Published in Lectures Notes in Artificial Intelligence 1010, pp. 121-132

A Case-Based Approach for Developing Writing Tools Aimed at Non-native English Users

Sandra M. Aluísio[1] and Osvaldo N. Oliveira Jr.[2]

[1] Universidade de São Paulo
Departamento de Ciências de Computação e Estatística
CP 668, 13560-970, São Carlos, SP, BRAZIL
e-mail:sandra@icmsc.sc.usp.br
[2] Universidade de São Paulo
Instituto de Física de São Carlos
CP 369, 13560-970 São Carlos, SP, BRAZIL
e-mail:chu@ifqsc.sc.usp.br

Abstract

A writing tool has been developed for helping non-native English users to produce a first draft of Introductory Sections of scientific papers. A corpus analysis was carried out in 54 papers of Experimental Physics which allowed one to identify the schematic structure of Introductions and 30 rhetorical strategies generally employed. Each one of the Introductions analysed constituted a case. The user chooses from menus features related to the rhetorical strategies for each component and gives the intended order for his/her Introduction, thus forming the requisition. Using three types of metric, the tool recovers the best-match cases that can be later modified in a revision process. Preliminary experiments showed that high precision and recall will only be obtained if the number of cases in the case base is considerably increased. In the revision process, four operations are suggested which consist in modifying/adding/deleting the different rhetorical messages that constitute the strategies of the chosen case.

1 Introduction

Writing scientific papers in English is often a severe limitation for the career of many a non-native English user. In addition to difficulties in choosing appropriate rhetorical, grammatical and semantic itens, non-natives also face problems related to interference from their mother tongue (L1 interference). In order to overcome some of these difficulties, we proposed an empirical resource based on offering the writer a set of standard sentences and collocations which appear with high frequency in scientific texts [Oliveira-91]. A reference writing tool integrated in the AMADEUS (AMiable Article Development for User Support) environment was then developed [Caldeira-92] which provided excellent results when applied by some Brazilian graduate students [Fontana-93], for they were able to get started with their writing task and could produce good pieces of text. These particular users generally already had a good reception of the English language. However, the tool was not successful in helping less experienced writers, with poorer knowledge of English. This occurred mainly because such users had difficulties in localizing the collocations that were appropriate to their needs and, worst of all, they could not make sensible choices of cohesive links for building up a paragraph. It then became clear that these less

experienced users require a writing tool which deals specifically with collocations in real contexts so that cohesive devices may be readily apparent. The utilization of reusable texts has already been exploited for producing standardized texts [Born-92], for reducing the duration of writing tasks [Buchanan-92] and in natural language generation or interpretation systems by way of phrasal lexicons [Kukich-83; Jacobs-85; Hovy-90; Smadja-91], but to our knowledge no tool has been devoted to non-native English users.

We take the view that a tool with the characteristics mentioned above could make use of the case-based approach, as a case would not only illustrate the rhetorical structure[1] of a given piece of text but would also instantiate the cohesive devices with real text. While the case-based approach applied to a writing tool for non-native English users is new, there are few existing case-based systems for natural language generation. One example is a case-based system that stores cases of actual utterances and uses them in order to generate sentences in the target language of a speech-to-speech dialog translation system [Kitano-90]. Another one is the LetterGen system for the generation of business letters in four languages [Pautler-94].

Our approach required a systematic corpus analysis of a number of papers, which is described in Section 2. The development of a writing tool for Introductory Sections of Experimental Physics papers is reported in Section 3. Upon selecting from menus various intended features for the Introduction, the user is returned real Introductions from the corpus which better map onto his/her requisition. It must be stressed that even if the recovered cases reflect the chosen features, this is still not sufficient. The user must be able to adapt/revise the case to input his/her own material. Suggestions are given to the user of the possible ways for revising the first draft. Section 4 discusses these issues, highlighting the strengths and limitations of the case-based approach for the development of this specific type of writing tool.

