ED 413 753 FL 024 774

AUTHOR	Volz, Norbert
TITLE	CORDON - A Joint Venture Case Study.
PUB DATE	1995-00-00
NOTE	10p.; In: Language Resources for Language Technology: Proceedings of the TELRI (Trans-European Language Resources Infrastructure) European Seminar (1st, Tihany, Hungary, September 15-16, 1995); see FL 024 759.
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Computational Linguistics; *Computer Software; Computer Software Development; Foreign Countries; *Language Patterns; *Language Research; Language Usage; *Languages; *Lexicography; Linguistic Theory; Program Descriptions; Word Frequency
IDENTIFIERS	European Union; *Language Corpora; Neologism

ABSTRACT

CORDON is a computerized system for tracing and documenting changes in the use and frequency of word forms within textual language corpora. In its new form currently under development, CORDON will address the emergence of neologisms, both new words and new terminology within special areas (generally new uses of existing words). The system detects a set of candidates for neologisms within a textual corpus through text annotation, lemmatization, and disambiguation and statistical analysis. The CORDON system is the effort of a consortium of five academic and five industrial partners or subcontractors. The majority of funding for the 2-year software development project will come from the European Union. The resulting product will be a useful and highly efficient tool, but will require regular maintenance to keep it current and close cooperation between academic and industrial partners. (MSE)

CORDON – A Joint Venture Case Study

Norbert Volz

Institut für deutsche Sprache
Abt. Sprachentwicklung in der Gegenwart
Postfach 10 16 21
D-68016 Mannheim, Germany
Tel.: +49 621 1581-437
Fax: +49 621 1581-415
E-Mail: volz@ids-mannheim.de

1. Introduction

Since I first presented CORDON at the Tihany seminar last Summer, some changes had to be made to its structure as the project became more concrete. In order to come up with a realistic, promising proposal, some early concepts had to be dismissed, others had to be revised, and where necessary, new ideas had to be followed. Still the main idea behind CORDON, to facilitate the work of the lexicographer or terminologist by using automatic procedures in order to trace and document changes in use and frequency of word forms within textual corpora, remained the same. In the following sections, I will give a short overview on the underlying concept of neologisms, the market situation and the user needs that called for the CORDON project, its scientific background, and the techniques applied for its translation into action.

2. Neologisms – Birth and re-birth of new words

Somehow, words are like people. They are born, they live, and change, and eventually they die. Language is changing constantly, responding to developments in culture, politics, society, industry, and science. New concepts emerge, requiring words for their expression.

> "What is a new word? This, of course is a question which can never be answered satisfactorily, any more than one can answer the question 'How long is a piece of string?' It is a commonplace to point out that the language is a constantly changing resource, growing in some areas and shrinking in others from day to day." (*ONW* 1991, p. v)

Sometimes, these new words are coined by the inventors of new concepts. More often, however, already existing word forms (single or multiplex) will undergo certain shifts or changes in their usage and meaning, thus enabling the integration of those new concepts in the communication process.

It is those new-born or recycled word forms that are usually called "neologisms" – in short: new, hitherto unrecorded lexical items – regardless of their meaning – that are neither proper names or acronyms, nor typographic errors, nor occasionalisms (cf. Teubert 1996).

A further distinction can be made with regard to the area in which these neologisms are emerging: On a large, common-language scale, we have the appearance of *new words* in general. In industry and science, we witness the emergence of *new terminology* within special areas, for instance: *mouse* in the meaning "a pointing device for computers".* CORDON will deal with both areas as it is designed to cater for general as well as specific applications; therefore, we do not have to go into detail with this issue here.

The previous distinction, however, is more important: The first type is what we have already identified as "new-born" neologisms. These are mostly single words that have been newly created in a particular language, either without a previous model as is the case with most acronyms or so-called "nonce" words (e.g., *quark* in English), or by borrowing words or parts thereof such as roots, affixes, etc. from other languages such as Greek, Latin or nowadays English. To detect this type of neologism, one can countercheck the textual data against a list of already known word forms that is regularly being updated and look for suitable candidates for new word forms - a labourous and tiresome, yet fairly straightforward, procedure in itself.

The second type are the "recycled" word forms, i.e., already existing lexical items – either single words or multiple forms – whose usage and meaning have been shifted, changed, or extended. Obviously, the detection of these word forms proves even more difficult as with the first type. Just searching a full-form lexicon will not detect those lexical items that have maintained their spelling whilst changing their meaning. To detect the candidates of this type of neologism, a statistic approach, based on time-structure, frequency, and context analysis is needed.

