommended rewrite for any sentence containing errors with a strong diagnosis.

Other validating parsers include the MULTILINT Controlled German Checker (Reuther and Schmidt-Wigger, 2000), the Caterpillar Technical English ClearCheck system (Kamprath, Adolphson, Mitamura. and Nyberg, 1998) and the Diebold Controlled English Checker (Moore 2000). MULTILINT works with technical documents in German and is capable of checking spelling, syntax, style and terminology. ClearCheck is an interactive checker which helps authors decide whether their writing conforms with Caterpillar Technical English. If it does not, the system suggests how to achieve conformance. The checker was developed in the 1990's and has undergone several updates. The two main features of the Diebold Controlled English Checker are the browser interface and the parsing software. While the checker gives recommendations for a sentence, the writer needs to edit the material at the level of meaning.

3 Specification of STS

3.1 Method

The first stage in producing a specification for STS was to decide on the reference corpus to be used. As Construcciones Aeronaúticas Sociedad Anónima (CASA) are the leading aircraft manufacturer in Spain an extract from a maintenance document totalling 1.45 Mb was obtained from them (CASA, 1997). The corpus consists of fifteen extracts each describing a particular procedure.

In order to develop the STS Writing Rules, we followed the following steps: First, we considered a potential writing rule for STS taken from the SE AECMA Guide. The 1998 version of this was used (Revision I) because this was what was available to us at the time when the research was carried out. Secondly, we decided if the rule was applicable to Spanish. In order to reach a decision, we tried to find evidence in the reference corpus. If we found examples in which the rule was infringed and the sentence in question caused ambiguity, the rule was adopted in STS. If on the contrary, the nature of Spanish language did not allow the rule to be applied, it was disregarded.

During the development of the STS Writing Rules, we came across some cases which required particular attention in the Spanish language. One example is the use of accents (´). Such cases gave us evidence to produce new rules which are only applicable to STS.

3.2 General Structure

We now discuss the results of our analysis. SE contains 57 Writing Rules, divided into nine families: 1. Words, 2. Noun Phrases, 3. Verbs, 4. Sentences, 5. Procedures, 6. Descriptive Writing, 7. Warnings and Cautions, 8. Punctuation and Word Counts, 9. Writing Practices. By contrast, STS contains 39 such rules divided into nine categories which are not exactly the same as those in SE: 1. Words, 2. Noun Phrases, 3. Prepositional Phrases, 4. Verbs, 5. Sentences, 6. Procedures, 7. Descriptive Writing, 8. Warnings and Cautions, 9. Punctuation and Word Counts. A tenth family (Writing Practices) is planned for the future when there is more practical experience of writing in STS. In the next paragraphs, we analyse those rules related to words, noun phrases and verb groups in SE and explain what decisions were involved in accepting or disregarding each of the rules in STS. The rules relating to other categories have not been explicitly commented upon as they are essentially the same in SE and STS.

3.3 Words

There are 13 rules in SE related to words:

1.1 Choose the words for procedures from:
-Approved words in the Dictionary,
-Words that qualify as Technical Names,
-Words that qualify as Manufacturing Processes.

1.2 Use approved words from the Dictionary only in the part of speech given.

1.3 Keep to the approved meaning of a word in the Dictionary. Do not use the word with any other meaning.

1.4 Use only those forms of verbs and adjectives shown in the Dictionary.

1.5 You can use words that are Technical Names.

1.6 Use a Technical Name only as a noun or an adjective, not as a verb.

1.7 Use the official name (shortened if necessary).

1.8 Do not use different Technical Names for the same thing.

1.9 If you have a choice, use the shortest and simplest name.

1.10 You can use verbs that are Manufacturing Processes.

1.11 Use Manufacturing Processes only as verbs, not as nouns or adjectives (unless the noun form qualifies as a Technical Name).

