ANESTTE: a writer's assistant for a specific purpose language

Juan M. Montero

Departamento de Ingeniería Electrónica

ETSI Telecomunicación
Universidad Politécnica de Madrid
Ciudad Universitaria S/N
28040 Madrid Spain
juancho@die.upm.es

M. Mar Duque

Departamento de Filología Inglesa aplicada a la Ciencia y a la Tecnología

ETSI Telecomunicación
Universidad Politécnica de Madrid
Ciudad Universitaria S/n
28040 Madrid Spain
marduque@etsit.upm.es

Abstract

This paper presents a new automatic tool for assessing the linguistic quality of scientific papers written in English. A set of complex lexical and syntactic surface-level rules compute more than 80 style-related variables. Their combination defines the score of a text in the four dimensions of style assessment for scientific papers: clarity, variety, conciseness and conviction. The software has been tested on 60 published articles and incorporates an animated agent that acts as a personal assistant and explains the results to the writer and how to improve them.

1 Introduction

After the emergence of English as a global language, it has become the predominant language in science and technology publications. Proficiency in written English is a necessary skill for current scientists all over the world, as they must publish their research results in international conferences and journals. Similar remarks can be applied to other areas of activity such as business or education: to succeed, people need to express their written ideas effectively and in English. As this is a difficult task for non-native English speakers, it has provoked a rapid increase in the interest on studying English for Specific Purposes.

One of the best ways to improve the quality of texts would be the use of good automatic computer-assisted text analyzers and style checkers. These software tools could aid in the creation of accurate English-language documents. They could be especially useful when adapted to the rules and conventions of a specific genre, such as technical or scientific articles. The top performance would be achieved when these programs are trained to verify a simplified or a controlled language, making the translation easier.

Nowadays spell checkers (software tools that analyze the lexical component of a text, and search for the use of misspelled words or mechanical errors), are usually part of any state-of-the-art word processing suite. These computer programs also incorporate some kind of grammar checker (number agreement errors, the use of unacceptable syntactic structures…) that outputs one or several simple readability measures such as Flesch Reading Ease index (Flesch 1948) or Flesch-Kincaid Grade Level. Among the complementary linguistic information that is taken into account by these programs, we can find: the length of noun clusters, the balance between the use of active and passive voice, the structure of the clauses and sentences (simple vs. coordinate or subordinate), etc. Nevertheless, thorough computer-based style analyzers are still a matter of NLP research and there is much room for improvement.

Authoring and proofreading tools can also incorporate some standardization of the typographical style of the documents (fonts, margins…), or can include facilities for sharing documents or checking consistency in a collaborative writing environment (Glover 1996). Finally, they are usually augmented with additional help documentation on grammar basics or examples of use (although in most cases they are designed with native speakers of English as the target users).

In this context of writer's assistants, we have developed a new software tool called ANESTTE (ANalyzEr of Style for Technical Texts in English). Our work has been focused on reviewing the text

style for scientific and technical writers, especially when they are non-native and need to write research articles for international journals in English. We have not developed any tool for spelling or grammar checking, because our interest is in a higher-level concept of style as described below.

By addressing this genre or this specific language, we aim at providing a deeper and more complex analysis. Although including rules and formulas developed for a specific purpose language, the tool is fully configurable and could be easily customize to any other register or genre.

Another important feature in our system is that incorporates an HTML guide on style excellence and an animated agent. This agent explains the results of the analysis to the users by means of speech synthesis and speech recognition (Montero 2000), and explains how to achieve a better overall quality in their writing style. The critiquing agent comments which concrete linguistic variables must be increased or decreased to improve the style of the text, and it can display the HTML pages that explain the meaning of the variables in the native language of the writer (our system is mainly aimed at Spanish technical or scientific users).

2 Style analysis

Modern stylistic studies comprise two areas. One is called Literary Stylistics, and is related to the description of the use of lexical and grammar variations in literary works. The other area is Linguistic Stylistics, that describes the formal characteristics of a text in terms of linguistic variation, by means of qualitative and quantitative analysis.

2.1 Linguistic Style Analysis

The reference frame for Linguistic Style Analysis can be a text genre (for instance, the scientific or technical research article), a register or even a specific author.