2 Schematic Structure of Introductory Sections

A corpus-based analysis was carried out for Introductions of experimental physics papers. Fifty-four (54) papers were selected at random from the years 1992-1994 from the journals Physical Review Letters (33 papers) and Thin Solid Films (21). The choice of these journals was based on their quality and also on the fact that their Introductions are reasonably short and standardized. In spite of its length, the Introduction in Physical Review Letters is more elaborate than in Thin Solid Films, for the intended audience is broader in the former journal. The analysis basically confirmed what has been proposed by Weissberg [Weissberg-90] and Swales [Swales-90] in that the most common schematic components of Introductions are *setting, review, gap, purpose, methodology, results, value and layout*. Some of these components are optional and may depend on the journal style and length of the article. The *layout* component, for instance, generally does not appear in Letters. In the corpus analysed the first four components above had very high frequency. The *gap* component is a strategic one since it points out the weaknesses and limitations of studies discussed in the *review* component, thus paving the way for a clear statement of the paper's intended contribution in the *purpose* component.

[1]The term rhetorical structure is used here for referring to the structuring of a particular text which may (or may not) obey strictly the canonical order of components of the schematic structure of a given genre of text.

The corpus analysis also served to identify 30 rhetorical strategies (based on works by [Maybury-91; Huckin-91; Trimble-85, Weissberg-90; Swales-90]) for organizing 45 different types of information (called here as messages) that make up these strategies. Only papers written by native speakers were considered in order to avoid possible insertion of non-idiomatic linguistic structures into the corpus.

Some of the rhetorical strategies possess common characteristics and could be generalized. But they receive distinct names for helping the user in the gathering of features as they appear in different components. For example, both strategies *Historical review* from the Review component and *Listing the issues to be reported* from the Layout component are chronological strategies for narrative texts. They are temporally organized. The historical review pattern is used to recount a sequence of past events:

Deposition of phthalocyanine derivatives as thin films by the Langmuir- Blodgett technique was ***first*** *described by Roberts et al. who demonstrated that tetra-t-butyl substitution of the ring system facilitated deposition [1].* ***Subsequently****, a number of groups worldwide have derivatised the ring system with a variety of substituents and substitution patterns in efforts to obtain improved LB films containing a high degree of molecular order [2].* ***Recently*** *we described the film forming properties of the amphiphilic phthalocyanine (1) [3, 4].*

(Chesters,M.A et al. Structural evaluation of Langmuir-Blodgett films of amphiphilic phthalocyanines using infrared spectroscopy. Thin Solid Films, 210/211, (1992), pp. 538-541)

The layout pattern, on the other hand, describes step by step the parts of the article using a temporal series of adjuncts:

The rest of this paper is organized as follows. We ***next*** *describe our model of activation over fluctuating barrier and indicate the modes of analysis that are brought to bear its study. Exact and Monte Carlo results are* ***then*** *presented to both illustrate the resonant activation and provide an understanding of the mechanism responsible for this behavior. We* ***conclude*** *with a brief discussion of our results, pointing out some open questions.*

(Doering,C.R. et al. *Resonant Activation over a Fluctuating Barrier*. Physical Review Letters, October, 1992, pp. 2318-21)

In order to further illustrate how the strategies were identified and named, we describe the characteristics of the strategies from the Review Component. Similarly to *Historical review*, *Current Trends* uses a temporal organization but the main rhetorical pattern is one of contrast. The *General to particular ordering for citations* is a descriptive strategy and organizes topics/citations from distant (general area) to close (research topic). That is to say, it is topically organized. The *Citations grouped by approaches*, on the other hand, uses the listing pattern for organizing topics at the same level of closeness. It is similar to the *State of the Art* in grouping facts at the same level of closeness except that the facts reported in the latter are known facts and usually come without citations. Conversely, *Progress in the area* and *Requirements for the progress in the area* are argumentative strategies.

The resulting cases consist therefore of whole introductions, also encompassing information on the associated features and categorization of the strategies. Fig. 1 shows one of the 54 cases.