1.12 Once you choose the words to describe something, continue to use these same words (particularly Technical Names).

1.13 Make your instructions as specific as possible.

The rules applicable to Spanish are 1.1, 1.3, 1.8 and 1.13. Rule 1.1 has changed slightly in STS. While in SE it only applies to procedural writing, in STS it relates to both procedural and descriptive writing. The reason we have not included Rule 1.2 is that in Spanish the problem of a word representing several categories is not so common. A noun and a verb will never be mistaken as they have different forms, in contrast to what happens with the English language. Rule 1.4 had to be left out. Showing all the approved forms of verbs and adjectives is possible in English because there are not very many, but Spanish is an inflected language with many different forms of the same base word. We have decided not to include Rule 1.5 *You can use words that are Technical Names* because this rule is already implied by Rule 1.1. Rule 1.6 *Use a Technical Name only as a noun or an adjective, not as a verb* does not apply to Spanish because a Technical Name cannot function as a verb. Rule 1.7 *Use the official name (shortened if necessary)* is not included as the dictionary of common words and the domain-specific dictionary are restricted to official names and are therefore not supposed to contain any unofficial ones. Regarding Rule 1.9 *If you have a choice, use the shortest and simplest name*, we have ensured that this is the case by constructing a lexicon in which only the shortest and simplest names have been included. As a result, Rule 1.9 is redundant. We have decided not to include Rule 1.10 *You can use verbs that are Manufacturing Processes* because this rule is already implied by Rule 1.1. Rule 1.11 *Use Manufacturing Processes only as verbs, not as nouns or adjectives (unless the noun form qualifies as a Technical Name)* is not applicable to Spanish because a Manufacturing Process can never be a noun or adjective in Spanish. Rule 1.12 *Once you choose the words to describe something, continue to use these same words (particularly Technical names)* is not necessary. There is simply no option to use different words to explain the same thing because synonyms are not allowed in STS. The lexicon cannot include any synonyms, and all words that are allowed are contained in the lexicon.

3.4 Noun Phrases

There is a separate section in the SE Guide for rules relating to noun phrases. They are:

2.1 Do not make noun clusters of more than three nouns.

2.2 Clarify noun clusters that are Technical Names with one of these two methods:
-Use hyphens to show the relationship between the most closely related words;
-Explain the noun cluster. Then, if possible, use a shorter name after the initial explanation.

2.3 When appropriate, use an article (the, a, an) or a demonstrative adjective (this, these) before a noun.

The STS Guide has an equivalent section to Noun Phrases to cover Rule 2.3. The other rules of the section have been included in a new one called Prepositional Phrases because the difficulties inherent in noun clusters are different in Spanish. In English technical language, it is common to see phrases made from several nouns which are called noun clusters. The headword is usually at the end of the noun cluster and if this is very long, it can be confusing for the reader because it is difficult to distinguish which are the nouns that modify or describe the main one. An example is:

Runway light connection resistance calibration

In Spanish, nouns are not really clustered together. The main word comes first and the modifiers come afterwards in the form of prepositional phrases. When there are more than three prepositional phrases together it is difficult to understand which prepositional phrase modifies which. As a result, the overall meaning of the sentence is hard to comprehend. An example is shown below with prepositional phrases delimited by square brackets:

Se enciende el módulo ENG FIRE [en la central] [de aviso] [de fallos] [de la cabina] [de pilotos] [...] *[Doc_26-11-00]*

<*Itself lights the caption ENG FIRE on the panel of warning of failure of the cockpit of pilots [...]*>

{"ENG FIRE" caption comes on in the cockpit warning annunciator panel and warning signs flash.}

There are ways of simplifying this complex structure. For example, you can use relative clauses:

Se enciende el módulo ENG FIRE en la central de aviso de fallos que hay en la cabina de pilotos.

<*Itself lights the caption ENG FIRE on the panel of warni*ng of failure which is in the cockpit of pilots.>

3.5 Verbs

SE has 7 rules in relation to verbs:

3.1 Use only those forms of the verb that are listed in the Dictionary.

3.2 Use the approved forms of the verb to make only:
-The infinitive,
-The imperative,
-The simple present,
-The simple past tense,
-The future tense.

3.3 Use the past participle only as an adjective, either with a noun or after the verbs TO BE, TO BECOME.

3.4 Do not use the past participle with a form of the verb HAVE to make an unapproved tense.

3.5 Do not use the past participle of a verb with a helping verb to make a complex verb.

3.6 Use the active voice. Use only the active voice in procedural writing, and as much as possible in descriptive writing.