3. The CORDON Modules

CORDON combines these two complementary approaches within a modular, step-by-step solution whose parameters can be fine-tuned to different application environments. It is centered around three main steps:
- Corpus Annotation based on full-form lexicons
- Context-based Detection Module
- Time-structure-based Detection Module

The goal of CORDON is to detect a set of "candidates" for neologisms within a textual corpus, either general language or terminology. The corpora

* However, it is clear that nowadays *mouse* in its specialised meaning is no longer to be regarded as a neologism in the strict sense, as the word became part of the everyday language of computer users. This aptly illustrates the aforementioned "life-cycle" of words.

used by CORDON will be monitor corpora according to John Sinclair's typology (Ball/Sinclair 1995), that is, collections of texts that can be acquired cheaply and without greater effort in relation to their size, e.g., machine-readable texts readily available on CD-ROM, typesetting tapes, or other electronic media. Periodical texts, for instance, newspapers or magazines are particularly useful for that purpose.* They will be linguistically annotated, i.e., lemmatised and POS-tagged, to reduce ambiguities, as this will further increase the percentage of genuine candidates for neologisms (cf. Roche 1993).

At the end of the annotation process, we now have a concordance list of unrecognised words, containing, besides proper names, typos, or special forms, the potential candidates for neologisms. To minimise "noise", i.e., remove those proper names, typographic errors, and other quasi-linguistic material and to maximise the output of genuine neologisms, a complete lexicon is to be regarded as crucial to the project.

In order to browse the texts accumulated within a month from an average daily newspaper, a lexicographer would have to examine about 40 000 word forms, of which only 2% prove to be genuine neologisms (cf. Maier-Meyer 1995, p.196). Therefore, it is necessary that proper names and alphanumeric sequences are recognised as such and not as potential candidates for neologisms. A complete lexicon will allow these sequences to be filtered, producing less noise, thus optimising the input for the subsequent statistic procedures. The lexicons used by CORDON, therefore, have incorporated lists of proper names and abbreviations. Furthermore, they include segmentation algorithms that separate alphanumeric sequences into the numeral part and the alphabetic part. The number of unknown words is considerably reduced.

For the context analysis, CORDON uses the statistics-based context analyser ENV-LINE developed at the School of English, University of Birmingham under the framework of the MECOLB project. ENV-LINE assesses the statistic significance of co-occurring words in a defined series of words and evaluates the significance of the entire series. It disambiguates words through the context and detects characteristic "example series". ENV-LINE tries to locate combinations of words that show a high degree of attraction within a file of concordance lines, tracking them down to collocates which

* At the Institut für deutsche Sprache, monitor corpora have been recently included to the corpora available within COSMAS, the IDS' corpus storage and retrieval system. There is now access to a so-called "text pool" of a regional newspaper, the *Mannheimer Morgen*. This pool consists of all newspaper articles considered for publication within the next few days and is constantly updated.

are attracted by a certain node-word. The underlying assumption being that words in a text attract each other to a certain degree, characteristic occurrences for a certain word, therefore, may be found. The program was originally conceived as a tool for lexicographers to find suitable examples of words in use; but when operating ENV-LINE, it was found that the result of the processing may also be helpful for the disambiguation of different senses of a word. The crucial point in assessing the output is the clustering of significant collocations. A word with several different meanings will show several corresponding groups of collocates; hence, a change or shift in meaning can be tracked by the software.

The core of the Time-structure-based Detection Module is a statistical analysis of the distribution of textual features within time-interval segments (Belica 1996). The corpus material used has to be suited to that purpose: Any text within the corpus has to refer to a certain sampling interval, that is, to one of a certain number of disjunct segmentation phases, according to its date.* The sampling intervals depend on the following three factors: text size, dispersivity, and homogeneity: The smaller the text size, the lesser saturated the samples will be, and the lesser significant irregularities will be detected. Dispersivity means the duration of the particular language within a selected text regardless to its position on the time scale. A newspaper article, for example, is less dispersed than a large novel. The smaller the dispersivity, the smaller the sampling intervals.

The ideal corpus for the above approach would have to be homogeneous, that is, the texts only differ in their position on the time scale, whereas all other parameters such as text type, style, genre, subject, author profile, etc., are identical. In reality, however, those ideal corpora do not exist. Therefore, the flaws caused by internal inhomogenity have to be eliminated by adjusting the corpus composition.

