LEXICOGRAPHICA

International Annual for Lexicography
Revue Internationale de Lexicographie
Internationales Jahrbuch für Lexikographie

Edited by

Antonín Kučera, Alain Rey,
Herbert Ernst Wiegand, Ladislav Zgusta

in conjunction with the
Dictionary Society of North America (DSNA)
and the
European Association for Lexicography (EURALEX)

Editorial Assistant: Werner Wolski

Guest Editor: Laurence Urdang

4/1988

Sonderdruck

Max Niemeyer Verlag
Tübingen

Working Paper Machine Translation System. In: Machine Translation: The State of the Art, M. King (ed.), Edinburgh Press, 1986.

van Sterkenburg, P./W. Martin/B. Al: A New Van Dale Project. In: Goetschalckx, Rolling, (eds.), Lexicography in the Electronic Age, North Holland, Amsterdam, 1982, 221–237.

Teubert, Wolfgang: Applications of a Lexicographical Data Base for German, Proceedings of Coling 84, 34–37, Stanford, 1984.

Thurmair, G.: Linguistic Problems in Multilingual Morphological Decomposition, Proceedings of Coling 84, 174–177, Stanford, 1984.

Vasconcellos, M./M. Leon: SPANAM and ENGSPAN: Machine Translation at the Pan American Health Organization, Computational Linguistics, vol. 11, Nr. 2–3, 122–136, 1985.

Vauquois, B./C. Boitet: Automated Translation at Grenoble University, Computational Linguistics, volume 11, Nr. 1, 28–36, 1985.

Vollnhals, Otto: Technical Dictionaries Retrieved from a Database, META XXVII,2, 1985.

Wehrli, E.: Design and Implementation of a Lexical Data Base, Proceedings of the 2nd Conference of the European Chapter of the ACL, 146–153, Geneva, March 1985.

Weiner, E.: New uses for the new OED, The Bookseller, 3–6, January 1986.

Wheeler, P.: LOGOS, Sprache und Datenverarbeitung, 9. Jahrgang, Heft 1, 11–21, 1985.

Whitelock, P. J./K. J. Kilby: Linguistic and computational techniques in Machine Translation system design, CCL/UMIST report 84/2, University of Manchester, 1984.

Wilks, Y.: The Stanford Machine Translation Project. In: Natural Language Processing, R. Rustin (ed.), 243–290, Algorithmics Press, New York, 1973.

Zimmermann, Harald H.: Nutzbarmachung und Nutzung maschineller Wörterbücher in der Fachinformation und im Büro. In: Sprache und Datenverarbeitung No. 1/2. 1982, 5–10.

Dr. Susan Warwick, University of Geneva, ISSCO 54, rte des Acacias Geneva, Switzerland

Nicoletta Calzolari/Antonio Zampolli

From Monolingual to Bilingual Automated Lexicons: Is There a Continuum?

Abstract

Lexical Data Bases (LDB) are investigated as essential tools in the storing and processing of lexical data for their capability of providing direct and immediate access to sets of words sharing specific properties or features. LDBs are considered as "multifunctional" in many respects: with respect to the data, to the applications, and also to different linguistic theories.

The Italian LDB and its future developments are seen as an example of implementation of a LDB according to the lines described. The aim of the project is to create a large repository of lexical data, where access is provided at the various levels of lexical units, properties, and relations. We are tending towards an automated dictionary which represents lexical information by relations from words to words, or from words to metalinguistic codes, using existing dictionaries as one of the sources for the raw data. One of the purposes of implementing these relations is to transform a particular natural language text, i.e. definitions, into a knowledge base, and to relate natural language words to an underlying and probably interlinguistic taxonomy or network of concepts. A suitable structure makes it possible to obtain wide access to lexical information both in breadth and in depth, and for a number of foreseeable applications. In the perspective of realizing a "neutral" syntactic component, we are investigating the possibility of representing in the lexicon the linguistic information used in parsers and generators in such a way that it can be reutilized in many different theoretical frameworks.

A project is also described which uses bilingual machine readable dictionaries as a "bridge" connecting two otherwise independent monolingual lexical data bases. One of the objectives is to integrate the different types of information traditionally contained in monolingual and bilingual dictionaries, so as to expand the informational content of the single components in the new integrated bilingual system. In the new bilingual system, there would no longer be a source language and a target language, since the look-up and access procedures are independent and neutral with respect to direction. Bidirectional cross-references will also be automatically generated for the information contained at each sense level as semantic indicators (i.e. synonyms, hyperonyms) or contextual indicators.