SETTING: FAMILIARIZING TERMS OR OBJECTS OR PROCESSES

1) Conductive LB films of charge transfer complexes based on TCNQ, TTF and related donors [1], are characterized by strong one-dimensional interactions and are therefore susceptible to stacking defects and the Peierls instability.

REVIEW: REQUIREMENTS FOR THE PROGRESS IN THE AREA

2) It should be possible to overcome such defects and thus produce films of higher conductivity if a complex of higher dimensionality is chosen.

3) Indeed, Nakamura et al. [2] have shown that monovalent salts of a range of long-chain alkylammonium salts of M(dmit)2, where M = Au, Ni, Pd, or Pt may be deposited as LB films and oxidised with bromine or iodine to yield molecular films with conductivities in the range 0.001-30S/cm.

GAP: UNRESOLVED CONFLICT OR PROBLEM AMONG PREVIOUS STUDIES

4) However, owing to their instability at the air-water interface, it was necessary in most cases to deposit the complexes as 1:1 mixtures with icosanoic acid.

PURPOSE: PRESENTING AN EXTENSION OF A PREVIOUS AUTHOR'S WORK + INTRODUCING MORE PURPOSES

5) § In a preliminary report [3], we have shown that it is possible to deposit LB films of the pure didodecyldimethylammonium salts of the divalent metal complexes [M(dmit)2]2- and [M(mnt)2]2-.

6) In this report we describe further investigations into the behavior of these compounds (complexes 1 and 2, respectively) at the air-water interface and extend our studies to the monovalent salt (complex 3) for comparison.

(Taylor,D.M.et alli. *Monolayer characterization and multilayer deposition of conducting Langmuir-Blodgett films*, Thin Solid Films, 210/211 (1992) pp.287-289)

Fig. 1. One of the Introductions of the case base

It can be seen that some fragments are underlined which correspond to the reutilizable parts of the text. The sentences are numbered because the basic unit of analysis is a sentence that is codified in a phrasal lexicon[2]. Information specific of the authors' work comes in normal text. The corresponding representation in Prolog is illustrated in Fig. 2.

```
case( tsf4,
    [c(setting,s(familiarizing_terms_or_objects_or_processes,
                        [m(semi_formal_definition, tsf4,1)])),
     c(review,s(requirements_for_the_progress_in_the_area,
                        [m(topic_complexity, tsf4,1),
                         m(requirements_progress,tsf4,2),
                         m(evidence,tsf4,3)])),
     c(gap,s(unresolved_conflict_or_problem_among_previous_studies,
                  [m(conflict_or_problem,tsf4,4), m(cause,tsf4,4)])),
     c(purpose,s(presenting_an_extension_of_a_previous_authors_work,
                   [m(author_prominent,tsf4,5), m(extension,tsf4,6)])),
     c(purpose,s(introducing_more_purposes, [m(addition,tsf4,6)]))], _ ).
```

Fig. 2. Rhetorical features represented in Prolog

[2] The phrasal lexicon indexes several sentences and clauses, as well.

A case is represented by the Prolog structure case(Case_Name, ListofComponents, ListofPragmaticFeatures). Each of the components in the ListofComponents is represented by c(Component_Name, Strategy). For each component only one strategy is chosen each time. The strategies have the Prolog structure s(Strategy_Name, ListofMessages) which should include a few messages (typically 3 or 4) taken from the 45 different types. Some of the messages are specific to the genre of scientific article, such as *extension, topic_complexity*, *requirements_progress* and *conflict_or_problem*. Some are rhetorical relations which appear in other genres as well (such as *motivation*, *cause*, *addition*, *evidence)*. Messages may also be recognized by the conventional order of their semantic components. In the latter class one may include *formal* or *semi_formal_definition*, and *author_prominent*. The Prolog structure of the messages is m(Message_Name, Case_Name, Sentence_Number). The messages have pointers to the phrasal lexicon entries. As already mentioned, our current implementation includes 45 message types, some of which are stand-alone messages. Other messages can be combined to form compound messages and there are also those that can be written in both stand-alone and compound fashions.