A statistical approach in Stylistics (e.g. to compute the mean length of words or the percentage of passive-voice sentences) can be very useful for the development of automatic applications such as forensic ones: authorship attribution that determines the identity of the author of a given text. It could be also helpful in writer's assistants, for the determination of the estimated reading level needed to comprehend a written text, or the adequacy of the text to the publishing context: journal, conference… (Sharples 1992). Another important area of application is the automatic detection of genre: to assign electronic documents to a taxonomy of pre-defined subject categories (Kessler 1997).

For this style evaluation process, a certain norm must be defined. If we are studying the appropriateness of one particular paper, we need to define a set of measures to determine the distance between our sample paper and the abstract good-quality technical paper. Mathematically, this distance (based on measurable characteristics) is the norm of the style for this genre (probably part of this norm is general and, therefore, applicable to other genres or registers).

Previous approaches made use of several natural language strategies that now we can combine:

- **If-then-else rules**: they establish conditions on linguistic variables, such as the mean number of words per sentence or the maximum number of words per sentence (McGowan 1992). These rules should depend on the genre of the text, the audience of the text… In our case (scientific and technical papers), the context is homogeneous, and can be modeled with only one set of rules.
- **Parametric formulas**: as in Critique (Jensen 1993), that uses non-Boolean acceptability limits and weighting factors (for instance, applied to the assessment of the distance between subject and verb). Thus, we can provide a more detailed analysis that allows relaxed or fuzzy conditions.
- **Lexical analysis**: as included in most proofing tools. It is used to reject the use of informal words, clichés or contractions in a formal style, to detect long words that make a text more difficult to read, etc.

2.2 Technical writing style features

The language of scientific and technical papers as published in international journal and proceedings, constitutes a Language for a Specific Purpose. It must provide an efficient means of communication between research workers and must stimulate the development of future research activities (Lundberg 1994).

Although the contents of the articles are the central point for achieving these objectives (guided by experimentation, innovation and the scientific method), the linguistic aspects of the paper also play an important role for providing efficiency and attractiveness to the text.

Four sub areas of linguistic analysis can be distinguished (Duque 2000):

- **Clarity**: the meaning must be transparently conveyed in an effortless manner. Among the enemies of clarity, we can list:
  - long noun or preposition clusters,
  - long sentences,
  - an excessive use of passive voice,
  - the abundance of abstract nouns or static verbs,
  - the lack of connecting elements, that bring cohesion to the discourse.
- **Conciseness**: the use of the right words and just the right words. Against this virtue we can find:
  - an excess of nominalizations,
  - the use of redundant or unnecessary expressions.
- **Variety**: it brings attractiveness to the communication process. Monotonous features that impoverish a text can be:
  - the repetition of the same sentence structures,
  - an excessively impersonal tone,
  - the lack of richness in terms of sentence and paragraph lengths, or in terms of types of verbs.
- **Conviction**: it expresses the confidence of the author in the information that the text conveys. The abuse of certain modal verbs and expressions of doubt (or probability) can decrease the conviction of a paper, transmitting the feeling that the author lacks security in his/her assertions.

In addition to this, semantic constraints should be imposed, constraints that would check for textual coherence or textual progression. As the state-of-the-art automatic semantic analysis is not reliable yet, our system will not model the meaning of the words or the logical flow, but the syntactical structure of the sentences.

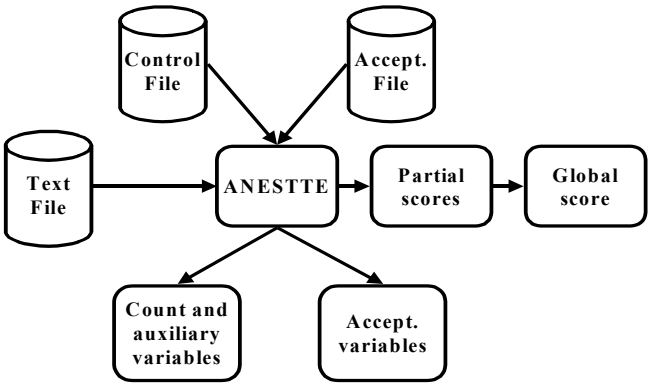

Figure 1. Process of analysis

3 ANESTTE system

Our system follows a knowledge-based approach, using a set of specific linguistic rules to count the occurrence of most of the features that are relevant for the analysis of style. Upon these variables, a set

of acceptability formulas is applied. Thus, using a two-phased approach (Figure 1), the system arrives at one general style score of the text, and four particular scores in the sub-areas of analysis, as mentioned above.