3.7 If there is an approved verb to describe an action, use the verb (not a noun or other part of speech).

The verb system in English is quite different from the verb system in Spanish. In consequence, the rules applicable to STS are Rule 3.2, Rule 3.6 and Rule 3.7. Many new rules had to be added in order to cover the requirements of Spanish. A summary is indicated below.

Rule 3.1 is a repetition of Rule 1.4. Therefore, it has not been included in the STS Guide. We decided not to list the forms in the STS Dictionary, as discussed in Section 4.5.2. Rule 3.2 was modified greatly because the Spanish verb system is quite different from the English one. With reference to the personal forms of the verb, the allowed tenses are the present and the future. There are no instances of the simple past tense in the reference corpus and therefore it has not been considered necessary. In fact, procedures will require mainly the use of the infinitive with an imperative value (instructions). For descriptions technical writers use the present simple and in some instances the future, but not the past simple, which is more used in literature than in technical documentation.

In contrast to English, the imperative in Spanish does not coincide in form with the infinitive. However, technical writers use the infinitive form to express an imperative value in Spanish technical documents. Therefore, the imperative tense has not been included.

With reference to the non-personal forms of the verb, the allowed tenses are the infinitive, the past participle and the gerund. The infinitive is allowed only in those cases in which it denotes an imperative value or a value of purpose. An example of the infinitive with an imperative value (instructions) is:

No exponer el cilindro extintor a temperaturas por encima de los 70 ºC. *[Doc_26-21-11]*

<*Not expose the cylinder extinguisher to temperatures as above of the 70 ºC.*>

{Do not expose the extinguisher cylinder to temperatures above 70 ºC.}

An example of the infinitive with a purpose value is:

El piloto tira de la palanca para energizar el sistema. *[Doc_26-21-00]*

<*The pilot pulls the handle in-order-to supply-power the system.*>

{The pilot pulls the handle to supply power to the system.}

There is ambiguity in the two cases described above. While the infinitive with a purpose value will always have the "para" particle {to}, the infinitive with an imperative value does not have any particle in front of it. In order to avoid ambiguity, the infinitive used with any other value should be changed into a personal form of the verb.

The past participle is only allowed with an adjectival value. An example is:

> Las tarjetas impresas están colocadas en paralelo dentro de una caja de acero inoxidable herméticamente cerrada. *[Doc_26-11-00]*
>
> *<The cards printed are placed/positioned in parallel inside of a case de steel stainless hermetically sealed.>*
>
> {The circuit boards are installed in sandwich form within a hermetically sealed stainless steel case.}

The gerund and progressive participle are not permitted in SE. However, in Spanish we find it very useful when we need to express the way in which to perform an action. Therefore, the gerund is allowed in such cases in STS. An example is:

> La alarma acústica puede anularse, pulsando cualquiera de los interruptores de las luces principales de aviso de fallos. *[Doc_26-11-00]*
>
> *<Le alarm acoustic may-be cancelled, pressing any of the switches of the sign master of warning of failure.>*
>
> {The audible alarm may be cancelled by pressing either of the master warning sign switches.}

In order to avoid ambiguity, the gerund used with any other value should be changed into a personal form of the verb.

In Spanish there are constructions called Perífrasis Verbales {Verbal Periphrases} which have also been accepted within the STS guide. A verbal periphrasis is a syntactical construction which consists of two or more contiguous verbs, none of which are any of the auxiliary verbs used to form complex verb tenses. Of all the combined verbs, one is the main verb (a non-personal verb form: infinitive, gerund or past participle) and the others are helping verbs. At present, the allowed verbal periphrases in STS are V. ESTAR + PAST PARTICIPLE (Only the forms está and están) and V. PODER + INFINITIVE (Only the forms puede and pueden).

With regard to the approved moods, the subjunctive was prohibited. The nature of technical documentation requires the use of the indicative (descriptions) and the imperative (instructions), as they add accuracy to the text. The use of the subjunctive mood can lead to unspecific statements.