The last component of the structure of a given case, ListofPragmaticFeatures, is not being used in developing the present tool because the articles in the corpus possess several characteristics in common — such as the length of the Introduction, style of the journal, and to some extent, the intended audience. This component would be very important, however, when the approach is applied to other sub-genres or broader classes of scientific work. An Introduction of a review article or of a book chapter, for instance, is much longer, it has another purpose, being directed to another target reader, and therefore it possesses different components and/or arrangements.

3 The Writing Tool

The tool is being implemented in a Sun SparcStation using Sicstus Prolog with the interface being built with the XView toolkit and the Netscape browser. It is integrated into the AMADEUS environment which already offers a reference writing tool for more experienced writers [Fontana-93]. To use the tool in writing an Introduction the user follows a three-step procedure: i) gathering of features, in which the user selects from several menus the features intended for his/her introduction; ii) selection of the best-match case, following the case recovery by the system; iii) revision on the selected case. These stages are shown in the architecture of the writing tool (Fig. 3).

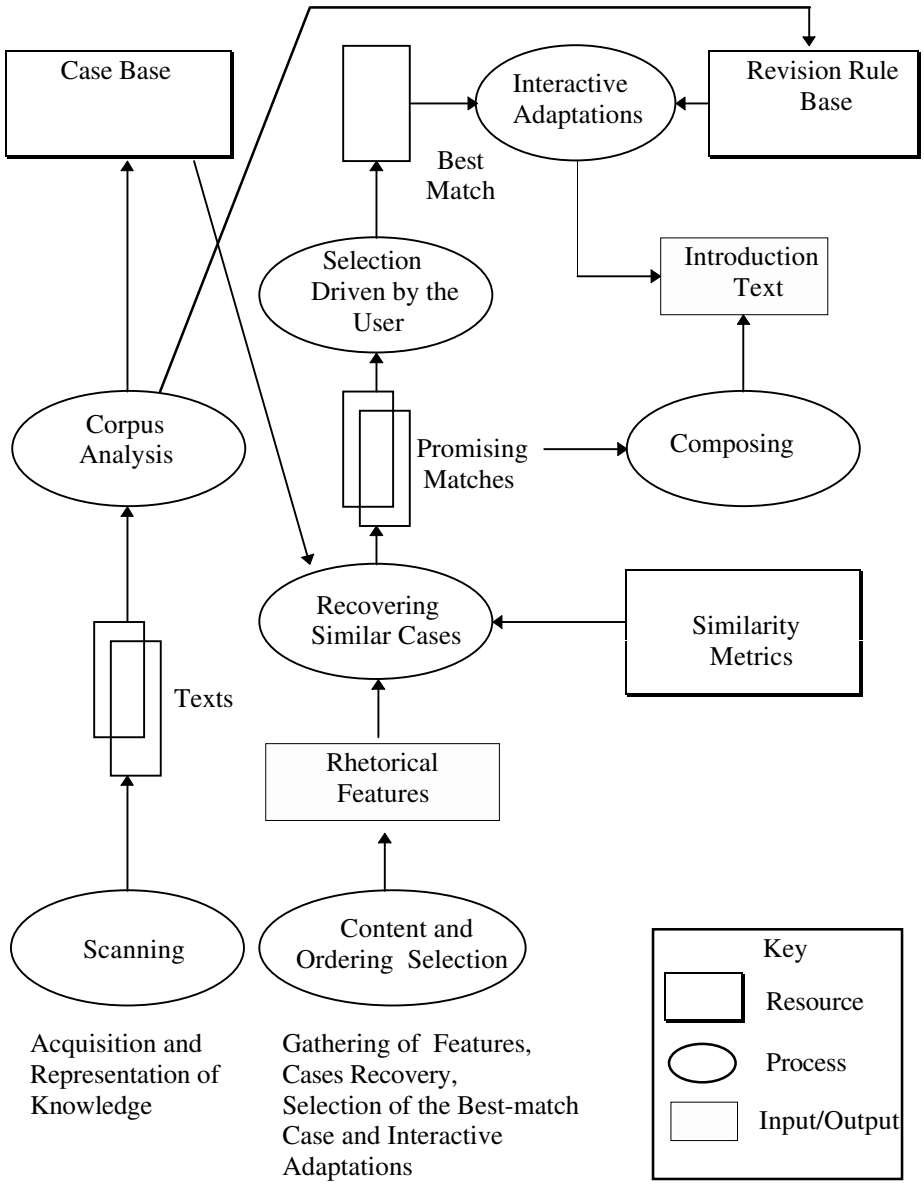

Fig. 3. The architecture of the case-based writing tool

3.1 Gathering of Features

In this stage the user selects in a menu system the components and strategies (and their order of appearance) for his/her introduction, forming the so-called requisition. The user also provides the system with information on the degree of certainty of the chosen order of components. There are 3 degrees of certainty which affect the choice of the metrics of similarity to be employed in recovering the cases: *sure about the chosen order*, *some doubts*, *many doubts*. Three distinct metrics were therefore implemented which only take into account the structural features of the Introduction:

components, strategies, and their order of appearance. Messages are only considered at the Revision step (see below).

3.2 Cases Recovery

Three ways of pattern matching are used: perfect match (*equal lists*), proper undermatch (*sublist*) and non-proper undermatch (*intersection*). The tool selects cases to be returned to the user by employing these three metrics which are basically related to the degree of certainty on the part of the user about the order of the components and strategies (see Fig. 4). The number of returned cases (real introductions) is at most 4. But the system may pick them up from more or less wide sources. The choice "Sure about the chosen order" activates *equal lists* and also recovers cases which contain the requisition (*sublist*). The choice "Some doubts about the chosen order" returns cases using the two previous types of search and other ones that are completely contained in the requisition (*sublist*). Finally, for the choice "Many doubts about the chosen order" the user is returned cases resulting from the previous searches and also from the intersections — requisition partially contained in the case, and vice-versa. Therefore, the search becomes more flexible when the degree of certainty is decreased.

