ED 413 743 FL 024 764

AUTHOR Gellerstam, Martin
TITLE Lexical Resources and Their Application.
PUB DATE 1995-00-00
NOTE 9p.; 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; Descriptive Linguistics; *Dictionaries; Foreign Countries; Grammar; Information Sources; Information Utilization; *Language Research; *Languages; *Lexicography; Linguistic Theory; Pragmatics
IDENTIFIERS Trans European Language Resources Infrastructure

ABSTRACT

This paper discusses computer-based resources for lexical data and their uses. First, the kinds of lexical data available are described, including those related to form (spelling, pronunciation, inflection, word class), meaning (definition/equivalent, synonyms/antonyms/hyperonyms, thesaurus classification), context (grammatical collocations, lexical collocations, idioms, valency), and pragmatics (distribution, frequency). Different forms in which lexical data are collected are examined, including: word frequency lists; printed dictionaries in machine-readable form, with and without linguistic codes and classification; machine lexicons; lexical databases; and computational linguistics lexicons. The paper then notes the resources available or in production through the Trans-European Language Resources Infrastructure. Finally, areas in which lexical resources can be used are sketched, including lexical research itself, field-specific research, lexicography, creation of writing aids (spell and style checkers), computer aided translation, language instruction, and information retrieval and artificial intelligence. (MSE)

Lexical Resources and Their Application

Martin Gellerstam

Göteborg University
Department of Swedish
S-412 98 Göteborg
Tel.: +46 31 77 34 544
Fax:. +46 31 77 34 455
E-mail: gellerstam@svenska.gu.se

1. Lexical data as linguistic resources

Lexical data are valuable resources in a knowledge society. An American computer linguist, Martin Kay, in a paper about "The Dictionary of the Future and the Future of the Dictionary" (Kay 1984), compares big dictionaries with "Rolls Royce cars" and "country estates" . The metaphor may be a bit surprising, but it certainly takes a substantial amount of effort and skill to produce a good dictionary, and the ongoing discussions in computational linguistics about "reuse" of lexical data is a reflection of this fact.

Dictionary data have not always been considered as "resources" in the way this word has been commonly used up to now. The typical context for the word is "natural resources", things like water, timber, ore, etc. A widening of the concept to the field of linguistics came with computers and corpus linguistics. If you consult a printed dictionary to see how a word is spellt, you probably do not think of your dictionary as a resource; however if you use an automatic programme that spots all your misprints you start looking at lexical data in a new way. Thus, a spelling correction programme is just a start. We know that the dictionary has a greater potential than that. Let me just quote a leading lexicographer (Swanepoel 1994):

> The computer systems and tools that are becoming available both to the researcher, the practical lexicographer and the human user are opening up a myriad of possibilities for the presentation and utilization of masses of lexical information.

Therefore, it is not just a question of computer people handing over practical tools to lexicographers. Without lexical data, collected and systematized by skilled lexicographers, there would not be much to "reuse" by computational linguists. And the lexicon is a sine qua non for computational applications:

> The lexicon can be conceived as the point of conjunction of the different types of information to which any NLP system must have access: morphological, syntactic, semantic, pragmatic. (Calzolari 1989)

According to EU terminology, linguistic resources are divided into corpus resources, lexical resources, and tools. The borderline is not very distinct. It is a question of perpective if your resource is in horizontal order (textual data) or in vertical order (lexical data), and the same tool can be based on a tagged text or a lexicon. Furthermore, texual data form an important part of a dictionary (as sentence examples, etc). In fact, a corpus could be seen as an extension of the textual data actually used in a dictionary. In a CD-ROM version of a dictionary, the user could very well find a link to all the examples he or she could possibly need. This method is used already by dictionary publishers. Thus, corpora are the stuff that dictionaries are made of, and dictionaries can refer to extended text reservoirs in the form of corpora.

So, what is the implication of the concept "lexical resources"? I would like to define it as "lexical data, preferably in machine-readable form, that can be used in lexical research and/or form the basis of commercial products". And "commercial products" could mean almost anything connected with words that is worth putting money into: dictionaries (computer-based or not), computer games, automatic hyphenation, spellchecking, writing aid, computer-aided learning, computer-aided translation, etc.)

2. Types of lexical data

As a starting point for our discussion of lexical resources, we will take a quick look at the lexical data involved. The following lexical facts – well-known to all lexicographers – reflect different aspects of lexical information necessary to describe the usage of a word. Owing to the tyranny of alphabetic order, dictionary publishers often take the opportunity to portion out the different types of lexical information into smaller dictionaries, specialized in one or two lexical aspects: dictionaries of spelling, pronunciation, definition, synonyms, idioms, etc. It remains to be seen if dictionaries stored in computer form will change this publishing tradition.