These linguistic context-sensitive rules perform a robust parsing and disambiguation of unrestricted text (we are not dealing with a controlled language, but with the complexities of technical texts). In this analysis, the local grammars used by the rule engine, concentrate on the most common surface patterns of scientific English.

3.1 Counting Phase

This step of the analysis process is guided by the general control script file. The file has already been preprocessed: detection of phrasal verbs, contractions and abbreviations, using a special configurable preprocessing file. The control file contains a sequence of commands that will be processed by the analyzer in a sequential way. The current version of the ANESTTE control file contains up to 80 style-related variables. All the files admit comment lines.

Two types of commands are available:

- **Counting commands**: they specify the grammar rule file that must be used to compute a certain linguistic primary variable:
  - Percentage of abstract nouns.
  - Percentage of…
- **Computing commands**: they allow to define new auxiliary or secondary variables, and to apply arithmetical, logical and comparison operations on pre-computed counting variables. These operations are: addition, subtraction, multiplication or division, conditional assignment, percentage, etc.

In addition to these commands and user-defined variables, up to 14 predefined system variables can be used (ranging from number of words, sentences or paragraphs, to Gunning or Flesch indices).

An example of computation is:

GlobalScore= ConvictionScore + VarietyScore +

ClarityScore +

((*ConcisionScore*>7) @ (C*oncisionScore*-7))

In this equation, the variable *GlobalScore* is the sum of the other scores (with the exception of *ConcisionScore*, that needs to be scaled down when it is greater than seven).

3.2 Grammar Rule Files

The core of the grammatical system is the set of contextual hand-coded rule files (Duque 2000). They specify the conditions to increment the count of the associated variables.

The text is iterated on a word-by-word basis and each rule in each file is checked against the stream of input words. Each line of the rule files defines a condition for incrementing the counter of a variable. Every time a line matches the input text at current position, the counter is incremented. A line contains a sequence of elements such as:

- **Words or punctuation marks**, that should be present or should not be present (the matching can be carried out in a case-sensitive or in a case-insensitive way).
- **Endings of words**: we can specify conditions that force the presence or the absence of a certain ending in the input words.
- **Syntactical tags**: according to a dictionary search.
- **Another rule file**: rule files are hierarchical and one rule line can contain a reference to another whole rule file, allowing a reuse of rules in several files, complex recursive patterns, etcetera. A rule file element is satisfied when at least one of its rule lines matches the text at current position.
- **A logical combination of conditions**: we can impose the presence (or absence) of a word (or a word ending or a syntactical tag) and, at the same time, the satisfaction of another rule file.

We can ask for the positive or negative satisfaction of two rule files, or for the positive matching of one rule file and the negative matching of another one.

Although the language of rule files is mainly declarative, some procedural elements are necessary for the detection of certain structures:

- **A sequential search** in the text for the matching of a file (it usually contains conditions for the determination of the end of a sentence). While the system is going forward, it must also count the occurrence of the elements from another rule file. For the satisfaction of the complex rule at the end of the search, the counter must have a certain configurable value (it must be zero, or greater than a minimum value…). This rule is not applied in a word-by-word basis. Each search moves forward the current analysis pointer.
- The same rule as above, but with **word-by-word processing** if the element is not satisfied.
- **A modifier** for an element that checks for the presence of this element, but does not change the analysis pointer.

More than 80 counting and computation rules have been written. They reference 324 carefully designed rule files.

Some examples of rule lines (that describe infinitive phrases) are:

- [Infi-0] **to** +*infinitive*
- [Infi-0] +*infinitive*
- [Infi-0] [Object] +*infinitive*
- [Infi-0] [Object] **to** +*infinitive*
- [Infi-0] **for** [Object] **to** +*infinitive*

The Infi-0 rule file contains a list of verbs that can be followed by **to**, followed by *infinitive*. The file called Object describes typical grammatical-object structures, and it contains lines such as:

- +Pronoun
- [Det1]+*adjective*_<Adject-0> +*noun*_<Noun-0>

The element +*Infinitive* specifies that the last word in the sequence has to be tagged as a verb in the dictionary.

Emboldened words are words that must be literally present in the text at the specified relative location (after the first infinitive, before the last one, etc.)