Basically, the statistic analysis can be divided into 10 steps:

Step 1: Definition of sampling intervals (e.g., daily, monthly, annually, or based on individual events such as elections, etc.)

Step 2: Definition of the significance factor

Step 3: Selection of the statistic features (e.g., frequency of word forms, average word length, frequency of annotates, etc.)

* The specifications for the sampling intervals have to be set according to external, problem-based criteria; for instance, the granularity, i.e. the duration and frequency of the sampling intervals, is subject to the linguistic phenomena to be analysed.

Step 4: Postulation of the zero-hypothesis: “The selected statistic feature shows a normal distribution within N intervals“.
Step 5: Calculation of the statistical distribution over sampling intervals
Step 6: Validation of the zero-hypothesis for all samples within the intervals (Walsh-Test, χ^2-Test, Fisher-Test)
Step 7: Selection of the “irregular” or “abnormal” samples
Step 8: Quantification of the distribution (Kendall-Coefficient, Difference-Coefficient, Concordance-Coefficient)
Step 9: Analysis of the distribution (Context analysis, pattern recognition, clustering)
Step 10: Linguistic interpretation

At the conclusion of the above procedures, possible “candidates” for neologisms have been identified that can be further processed, for example, stored in a continuously updated database or translation memory, as is the case with the CORDON Demonstrator application.

The front-end used for CORDON will be based on the user interface designed for the machine-based translation system STA that has been developed by one of the consortium partners, THAMUS Spa. As this software product is based on Microsoft Windows®, large portability and affordability of the final CORDON product is guaranteed.

4. The CORDON Consortium

The CORDON consortium consists of five academic and five industrial partners or subcontractors. The academic partners are:

BIR	University of Birmingham, School of English, Birmingham, U.K.
CIS	Centrum für Informations- und Sprachverarbeitung der Universität München, Munich, Germany (Subcontractor)
GOT	University of Gothenburg, Department of Swedish, Gothenburg, Sweden
IDS	Institut für deutsche Sprache, Mannheim, Germany
PAR	Laboratoire de Linguistique Informatique Université Paris XIII, Paris, France

The industrial partners are:

CBD	COBUILD Ltd., Birmingham, U.K.
GCP	GECAP Gesellschaft für technische Information mbH, Munich, Germany

KFT	Krupp Fördertechnik GmbH, Duisburg, Germany (Subcontractor)
MDD	La Maison du Dictionnaire, Paris, France
THA	THAMUS, Consorzio per la Linguistica Computazionale, Salerno, Italy (Coordinator)

The project duration will be two years with an assigned manpower of 184 person-months. Estimated project costs are 2061 kECU of which 1486 kECU are expected to be EC-funded.

THA will be responsible for the coordination of the project. As a company with proven expertise in the area of language engineering, they will ensure a smooth and efficient overall organisation of the project under an experienced and professional management. They will organise the semiannual meetings and will also be responsible for information flow and decision enforcement among the partners and accounting of the project.

KFT with their large experience in LR applications on terminology and translation will be responsible for the specification of validation criteria in order to ensure maximum response to user needs and to minimise the marketing risks for the final product. For the tools and resources available within the CORDON consortium, the main evaluation criteria are their generic applicability as modules for the tasks encompassed by the project, their conformance to already existing standards, and their response to user needs as denoted by the industrial user community. For the CORDON demonstrator itself, the validation process will help to calibrate parameters, specifications and algorithms according to the characteristic features of a limited number of applications, i.e., lexicon development, translation, and terminology management.

Assessment and evaluation are undertaken at defined stages ("evaluation breakpoints") within the overall project (months 12 and 18), that is, after completion of the corpus and functional specifications. In addition to these evaluation stages, a "dummy version" of the CORDON demonstrator is made available already at a very early stage of the project to allow for effective verification and quick response to user needs, especially in the areas of ergonomy and front-end performance. This "dummy version" will successively be upgraded or replaced by the inclusion of more modules.

From the very beginning of the project, one of the industrial partners, CBD, will be responsible for the establishment of an Industrial Interest Group (I.I.G.) which is to be involved in every stage of the project in order to ensure verification and alignment to the needs of the user community. The I.I.G. will represent the typical prospective users of CORDON. It will include end users such as translators, producers of technical or scientific texts, etc., as well as processing users such as terminology providers, and lexicon and dictionary publishers. The I.I.G. will meet on a semiannual basis, allowing on the one hand the users to influence and verify results and performance during the project work while on the other hand enabling future users to make themselves familiar with CORDON while the project is still running. This has two main advantages: It will speed up the start of usage of the final software product and will also reduce the associated marketing risks once the project is finished. The final CORDON demonstrator, therefore, will be the result of a continuous process, involving potential users at every stage.