Words in a dictionary could be described according to the following principal aspects:

FORM	spelling
	pronunciation
	inflexion
	word class
MEANING	definition/equivalent
	synonyms (antonyms, hyperonyms, etc)
	thesaurus classification
CONTEXT	grammatical collocations
	lexical collocations
	idioms
	valency
PRAGMATICS	distribution (domain, register, style)
	frequency

If you look at these lexical categories – most of which have a published dictionary counterpart – you will find that they are more or less suited for computational application. The bestsellers today are computer programmes that check your spelling or your grammar and style and give you information about synonyms. And for these programmes to be able to handle your text, they have to have access to inflexion and word class. Other categories (like valency and collocations) will be of great importance in automatic text analysis. Frequency could also be used in this context.

3. Carriers of lexical data

Lexical data will reach the language user in a variety of machine-readable dictionaries and computer programmes, ranging from simple spelling-checking devices to sophisticated products of computational technology. The following categories reflect a hierarchy from simple carriers of lexical data to more complex ones. The hierachy also reflects the degree of explicitness: from dictionaries where the human reader can pick the details for his or her understanding of a text to formalized lexicons where the information must be explicit.

1. Word frequency lists (word form -> lemma)
2. Printed dictionaries in machine-readable form
3. Printed dictionaries in machine-readable form with linguistic codes and classification; also recently published CD-ROM versions
4. Machine lexicons, classified, encoded, and with selected information (often designed for automatic lemmatization)
5. Lexical data bases
6. Computational linguistics lexicons

Word frequency lists were produced as a result of many frequency counts in the sixties and seventies. One example is the American Brown Corpus (Kučera & Francis 1967) which has had a marked influence on later corpus investigations, especially as a model for corpus collection.

Printed dictionaries in machine-readable form can also be dated back to the sixties when the Merriam-Webster Seventh New Collegiate Dictionary was put into machine-readable form by J. Olney and his colleagues.

Printed dictionaries with explicit linguistic codes and classification have been published since the first edition of the Longman Dictionary (LLDOCE 1981). Later editions of both this dictionary and Collins COBUILD dictionary have lately appeared also in CD-ROM form.

Thesaurus dictionaries date back to Roget's Thesaurus from the 19th century. Later efforts in the same direction have been published as pedagogical variants of regular dictionaries, from Longman's Lexicon of Contemporary English (1981) to Longman's Language Activator (1993).

Machine lexicons are not designed to be read by humans but provide explicit lexical information for performing specific tasks, e.g., automatic lemmatization. The words of a text are confronted with a list of forms listed in the dictionary. If a certain form is found in the text, the word is associated to the correct lemma (including part-of-speech).

Lexical data bases (LDBs) contain formalized information at many descriptive levels. It is one of the chief tools today for processing great quantities of lexical information. It can be used for various types of linguistic applications and for general research in the lexical field. A data base management system provides the user with tools which enable him to access the data without necessarily being familiar with the internal or physical organisation, but only with the type of information he can retrieve.

Computational linguistic lexicons are more complex tools for parsing, for artificial intelligence (question-answering) and for Machine Translation.

3. TELRI resources

A quick look at the TELRI Language Resources gives an impressive overview of lexical data from a wide range of European languages. Efforts to bring all of this data together and make it accessible – or at least make a catalogue accessible – should be a priority for TELRI. We certainly need a European counterpart to the Linguistic Data Consortium in the United States.

It is difficult to avoid a certain disagreement about lexical terminology. When you find a "database with morphological information" among the TELRI lexical data you do not know if you are confronted with a well-structured database containing various kinds of information about inflection, derivation, etc. or perhaps a machine lexicon for lemmatization. To be able to compare the different types of information among the TELRI countries, you should not really compare a database with inflection categories with a lemmatizer. To have a database does not imply that you have all relevant lexical information: it just means that you have stored the lexical facts in a certain form with multiple access. If you have a "phraseological dictionary", how does this relate to things like "collocation tools" (which is not a lexical resource), etc.?

After these reservations, the lexical resources situation could be summed up in the following way:

- There are quite a number of MRDs but few fullfledged databases with a variety of lexical information
- The majority of TELRI members have tackled the crucial question of lemmatization which often – but not always – correlates with the existence of spelling-checkers
- Semantic information is scarce: there are just a few dictionaries of synonyms and semantic tagging (e.g., tagging of definitions), and thesauruses are rare.