3.3 Revisions: The Interactive Adaptations

Four revision operations were envisaged for this stage of the writing process: i) changes in the lexical and syntactic material of the messages; ii) changes applied to the selected strategies recovering similar ones; iii) addition of messages to a specific strategy; iv) deletion of messages, the opposite operation to iii). Operations i) and ii) are illustrated in Fig. 4. In operation i), there is simply the replacement of material within a given template. In operation ii), however, the changes may also affect the thematic structure of the sentence. In the example given, the type of claim in an argumentation strategy has been changed from *The research is part of a lively area* to *The research is part of a significant area*. As one may see, one goes from a temporally-driven to a topically-driven pattern.

Operation (i)
Initial sentence pattern: In recent years a considerable amount of effort has been devoted to the study of <<ββ decay of various nuclei>>. Lexical and/or grammatical changes: (a) (Recently/In previous years/Currently) a considerable amount of effort has been devoted to the study of ...
(b) In recent years (a great deal of effort/ much attention/ a great deal of interest) has been devoted to the study of ...
Operation (ii)
Initial sentence pattern: In recent years a considerable amount of effort has been devoted to the study of <<ββ decay of various nuclei>>. Changes on the strategy: (a) <<topics>> have been the subject of renewed interest ... (b) <<topics>> have a wide-ranging importance in many areas of ...

Fig. 4. Examples of the revision process

The use of metrics of similarity employed for recovering cases (whole Introductions) can also be used for recovering strategies and ultimately messages from the phrasal lexicon, even if they belong to different cases.

3.4 A Sample Run

Fig. 5 and 6 show screendumps that illustrate how the tool is to be used. Fig. 5 shows the Gathering of Features stage in which the user chooses from a menu system the components and strategies for his/her Introduction. The user must also inform the degree of certainty for the order of components and strategies chosen. These pieces of information will guide the system with regards to the metrics to be employed. Because more than one case can be returned to the user, he/she has the choice of working on the cases returned by transferring linguistic material into his own working area (Composing process). But the user may also select the best returned case and activate the Revision process (in an another window not shown in Fig. 6) and then adapt it before taking material into the working area.