Rules 3.3, 3.4 and 3.5 are not stated explicitly, but they are implicit within the other rules related to verb tenses in STS.

3.6 General Vocabulary

Work on the STS General Vocabulary involved finding equivalents for the words approved in SE or rather for the notions that the SE words express. This implied the study of the words in their natural context and therefore a thorough analysis of the reference corpus. At present the dictionary consists simply of a list of 875 words. It is envisaged that definitions could be added in a future project.

When the translation for a word resulted in several synonyms, the most commonly used word in the reference corpus was chosen. "One word, one meaning" was the most important criterion in the choice of most SE words. There were cases in which a single word in SE corresponded to various words in Spanish, depending on the precise context of use. A good example is the verb remove:

> REMOVE (v) To "take" or move something away from its initial position

As a consequence, in SE you find:

(1) *remove an indicator from the aircraft* (i.e. remove the screws etc that attach it),

(2) *remove equipment after maintenance* (i.e. push it off),

(3) *remove sharp edges* (i.e. eliminate them).
In Spanish, one word is not enough to explain all these concepts.

In case (1) *remove* means to separate an item (a constituent part of the aircraft) from its next higher assembly. This involves removing the attaching parts and then taking the item away. The most suitable word in Spanish would be *quitar*. In case (2) *remove* means to move something so that it is no longer where it was before. Maintenance equipment is not part of the aircraft, thus the word *quitar* is not completely adequate. There is a better word which expresses that meaning: *retirar*. In case (3) *remove* means to make a substance or material disappear. For that, the most appropriate word in Spanish is *eliminar*.

There are also cases in which for two or even three words in English we have only one word in Spanish. Examples include Time (n)/ Weather (n) = Tiempo, For (pre) / During (pre) = Durante, Adjust (v) / Tune (v) = Ajustar, and Glossy (adj) / Bright (adj) / Shiny (adj) = Brillante.

Error type	Description
1	Sentence length > 25/30 words (procedural/ descriptive sentences) (Rule 6.1 & 7.1)
2	Word not in General Vocabulary or terminology database (Rule 1.1)
3	Passive construction used (Rule 4.3)
4	3 attached prepositional phrases (Rule 3.1)
5	> 1 commands in the sentence (Rule 6.2)

Table 1 STS Parser Conformance Checks

Sent. Type	Errors Found	Real Errors Found	Real Errors	Prec.	Rec.
Descr	104	96	98	0.92	0.97
Proc.	51	40	42	0.78	0.95
All	155	136	140	0.87	0.97

Table 3 Results under Metric 1

Error Type	Description
6	Word used with an unapproved meaning (Rule 1.2)
7	Inappropriate omission of an article or demonstrative adjective (Rule 2.1)
8	Verb tense is illicit or the infinitive, past participle or gerund have been used with an unapproved value (Rule 4.1)
9	Unapproved verbal periphrasis used (Rule 4.2)
10	Subjunctive mood has been used (Rule 4.3)
11	Noun used to express an action, rather than a verb (Rule 4.5)
12	Structures used between coordinated sentences differ (Rule 5.3)
13	List of > 2 items occurs which should be a tabular layout (Rule 9.1)

Table 2 Errors Undetectable by the STS Parser

It is also necessary to add that the choice of the STS writing rules has an impact on the choice of words. For example, if the subjunctive is not permitted in STS, there is no point in approving the conjunction "hasta que" as an equivalent for "until" or "a menos que" as an equivalent for "unless", because they are followed by a sentence with a verb in the subjunctive, and are therefore unusable.

4. A Validating Parser for STS

4.1 Architecture

The basis of our grammar checker was a Robust Layered Parser which we had developed in a previous project (Sutcliffe, 2000). It takes the form of a cascade of context-free recognisers operating on a text tagged for part-of-speech. Each pass looks for one kind of construct and uses as input a combination of terminal symbols (i.e. word / part-of-speech tag pairs) and non-terminal symbols (i.e. instances of constructs recognised in previous passes). The stages in parsing can be summarised as follows:

- Pre-processing to perform term recognition;
- Tagging using the Xerox Xelda system (Xelda, 2002);
- Three parsing parses: Simple Noun Phrases (snp), Prepositional Phrases (pp) and Verb Groups (vg);
- Post-processing to carry out conformance checking.