1. Introduction

Over the past few years automated lexicons have become increasingly attractive fields of research and development. From a theoretical point of view they are at the crossroads of many areas of research, from morphology and syntax and semantics to sociolinguistics and Artificial Intelligence. From an applicative point of view, the possibilities of use range from simple spelling checkers to parsers or knowledge representation systems.

Unfortunately, what has been lacking so far is an effort of harmonization and of convergence either at the level of linguistic structures or at the level of computational models.

We can however assert that a number of basic assumptions concerning monolingual automated lexicons can nowadays be taken for granted, e.g.:

- the necessity to utilize as invaluable sources of data printed dictionaries which are already in machine-readable form, e.g., for photocomposition;
- the necessity to organize dictionary data on a database structure provided with multiple access procedures;
- the necessity to integrate various levels of information, from the phonetic to the morphological, syntactic and semantic level;
- the importance of exploiting the peculiarities of the sublanguage of definitions in order to extract a number of semantic features and semantic relations to be implemented in the database;
- the property of multifunctionalism, i.e., the possibility of utilization, through different modes of access to differently organized virtual subsets, by different procedures for different applications or by human users;
- the property of multi-perspectiveness, i.e., the possibility of presenting the very same data under various perspectives, according to the specific relations between entries considered when querying the database;
- the property of being organized as a complex network of interconnected relational structures arranged in a multi-dimensional way with nodes and labeled links.

While examining the importance of these properties for automated dictionaries, we have to consider which of these notions, concepts, methods, structures and techniques are still valid when dealing with the design of a computerized bilingual dictionary, which ones must instead be changed and which discarded.

In particular we are concerned with the study of a model of a bilingual dictionary which can be a valid tool for the connection of already available large monolingual databases. In a certain sense, the bilingual dictionary should connect two descriptions of the world, i.e., two differently organized knowledge bases; a different formal syntax for the "bilingual entry" and a new explicit internal structure must be envisaged, certainly different from the standard entry of a printed dictionary; new types of manipulations of this structure must be allowed for different search strategies; if we consider each entry as a set of properties and of relations with other entries, the overall external organization of the entries should be conceived in terms of relations among subsets; given the importance of context in determining the exact correspondence between source and target language expressions, an integrated linguistic database (i.e., a lexical database plus a textual database) must be at disposal in those cases in which it is also necessary to have on-line access to large textual corpora through many search strategies.

2. Multifunctional Lexical Data Bases

Why lexical databases? Why multifunctional and multilingual lexical databases? By what methods? For which purposes? By whom? In which environment? For which applications? How have they developed so far? What is the present situation? What of future developments? What are the main features?

These are only a few of the many possible questions that can be posed with regard to electronic dictionaries and we shall try to answer some of them in this paper.

2.1. "Multifunctional" with respect to the data

As complex sets of phonological, morphological, syntactic, and semantic information, words have many properties that interact with each other and that should be accurately captured in a Multifunctional Lexical Database (MfLDB). Words must also be considered at each different level of linguistic analysis, and these levels can be integrated with each other in the many perspectives on the lexical entry offered by a MfLDB.

The first feature which makes LDBs essential tools in the storing and processing of lexical data is their capability of providing direct and immediate access to sets of words sharing a specific property or feature. This is possible because:

a) the lexical data either have been inserted in the LDB in a codified way (where this is feasible, as for grammatical category codes, style labels, field labels, POS, etc.), or have been organized in such a way that a number of morphological, syntactic, and semantic relations are normalized and/or formalized (as for synonymy, hyponymy, word-formation, transitivity, etc.);

b) there are procedures allowing for specified scanning of the data (also in the form of codes or of relations), and for the selection of specified subsets of entries which match given searching criteria. Section 3. describes some of the possible queries implemented in the Italian LDB.

2.2. "Multifunctional" with respect to the applications

The lexicon is obviously an essential component in any Natural Language Processing (NLP) system (for parsing, generating, machine translation (MT), question/answering (QA), information retrieval (IR), lemmatization, artificial intelligence (AI), etc.). The usual practice is to construct an ad-hoc lexical component for each NLP project. It is necessary to move towards large lexicons (both in extension and in depth of representation), where information is represented in such a way that it can be easily interfaced by different application procedures according to the different needs.

This means that the same set of data can be shared by the various applications. Each interface will project on the specific application only that aspect of the data which is relevant for the particular requirements.