The rule +*adjective*_<Adject-0> matches adjectives that are not included in the Adject-0 rule file.

For writing and checking all these rule files, we have created a development environment with an integrated editor and a rule file parser (that checks the syntax of the file), the analyzer software and final dialog windows that show the detailed results of the analysis (Figure 2).

3.3 Acceptability Phase

Using the primary and secondary variables computed at the first phase, the acceptability files contain a set of formulas that define the Boolean features that characterize a good writing style. These new variables will be true when a range-condition formula is satisfied.

These formulas specify the minimum and maximum acceptable values for the appropriate combinations of the first-phase variables. These acceptability formulas can be grouped in four main categories of analysis:

- **Clarity formulas**: they check which clarity related variables are within the acceptable range:
  - less than 30% of passive voice sentences,
  - less than 50% of abstract nouns.
  - …
- **Conciseness formulas**: that check which Conciseness-related variables are within the acceptable range:

- less than 1.3 compound prepositions,
- average sentence length between 20 and 28 words.
- …

- **Conviction formulas**: that check which conviction related variables are within the acceptable range:
  - predominance of strong modal verbs,
  - predominance of conviction adverbs over doubt adverbs.
  - …
- **Variety formulas**: they check which variety related variables are within the acceptable range:
  - A balance between short, medium and long sentences or paragraphs,
  - a balance between static, dynamic and modal verbs.
  - …

The number (or a range) of satisfied features determines the local score of the text in each category, and the combination of these four scores is the general evaluation of the text.

4 Evaluation of the analyzer

There have been two types of evaluation processes:

- **Objective evaluation**: using the debugging version of the software, we have checked the correctness of the linguistic analysis. This debugging version applies the rules to a text and outputs the rule files, the specific rules lines and the fragments of texts that match each rule. This output allows for an easy revision of the precision of the analysis (we look for maximizing precision, not recall).
- **Evaluation with published papers**: we analyzed 60 papers that were previously published in international journals (average number of pages: 12). The test confirms that mean global scores and their distribution for native and non-native papers are not significantly different under ANOVA and chi-square tests with 0.95 confidence level (all the papers were refereed and revised). More than 90% of the papers receive more 5 points in a 7-point scale, confirming the quality of the analyzer on published papers.

5 The dialogue component

Although the general and sub-area results from the ANESTTE analysis can be very useful for a technical writer, an assistant must provide an interactive help system.

The system can access up to 116 HTML connected pages that have been developed for explaining the elements of style and the acceptability rules that define a good style (Figure 2).

An animated agent (Gustafson 1999) explains the scores at the sub areas, both the good ones and the bad ones (Figure 3). A score is considered good (or remarkable) when it is above the global score (and above a certain minimum score: we should not praise a sub-area that is below this value, because we would be praising a really bad style). Using this strategy, most users receive a certain praise (some of the scores will be above the mean and, at least, are more positive than the others).

A score is considered not good (or bad) when it is below the mean (or below a certain minimum score). For these scores (some of them can be good in absolute values), the agent asks the user whether she/he wants more details. If the answer is positive, the agent establishes a new objective for the less positive sub-areas (not more than two: we cannot overflow the user with many faults) and suggests a set of variables for the improvement in the worst sub-area. The number of suggested variables depends on the specific score. For each variable, the agent shows the corresponding HTML help file (although a general-purpose search engine is also included in the environment).

As the rest of the system, the dialogue component has been designed in a fully configurable way (San-Segundo 2001). It receives a database of abstract results of analysis, and a set of reference values. A script file describes the sequence of multimodal interactions with the user, and specifies the

operations on the input database that determine the parameters of the dialogue (the good and the bad sub-areas, the degree of praise, an objective for improvement…).

6 Conclusions

We have designed, implemented and tested a new tool for helping technical writers improve the quality of their papers in terms of linguistic style. The program performs syntactical and lexical analyses to check a paper for clarity and conciseness of linguistic expression, and variety and conviction in writing style.

The tool has been tested on a corpus of published papers, verifying the correctness of the analysis and confirming the hypothesis that published papers from non-native speakers are stylistically very close to native-speakers papers.

The new writer's assistant includes a set of web pages that explain the relevant linguistic elements to non-native users, describing the acceptability rule that must guide the way they write. An animated agent makes comments on the results and suggests some weak points that should be improved.