The checks currently implemented in the last stage are:

- Sentence Length,
- Words in Dictionary,
- Passive Voice,
- Number of Prepositional Phrases Modifying a Noun,
- Multiple Commands in a Sentence.

The Sentence Length check counts the number of words in the sentence. It distinguishes between descriptive and procedural sentences. If the sentence is descriptive the maximum number of words allowed is 30. In the case of procedural sentences, the number of allowed words is reduced to 25. Punctuation is not counted in this process. If the number of words has been exceeded, the system outputs an error message.

The Dictionary check verifies that all words in the sentence are valid. To carry out this task this routine makes sure that all the words in the sentence, except those which have been recognised as technical terms, are in the STS Dictionary database. If there are words which are not in this database, an error message will be output for such words.

The Passive Voice check identifies those cases in which a passive voice sentence has been included. If one of these constructs is detected an error message is output.

Error Type	1	2	3	4	5	Oth -er
Descr.	21 14.7	51 35.7	11 7.7	15 10.5	0 0	45 31.5
Proc.	7 12.7	8 14.5	0 0	5 9.1	22 40.0	13 23.6
All	28 14.1	59 29.8	11 5.6	20 10.1	22 11.1	58 29.3

Table 4 Error Type Occurrences in Descriptive and Procedural Sentences. The upper figure in each box is the count while the lower figure is the percentage.

Sent. Type	Errors Found	Real Errors Found	Real Errors	Prec.	Rec.
Descr.	104	96	143	0.92	0.67
Proc.	51	40	57	0.78	0.70
All	155	136	200	0.87	0.68

Table 5 Results under Metric 2

The Number of Prepositional Phrases Modifying a Noun check searches for those cases in which there are more than three adjacent prepositional phrases in a sentence. The objective is to reduce cases of structural ambiguity.

Finally, the Multiple Commands in a Sentence check counts the number of commands in a procedural sentence. If there is more than one, an error message is output.

4.2 Evaluation

The evaluation of the parser was carried out in the following manner: Firstly, we selected 100 sentences from the reference corpus, 50 of which were descriptive and the other 50 procedural sentences. These contained examples of all error types detectable by the parser, although a few of them did not contain any of these errors.

Next, we went through each sentence by hand determining all the errors present. These included all the five types of error detectable by the parser, plus any other errors found. The following stage was to run the sentences through the parser. We then compared the output of the parser with the results of the manual analysis.

Sent. Type	6	7	8	9	10	11	12	13
Descr.	1 2.2	10 22.2	25 55.6	1 2.2	3 6.7	4 8.90	1 2.2	0 0
Proc.	0 0	7 46.7	3 20.0	1 6.7	2 13.3	0 0	0 0	2 13.3
All	1 1.7	17 28.3	28 46.7	2 3.3	5 8.3	4 6.7	1 1.7	2 3.3

Table 6 Undetectable Error Type Occurrences in Descriptive and Procedural Sentences

Two evaluation metrics were adopted. Metric 1 uses only those errors which the parser was designed to detect (see Table 1, Types 1-5). Metric 2 uses all errors found during manual analysis including those which the parse was not designed to detect (see Table 2, Types 6-13). Following standard practice in parser evaluation (Grishman, Macleod and Sterling, 1992), the Precision and Recall measures of information retrieval (van Rijsbergen, 1979) were used to summarise the results.

Table 3 shows the results under Metric 1. It can be concluded that the parser meets in some measure its design criteria having Precision 0.87 and Recall 0.97.

A breakdown of error types by sentence type (descriptive or procedural) is shown in Table 4. It shows that Error Type 2 (word not in general vocabulary or terminology database) is the most common in descriptive sentences (35.66%) while it is Error Type 5 (more than one command in the sentence) that is the most frequent in procedural sentences (40%). The least common error in descriptive sentences is Error Type 5 with 0 occurrences, while in procedural sentences it is Error Type 3 (passive construction used) also with 0 occurrences. If we observe all the sentences independently of their type, the most recurrent error is Type 2, followed by the "other" error type category with a very similar figure.