This is again possible if the lexical data have been organized as a database system, where access to the units of information can be differentiated along many perspectives, and many different paths can be traced among the recorded data. Each application has thus its own "view" on the common set of basic data. For example, a morphological analyzer will look only at that portion of the lexical entry which contains pieces of information pertinent to morphology, such as POS, paradigm class, and so on.

In this respect, an essential property of MfLDBs is to be easily extensible, i. e., it must be possible for different researchers to add their own idiosyncratic information consistently with the actual content of the LDB. For example, provision should be made for the addition of frequency information both to a headword and to all of its inflections and/or word-senses as they occur in a given corpus.

2.3. "Multifunctional" with respect to different linguistic theories

A large amount of work in computational linguistics (CL) is carried out on experimental lines, with consequently small-sized lexical prototype systems. Furthermore, emphasis is traditionally placed on the representation, organization and use of linguistic knowledge as encapsulated and expressed by linguistic rules and procedures. Lexical data seem to be considered of secondary importance or, at least, to be handled with relative ease.

At a recent workshop held at the Institute of Science and Technology, University of Manchester (UMIST) in 1985, (McNaught, 1986) an informal poll representing a good sample of today's computational linguists was conducted among the invited speakers to establish the average size of the lexicons used by their systems. With the exception of a prototype MT system, the average size was about 25 words. This is probably true of a large number of systems in the realm of computational linguistics, MT systems being the only apparent exceptions.[1]

Today we are forced to consider the following facts:

a) Our CL community has recently been faced with the request for large-scale NLP systems, owing to the recent advances in CL technology which make such applications feasible and to the interest expressed by supranational and national public and private organizations.

b) For "real-world" applications, it is of fundamental importance that a CL system be able to deal with tens of thousands of lexical items. The projects at present underway must be completed within reasonably fixed time limits. The preparation of Natural Language Processing dictionaries can be delayed no longer.

c) Various projects have been proposed for the same language. Up until now, it has been a fact that each system has had its own ideas and conventions with regard to content, organization, and structure of its lexicon. This makes it difficult or even impossible for various NLP systems to share linguistically relevant information for the same language.

d) Duplication of efforts may be a very "sad" fact. Building a comprehensive, consistent NLP dictionary is probably the most costly and time-consuming task in every NLP project. In this situation, it is natural that not only researchers and developers, but also the promoting and financing authorities should put forward the question as to whether it is possible to design a rich, powerful, and flexible LDB, in which different linguistic theories and CL systems can find the relevant lexical information required.

The problem of the feasibility of a neutral LDB and the assessment of even partial solutions are obviously of primary importance for us, as we are only just starting to define the content and the representation of the syntactic and semantic information which we have to implement in the Italian Linguistic Data Base (ILDB).

[1] In general, the MT systems have always been characterized by a strong applicative motivation, which requires a more or less "real-life-size" lexicon.

3. The Italian LDB

In Pisa we have developed a large lexical data base of the Italian language (see CALZOLARI, 1982, 1983a, 1984).

The aim of the project is to create a large repository of lexical data organized in database form, in which lexical units are stored together identifying many kinds of lexical properties and lexical relations and in which access is provided at the various levels of lexical units, properties, and relations. Already available, machine-readable dictionaries or the typesetting tapes prepared for photocomposition have been used as source of the data.

Programs that parse existing machine-readable dictionaries can extract the different types of information by decoding the typesetting codes and can distribute this information to appropriate locations in a model entry scheme.

Furthermore, that part of the information which in the common dictionaries is provided in natural language (i.e., definitions) in our linguistic data base is processed by definition parsing procedures and is transformed either into properties (for inherent features), into attribute/value pairs, or into qualified relations and pointers (for morphological and semantic relations). Thus, the entry is being formalized also at the semantic level.

As a matter of fact, we are now tending towards an automated dictionary which represents lexical information by relations among words, or from words to metalinguistic codes, using existing dictionaries as one of the sources for the raw data.

The lexicon will appear as virtually divided into as many subsets as the relations which have been determined and formalized. The values of some relations will range over restricted sets (e.g., of syntactic categories, usage labels, inflectional codes, etc.), while the values of other relations will range over considerably larger and less determined sets (i.e., the very words of the lexicon differently grouped and accessed according to their different relationships, e.g., by synonymy, antonymy, derivation, thematic role, etc.). By representing the lexicon as the set of all these relations, we can access the dictionary either by lexical items, or by features, or by relations; we can search the network to see where it matches with the query and retrieve different parts of the lexical content on the basis both of the access point and of the options activated at this point.