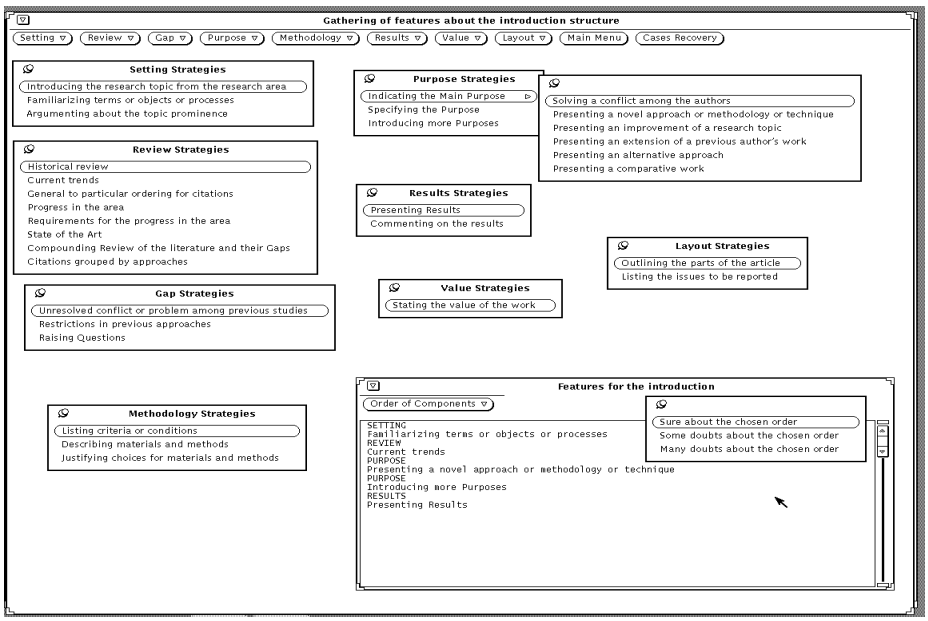

Fig. 5. Gathering of features

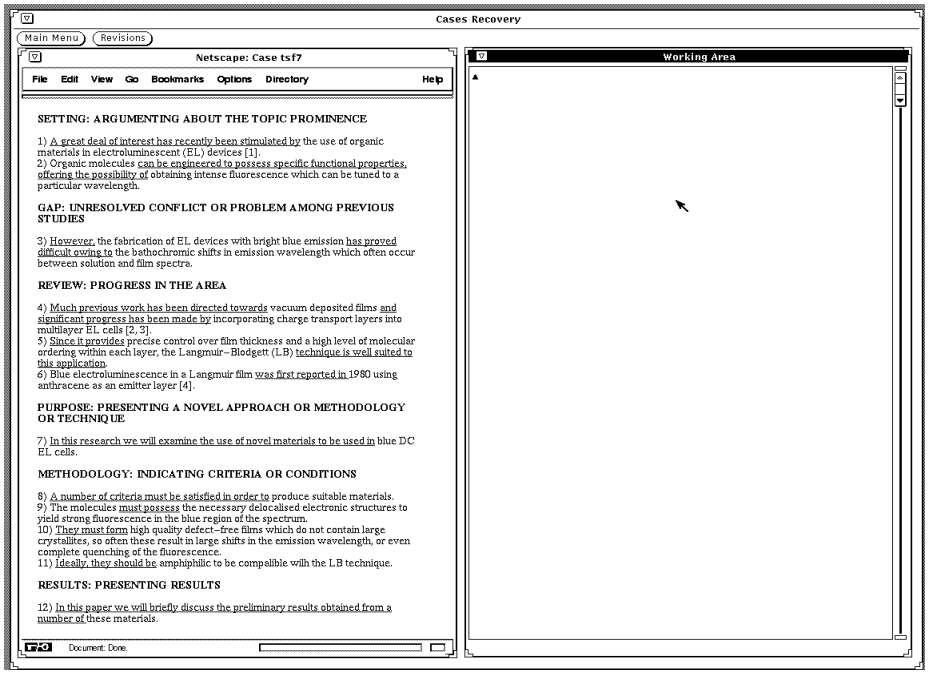

Fig. 6. Selection of the best match case

4. Discussion

The use of a combination of structural, semantic and pragmatic features make humans powerful agents in recovering information from analogous cases (within a domain and between domains) [Kettler-94]. It is therefore expected that the application of the case-based approach in a writing environment may not only help students in their writing tasks, but also teach them the most used strategies in writing a scientific paper. In particular because linguistic material is now offered to the user associated with the various strategies employed in an Introduction, including the cohesive links instantiated with real texts. This will probably obviate the major difficulties faced by less experienced users of the AMADEUS reference tool [Fontana-93].

As for the process of recovering cases, it cannot be quantitatively evaluated at the moment because the case base contains a relatively small number of cases, as it will be discussed later. An extension process for the base is underway. From the three parameters usually employed in such an evaluation - precision, recall and speed of retrieval - only the latter one may be considered now. Indeed, there is no problem of performance as the tool runs in a Sun Sparcstation and retrieval is practically instantaneous. We have nevertheless performed experiments by applying the three metrics to 5 cases in the case base. The results discussed below may be used in the future development of the tool.

The experiments showed in an unequivocal way that the current case base is too small for high recall and precision to be obtained. When the requisition consisted of a relatively long list, often only the identity case was recovered when the *sublist* and *equal list* were employed. The nearest-neighbor matching for these long lists, which provides the score of conceptual similarity, did not reach 50% in any of the cases when the *intersection* metric was employed. One feature that must be considered in recovering the cases is the length of the Introduction. Even though the papers are short (3 or 4 journal pages), their Introductions had lengths varying from 20 to 80 lines. In the longer Introductions some of the components may appear more than once; for instance, one may have a review about a general field and later on reviews on specific topics. Obviously these cases are represented by long lists of components which practically impair the use of the metrics *sublist* and *equal lists*.