5. The Market Situation

During the last decades, the progress within the area of information technology in general and language engineering in particular has shown a considerable impact on the acquisition, storage and maintenance of LR, and especially of corpus-derived lexical resources. Vast amounts of linguistic data material can now be accumulated at a fraction of the price of several years ago; and today, even small desktop PCs allow the access to CD-ROM databases many megabytes in size. As a result, today many more organisations, industrial enterprises, academic institutes, and private persons than ever before work with machine-readable textual data.

Machine-readable corpora can be described as "raw resources". They are used for the production of lexicons, which in turn form the central components for various other language applications such as translation memories, terminology databases, machine-readable dictionaries, etc.

Those raw resources are also capital investments. Unfortunately, textual resources become outdated rather quickly with natural language being a highly unstable communication system subject to changes affecting society, science, and technology. Therefore, excellent material is of utmost importance, as competition on the LR market is hard, and especially small and medium-sized enterprises (SMEs) that build the majority of Europe's LR and LT industry must ensure a product of high degree quality. Maintaining and enforcing the actuality of the resources used is vital for both the supplier and the end user of lingware products: It is important for the supplier as it will maintain the market value of the product. In other words: Old, out-dated applications and resources will not sell. It is important also for the user: Only actual, up-to-date material will guarantee a high degree of quality and competitiveness of the end product or service.

6. Conclusion

In this project, the European Community is expected to cover the major part of the costs. This is often the case with projects that are concerned with the development of innovative technologies. However, the participation and financial support from industrial partners is also very important and will become increasingly important in the future. Still, commercial enterprises will only invest in areas where they can expect to gain profit.

Therefore, a close cooperation between industrial and the academic partners is vital. The aim has to be the development of competitive products for the European and global LR and LT market. The academia will provide the resources and expertise needed, including generic software. The industrial partners will ensure that the final product meets the demands of the user community. CORDON shows that the above cooperation is possible and can be used to provide sophisticated generic tools that will facilitate communication among Europe's various languages.

So far, Europe still has a leading edge in LR and LT applications. Only cooperation between public domain research and industry will maintain and strengthen this leading position on the international market.

7. References

Ball, J. and J.M. Sinclair. 1995. "Corpus Typology". EAGLES working paper. Electronic document on the University of Birmingham FTP server. University of Birmingham.

Belica, C. 1996. "Analysis of Temporal Changes in Corpora". In: International Journal of Corpus Linguistics, Vol. 1,1 (forthcoming).

Guenthner, F. and P. Maier. 1994. Das CISLEX-Wörterbuchsystem. München: Centrum für Informations- und Sprachverarbeitung der Universität München (CIS-Bericht 94–76).

Maier-Meyer P. 1995. Lexikon und automatische Lemmatisierung. München: Centrum für Informations- und Sprachverarbeitung der Universität München, (CIS-Bericht 95–84).

Tulloch, S. (ed.). 1995. The Oxford Dictionary of New Words. Compiled by Sara Tulloch. Reprinted with corrections. Oxford, New York: Oxford University Press.

Roche, E. 1993. "Analyse syntaxique transformationelle du français par transducteurs et lexique-grammaire". These de doctorat en informatique, Université Paris 7, Paris.

Teubert, W. 1995. "Neologie und Korpus". In: Korpus und Neologie. (To be published)

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: TELRI - Proceedings of the First European Seminar:"Language Resources for Language Technology", Tihany, Hungary, Sept. 15 and 16, 1995

Author(s): Heike Rettig (Ed.)

Corporate Source:

Publication Date: 1996

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

☒

Check here
For Level 1 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) *and* paper copy.

The sample sticker shown below will be affixed to all **Level 1** documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 1

The sample sticker shown below will be affixed to all **Level 2** documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2

□

Check here
For Level 2 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but *not* in paper copy.

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at **Level 1**.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Sign here→ please

Signature:

Printed Name/Position/Title: Norbert Volz, M.A.
TELRI Project Manager

Organization/Address:
Institut für deutsche Sprache
R 5, 6-13 - 68161 Mannheim
Postfach 101621 - 68016 Mannheim

Telephone: +49 621 1581-437

FAX: +49 621 1581-4156

E-Mail Address: volz(at)ids-mannheim.de

Date: 28/11/97

(over)