Table 5 summarises the results using Metric 2, which is a measure of how effective the parser is at detecting non-conformance with the STS specification as a whole. Recall figures under Metric 2 are lower than those under Metric 1. 68% of all errors are detected by the STS Parser. This figure shows that although the parser is still useful, it could be improved if were to add new conformance checks able to detect errors of Types 6 to 13.

Table 6 shows the distribution of

undetectable error types in both procedural and descriptive sentences. We can observe that the most common error in descriptive sentences is Error 8 (verb tense illicit), while in procedural sentences is Error 7 (omission of article or demonstrative adjective). The least common error in descriptive sentences is Error Type 13 (list of more than two items not in a tabular layout) with 0 occurrences, while in procedural sentences it is Error Types 11 (noun rather than verb used to express an action) and 12 (different structures coordinated together) also with 0 occurrences.

8. Summary and Conclusions

The objectives of the work were to create a CL specification for technical documentation in Spanish and to develop a checker to help technical writers conform to it. In order to achieve these, previous work on CLs was investigated, a method for the design of a Spanish CL specification was established, research concerning the linguistic features of currently used CL Checkers was carried out, a specification for STS was drawn up, and a validating parser for STS was built and evaluated.

This project has contributed towards the improvement of technical documentation in Spanish as it is the first Simplified Spanish specification that has been developed for a technical domain. The STS Parser could form the basis of a tool for technical writers to help them implement the STS Rules.

Future work could improve both the STS Specification and the validating parser. With regard to the STS Specification, it does not currently include the meanings of the words contained in the STS General Vocabulary. These could be adapted from the dictionary of the Spanish Language Royal Academy which is the principal authority for the language. Secondly, we would need to associate a meaning with each of the technical terms in our terminology database. This would involve the study of each term followed by agreement among technical terminologists as to which meaning to assign to it. This should not be a very difficult task as the meanings of words belonging to technical domains are already restricted.

Improving the STS Parser would involve adding new checks to take account of the eight error types identified in Table 2. Some of these would involve overcoming a number of difficulties, while others would be relatively easy to implement.

References

AECMA 1998. *AECMA Simplified English, Issue 1, Revision 1. AECMA Document: PSC-85-16598. A Guide for the Preparation of Aircraft Maintenance Documentation in the International Aerospace Maintenance Language*. AECMA, Brussels .

Almqvist, I. and Hein, S. A. 1996. Defining ScaniaSwedish – A Controlled Language for Truck Maintenance. *Proceedings of the First International Workshop on Controlled Language Applications (CLAW96), Leuven, Belgium, Katholieke Universiteit Leuven Centre for Computational Linguistics 26-27 March, 1996*, 159-165.

Barthe, K. 1998. GIFAS Rationalised French: Designing one Controlled Language to Match Another. *Proceedings of the Second International Workshop on Controlled Language Applications (CLAW98), Pittsburgh, Pennsylvania, Carnegie Mellon University, Language Technologies Institute, 21-22 May, 1998*, 87-102.

CASA 1997. CN-235 *Manual de Mantenimiento*.

Douglas, S. and Hurst, M 1996. Controlled English Support for Perkins Approved Clear English (PACE). *Proceedings of the First International Workshop on Controlled Language Applications (CLAW96), Leuven, Belgium, Katholieke Universiteit Leuven Centre for Computational Linguistics 26-27 March, 1996*, 106-114.

Ericsson 2000. http://www.ericsson.com

Fuchs, N. E. and Schwitter R. 1996. Attempto Controlled English (ACE). *Proceedings of the First International Workshop on Controlled Language Applications (CLAW96), Leuven, Belgium, Katholieke Universiteit Leuven Centre for Computational Linguistics 26-27 March, 1996*, 124-136.

Godden, K. 2000. The Evolution of CASL Controlled Authoring at General Motors. *Proceedings of the Third International Workshop on Controlled Language Applications (CLAW 2000), Seattle, Washington, 29-30 April, 2000*, 14-19.