Acknowledgement

This work has been partially funded by Universidad Politécnica de Madrid through its A9904 project. Special thanks to José Manuel Pardo, for his comments; to Georgina Cuadrado, for her linguistic help; to Rogelio Vargas, Pilar Santamaría and Jesús Heras, who have helped in the programming tasks, and to all people that work with us at ETSIT-UPM.

References

Duque M.M. 2000. Manual de Estilo. El arte de escribir en inglés científico-técnico. Madrid, Paraninfo.

Flesch R. 1948. A new readability yardstick, *Journal of Applied Psychology*, 32: 221-233.

Glover A., Hirst G. 1996. Detecting stylistic inconsistencies in collaborative writing. In Sharples M., van der Geest T. (Eds.) *The new writing environment: Writers at work in a world of technology*. London: Springer-Verlag, pp. 147-168

Jensen K., Heidorn G. E., Richardson S. 1993. *Natural Language Processing: the PLNLP approach.* Boston: Kluwer.

Kessler B., Numberg R., Schütze, H. 1997. Automatic detection of text genre. In *Proc. of the 35th Annual Meeting of the ACL*, pp. 32-38.

Lundberg G. 1994. The Functions of Refereed Scientific Journals. In Weeks R., Kinser, D. (Eds.) *Editing the Scientific Refereed Journal. Practical, Political and Ethical Issues*. IEEE Press, pp. 1-6.

McGowan S. 1992. Ruskin to McRuskin: Degrees of interaction. In *Computers and Writing: State of the Art*, P. O'Brian and N. Williams (Eds.), Oxford: Kluwer, pp. 297-318.

Montero J.M., Córdoba R., Vallejo J.A., Gutiérrez-Arriola J., Enríquez E., Pardo J.M. 2000 Restricted-Domain Female-Voice Synthesis in Spanish: from Database Design to ANN Prosodic Modeling. In *Proc. of the International Conference on Spoken Language Processing,* Beijing.

San-Segundo R., Montero J.M., Colás J., Gutiérrez- Arriola J., Ramos J.M., Pardo J.M. 2001 Methodology for Dialogue Design in Telephone-Based Spoken Dialogue Systems: a Spanish Train Information System. In Procceedings of EUROSPEECH'01.

Sharples M., Goodlet J., Pemberton L. 1992 Developing a Writer's Assistant. In J. Hartley (ed.) *Technology and Writing: Readings in the Psychology of Written Communication.* London: Jessica Kingsley, pp. 209-220.

Figure 2. The result of the analysis and an HTML page that instructs on the use of the passive voice.

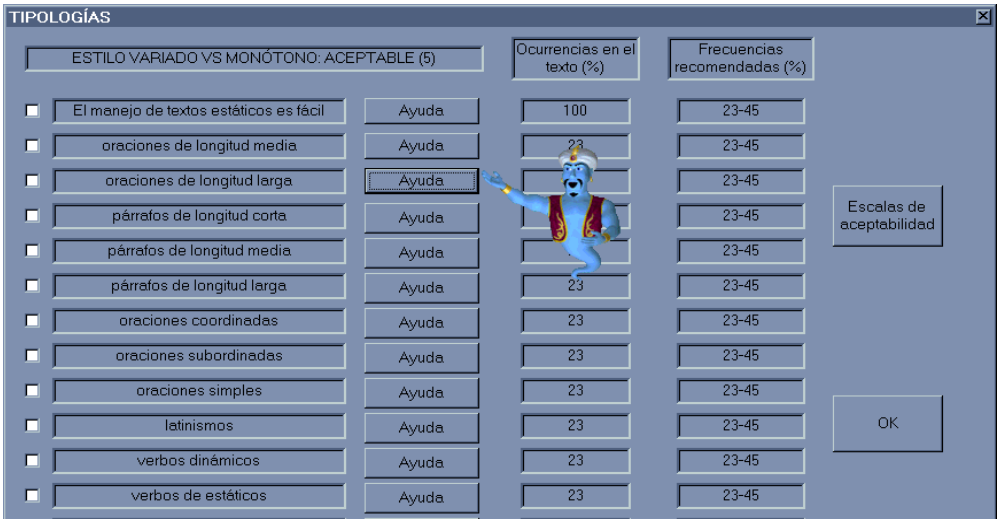

Figure 3. An animated agent explains the analysis results to the user.