The metric *intersection* must be used with caution, as it recovers a larger number of cases, some of which may not be appropriate. It is true that the user recognized that he/she did not know the best order for the components. But perhaps the most efficient way to deal with this problem would be to request the user to decide upon the best order, based on the several suggestions supplied by the system in initial runs. This metric would then only be used in a sort of screening process for the user to make up his/her mind with regards to the order of components.

In terms of implementation costs and portability of the present tool, the major difficulty lies in the need of corpus analysis. While the case-based approach including the metrics suggested here may be portable to other domains, the same is not true of the rhetorical structure of Introductions in other fields. Messages, strategies and even components are likely to depend to some extent on the idiosyncrasies of a particular research area [Crookes-86; Paice-90; Taylor-91]. In order to extend the approach to other Sections of a paper and for other research areas, we are currently developing a semi-automatic tool for compiling linguistic material and indexing it according to the messages and strategies already identified.

Finally, it must be stressed that the tool will only serve as a guide for the user to produce a first draft for his/her Introduction. No attempts to generate text automatically have been done. Nevertheless, even though the system cannot

guarantee the final quality of the text, in particular because the user is to input his/her own material, the approach favours the writing of cohesive, coherent text. The user will also be helped in modifying the text through the revision operations offered by the tool. It is hoped that a case-based writing tool may also have pedagogical implications as the user may learn by analogy to work at all levels of the rhetorical structure of an Introduction: from the components to phrasal lexicon entries.

Acknowledgments

The authors are grateful to FAPESP and CNPq for financial support. They are also grateful to Prof. Stan Matwin for helpful discussions.

References

[Born-92] Born, G. *A Hipertext-Based Support Aid for Writing Software Documentation*. In Computers and Writing - State of the Art, P. O'Brian-Holt & N.Williams (eds), Kluwer Academic Publishers, Dordrecht, pp. 266-277, 1992.