4. Areas of Application

To ask for the application of lexical resources is like asking for linguistic application in general. If the lexicon – at least in principle – is "the point of conjunction of the different types of information to which any NLP system must have access" (see above), you will find it difficult to say what is not an application of lexical resources. In this context, however, I will just point out a few obvious fields of application.

The first area that should be mentioned is the field of lexical research itself, which is not only producing applications but which is also an activity putting existing applications to the test. Such applications range from simple concordances and tagged texts to lexical databases and alignment methods (notice the fuzzy borderline between "lexical data", "textual data", and "tools").

Lexical resources are also used in various other fields of research, e.g., *psychology*, where lexical data is needed in fields like language learning, testing of patients with brain diseases, etc. In sociology, vocabulary data are used to reflect cultural and ideologic development in society.

Lexicographic practice is a field where lexical applications are put to continuous test. Applications cover the whole dictionary production line from corpus collection over the harbouring of lexical facts in a database with multiple access to the final production of a dictionary article. However, the one outstanding lexical application is the final dictionary itself – on paper, diskettes, CD-ROM, etc.

Even if the dictionary itself is the image of lexical resources, writing aids of different kinds are the most typical computational commercial product. To begin from basics, there are general text checkers that check practical things like starting a new sentence with a capital letter, spotting extra spaces between words, etc. *Spelling checkers* are usually based on a collection of wordforms representing an actual corpus or a list of wordforms generated from a dictionary. Text verification to find spelling errors and for automatic hyphenation is probably the number one commercial application. Spelling checking is a relatively easy task for a language like English with little morphology but becames a more complex task in a language with rich morphology. Spelling checking facilities are more or less standard ingredients in word processing today. This is also true of synonymy information (sometimes advertized to give you an impression that languages contain hundreds of thousands of synonyms). *Style checkers* have developed from the first simple spelling correction systems. Modern style checkers include checking of particular words from stylistic point of view ("why do you use the passive form?"), parsers for spotting grammatical errors (like congruence), and checking of contextual data ("have you used the right preposition after the verb?").

A more specialized type of writing aid is *computer-aided translation (CAT)*, where a computer programme looks up words and phrases and suggests translations to the human translator.

Language learning applications based on lexical data can be used in various types of interactive teaching of written language skills. Computer programmes can assist in tasks like sentence restructuring, checking of translation and dictation tasks, cloze testing (filling in omitted words in a text), and dictionary look-up.

Other fields of knowledge where lexical data can be applied is *information retrieval* (this is where the thesaurus comes in), various kinds of applications in *artificial intelligence* such as question-answering. The need for a comprehensive lexicon for *machine translation (MT)* is widely acknowledged today.

5. Final remarks

In this brave new world of possibilities and application of lexical resources, it may be necessary to add a few words of warning concerning the general handling of lexical data. Today, great efforts are put into *standardization of texts* and lexical data in the framework of the Text Encoding Initiative. The idea is of course that a home-made formalism for your lexical data is less than practical in a society where this sector is growing steadily and exchange of data is becoming more and more frequent. On the other hand, standardization is a laborious and expensive task – especially if we are talking about standardization of already existing corpora – and you may not quickly reap the full benefit of your effort. You must make a rational weighing-up of pros and cons.

Another problem is the *dissemination* of your lexical application. Be sure about the legal and economic implications before you offer your data free of charge or for a sum of money, for scholarly or commercial use. How many public domain dictionaries have you come across when surfing on Internet? Also, do not forget to protect your *copyright* in dealing with commercial partners.

References

Calzolari, Nicoletta. 1989. "Computer-Aided Lexicography: Dictionaries and Word Data Bases". In: Computational Linguistics. An International Handbook on Computer Oriented Language Research and Applications. Berlin & New York: Walter de Gruyter.

Collins COBUILD English Language Dictionary. 1987. London and Glasgow: Collins.

Kay, Martin. 1983. "The Dictionary of the Future and the Future of the Dictionary". In: Linguistica computazionale, 3 (1983).

Kučera, Henry & Francis, W. Nelson. 1967. Computational Analysis of Present-Day American English. Providence, Rhode Island: Brown University Press

Longman Dictionary of Contemporary English (LLDOCE). 1981. Harlowe and London: Longman Group Limited.

Longman's Lexicon of Contemporary English. 1981. Harlowe: Longman Group Limited

Longman Language Activator. 1993. Harlowe: Longman Group UK Limited

Swanepoel, Piet. 1994. "Problems, Theories and Methodologies in Current Lexicographic Semantic Research". In: Willy Martin & Willem Meijs & Margreet Moerland & Elsemiek ten Pas & Piet van Sterkenburg & Piek Vossen (eds), Euralex 1994 Proceedings. Amsterdam 1994, p. 11–26.

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