Grishman, R., Macleod, C., and Sterling, J. 1992. Evaluating parsing strategies using standardised parse files. *Proceedings of the Third ACL Conference on applied Natural Language Processing, Trento, Italy*, 156-161.

Holmback, H., Duncan, L. and Harrison, P. 2000. A

Word Sense Checking Application for Simplified English. *Proceedings of the Third International Workshop on Controlled Language Applications (CLAW 2000), Seattle, Washington, 29-30 April, 2000*, 120-133.

Kamprath, C., Adolphson, E., Mitamura, T. and Nyberg, E. 1998. Controlled Language for Multilingual Document Production: Experience with Caterpillar Technical English. *Proceedings of the Second International Workshop on Controlled Language Applications (CLAW98), Pittsburgh, Pennsylvania, Carnegie Mellon University, Language Technologies Institute, 21-22 May, 1998*, 51-61.

Kodak 2000. http://www.kodak.com

Lamiroy, B. and Gebruers, R. 1989. Syntax and Machine Translation: The METAL project. *Lingvisticae Investigationes* 13(2): 307-332.

Means, L. and Godden, K. 1996. The Controlled Automative Service Language (CASL) Project. *Proceedings of the First International Workshop on Controlled Language Applications (CLAW96), Leuven, Belgium, Katholieke Universiteit Leuven Centre for Computational Linguistics 26-27 March, 1996*, 106-114.

Means, L., Chapman, P. and Liu, A. 2000. Training for Controlled Language Processes. *Proceedings of the Third International Workshop on Controlled Language Applications (CLAW 2000), Seattle, Washington, 29-30 April, 2000*, 51-61.

Moore, C. 2000. *Controlled Language at Diebold, Incorporated. Proceedings of the Third International Workshop on Controlled Language Applications (CLAW 2000), Seattle, Washington, 29-30 April, 2000*, 51-61.

Reuther, U. and Schmidt-Wigger, A. 2000. Designing a Multi-Purpose CL Application. *Proceedings of the Third International Workshop on Controlled Language Applications (CLAW 2000), Seattle, Washington, 29-30 April, 2000*, 72-82.

Schachtl, S. 1998. Controlled Siemens Documentary German and TopTrans, *Technical Communication Forum*, http://www.tc-forum.org/topictr/tr9contr.htm

Smartny 2002. http://www.smartny.com/

Sutcliffe, R. F. E. 2000. Using a Robust Layered Parser to Analyse Technical Manual Text. *Cuadernos de Filología Inglesa*, 9(1): 167-189. *Número Monográfico: Corpus-based Research in English Language and Linguistics*, Murcia, Spain, University of Murcia.

van Rijsbergen, C. J. 1979. *Information Retrieval*. Butterworths, London:.

Wojcik, R. 1998. AECMA Simplified English and Controlled Language Checkers. *The ELRA Newsletter*.

Wojcik, R., Harrison, P. and Bremer, J. 1993. Using Bracketed Parses to Evaluate a Grammar Checking Application. *Proceedings of the Thirty First Annual Meeting of the Association for Computational Linguistics, Ohio State University, 22-26 June, 1993*, 38-45.

Wojcik, R., Holmback, H. and Hoard, J. 1998. Boeing Technical English: An Extension of AECMA SE beyond the Aircraft Maintenance Domain. *Proceedings of the Second International Workshop on Controlled Language Applications (CLAW98), Pittsburgh, Pennsylvania, Carnegie Mellon University, Language Technologies Institute, 21-22 May, 1998*, 114-123.

Xelda 2002. XeLDA: *Xerox Linguistic Development Architecture*, http://www.xrce.xerox.com/ats/xelda/

Xerox 2001. http://www.xrce.xerox.com

Zhang, W. and Shiwen, Y. 1998. Construction of a Controlled Chinese Lexicon. *Proceedings of the Second International Workshop on Controlled Language Applications (CLAW98), (Pittsburgh, Pennsylvania, Carnegie Mellon University, Language Technologies Institute, 21-22 May, 1998*, 159-173.