[Buchanan-92] Buchanan, R.A. Textbase Technology: Writing with Reusable Text. In Computers and Writing - State of the Art, P. O'Brian-Holt & N.Williams (eds), Kluwer Academic Publishers, Dordrecht, pp. 254-265, 1992.

[Caldeira-92] (Caldeira), Aluísio S.M.; De Oliveira, M.C.F.; Fontana, N.; Nacamatsu, C.O. and Oliveira Jr., O.N. *Writing tools for non-native users of English*. Proceedings of the XVIII Latinamerican Informatics Conference, Spain, p. 224-231, 1992.

[Crookes-86] Crookes, G. *Towards a Validate Analysis of Scientific Text Structure.* Applied Linguistics, Vol. 7 No. 1, 1986, pp.57-70.

[Fontana-93] Fontana, N.; (Caldeira), Aluísio S.M.; De Oliveira, M.C.F. and Oliveira Jr., O.N. *Computer Assisted Writing--Aplications to English as a Foreign Language*. CALL, Volume 6 (2), p. 145-161, 1993.

[Hovy-90] Hovy, E. *Pragmatics and Natural Language Generation*. Artificial Intelligence 43, p. 153-197, 1990.

[Huckin-91] Huckin, T.N. and Olsen, L.A. *Technical Writing and Professional Communication for Nonnative Speakers of English*. McGraw-Hill, In. 1991.

[Jacobs-85] Jacobs, P.S. *PHRED: A Generator for Natural Language Interfaces*. Computational Linguistics 11(4):219-242.

[Kettler-94] Kettler, B.P.; Hendler, J.A. Andersen, W.A; and Evett M.P. Massively Parallel Support for Case-Based Planning. IEEE Expert, pp. 8-14, February 1994.

[Kitano-90] Kitano, H. *Parallel Incremental Sentence Production for a Model of Simultaneous Interpretation*. In Current Research in Natural Language Generation, Dale, R., Mellish, C. and Zock, M. (eds), Academic Press, Boston, 1990, pp. 321-351.

[Kukich-83] Kukich, K. *Knowledge-Based Report Generation: A Knowledge Engineering Approach to Natural Language Report Generation*. PhD Thesis, University of Pittsburg, 1983.

[Maybury-91] Maybury, M.T. *Planning Multisentential English Text Using Communicative Acts*. (PhD Thesis) Tech. R. 239, University of Cambridge, 1991.

[Oliveira-91] Oliveira, Jr. O.N.; (Caldeira), Aluísio S.M. and Fontana, N. *Chusaurus: A Writing Tool Resource for Non-Native Users of English*, In

Proceedings of the XI International Conference of The Chilean Computer Science Society, pp. 59-70. Also In Computer Science: Research and Applications, R. Baeza-Yates and U. Manber (eds), Plenum Press, N.Y. pp. 63-72, 1992.

[Paice-90] Paice, C.D. *Constructing Literature Abstracts by Computer: Techniques and Prospects*. Information Processing & Management, Vol. 26, No. 1, pp. 171-186, 1990.

[Pautler-94] Pautler, D. *Planning and learning in domains providing little feedback*. AAAI Fall Symposium on Planning & Learning Notes'94.

[Smadja-91] Smadja, F. *Retrieving Collocational Knowledge from Textual Corpora. An application: Language Generation*. PhD Thesis, Computer Science Department, Columbia University, 1991.

[Swales-90] Swales, J. *Genre Analisys - English in academic and research settings.* Cambridge University Press, 1990.

[Taylor-91] Taylor, G. and Tingguang, C. *Linguistic, Cultural and Subcultural Issues in Contrastive Discourse Analysis: Anglo-amarican and Chinese Scientific Texts.* Applied Linguistics, Vol. 12, No. 3, 1991, pp. 319-336.

[Trimble-85] Trimble, L. *English for science and technology: a discourse approach.* Cambridge University Press, 1985.

[Weissberg-90] Weissberg R. and Buker, S. *Writing up Research - Experimental Research Report Writing for Students of English*. Prentice Hall Regents, 1990.