A suitable structure thus makes it possible to obtain wide access to lexical information both in breadth and in depth, and for a number of foreseeable applications. Even now we can, for example, retrieve on-line from the Italian LDB those lemmas with a given grammatical/syntactic code or only dialectal lemmas, and we can interactively ask for words with a given ending, or for synonyms, hyponyms, and so on along a number of different dimensions.

In a LDB a rather rudimentary form of semantic representation can be reached that allows commands such as:

– Give me a list of all the agent nouns.
– Display the verbs implying the use of an instrument.

- List the verbs of motion.
- Give me the adjectives which can be predicated of sounds.
- List all the derived words implying the notion of 'process' together with their verbal base-words.

These different types of searches demonstrate the fact that different modes of access give rise to lexical activation of differently related groups of entries.

Besides the practical applications in NLP where information of this type is obviously very useful (as for example in Information Retrieval Systems, MT, QA, computer-assisted instruction, etc.), a more theoretical application of the LDB is to use it as a powerful tool for long-term research into the structure of the lexicon itself. We are facilitated in investigations of the following type, e.g.:

- Which are the patterns characterizing the present-day semantic structure of the lexicon?
- Which are the most important semantic and lexical relation types?
- Which relations are given morphological evidence and which are not? and so on.

One of the purposes of implementing these relations is to transform a particular natural language text, i.e., definitions (in a certain sense a sublanguage, with lexico-grammatical restrictions that are very useful to exploit), into a knowledge base and to relate natural language words to an underlying (possibly and probably interlinguistic) taxonomy or network of concepts (typically the one which binds together the defining concepts in dictionary definitions).

4. Perspectives for a neutral syntactic component

As we have said above, we are now starting to define the content and the representation of the syntactic and semantic information which we have to implement in the ILDB.

It is a well-recognized fact that different linguistic theories and different methods of computational organization may have important consequences on the construction of a grammar. Less attention has been paid to the consequences on the lexicon. Although we feel intuitively that lexicons designed for different linguistic theories may contain information which from a certain point of view is identical, as it describes the same linguistic facts, we have to assess the validity of this intuition before starting to implement in the ILDB the information required by the NLP systems. A sound methodology for the evaluation of this intuition may consist in the following steps:

- to review the existing parsers and generators for various languages and, in order to assess the possibility of convergence, to examine the information contained in their lexicons and the way they are represented;
- to conduct a feasibility study on a representative subset of the Italian lexicon to assess the possibility of designing an ILDB which is "neutral" with respect to the major linguistic theories.

Let us briefly consider these two problems.

4.1. Comparison of existing lexicons

On the occasion of the workshop “On Automating the Dictionary”, organized in Grosseto in May 1986 (see D. WALKER et al., 1987), we have requested a comparative study of the lexicons used in computational parsers and generators to B. INGRIA (1987) and to S. CUMMING (1987), respectively.

A preliminary question to be answered is obviously whether and to what extent the directionality of linguistic processing, i. e., analysis or generation, influences the content of the lexicon. Some systems are explicitly planned to be bidirectional, i. e., to use the same lexicon for both analysis and generation; but, in practice, the two types of lexicons tend to be rather different.

> “The generation tasks set different priorities for the lexicon: roughly speaking, a generation lexicon has to put depth before breadth, while the reverse is true for understanding.” (CUMMING 1987)

The following are examples of the differences:

– Parsers must be able to accept a variety of inputs from the user: the grammar must be comprehensive at least with respect to the subset of the language covered; the dictionary must contain a large number of words and support all the syntactic distinctions that the grammar can make.

– A generator does not need a full range of syntactic capabilities (nor does it need a very large lexicon, e. g., one word for everything it needs to say, and fewer syntactic distinctions). Instead, it has to know more about the syntax lexicon: it needs a basis for choosing between syntactic alternatives and lexical items, so as to be not only conceptually appropriate and grammatical but also cooperative, idiomatic, non-redundant, etc.

An analogy can be made with the experience of learning of a second language by a human: typically, the range of appropriate language which the learner can produce is more limited than the range which he can comprehend.

The conclusion at the Grosseto workshop was that parsers and generators may in large part share the same bulk of lexical information. A LDB may easily contain the union of the knowledge requested by both. Some of the information will eventually be used in only one direction. We shall adopt this point of view for the ILDB.

From the point of view of the lexical information used by different parsers and generators, B. INGRIA divided the NLP systems into two general sorts of orientation:

> Syntactically oriented approaches: These systems typically categorize their lexical items in terms of traditional parts of speech and perform detailed syntactic analysis of input sentences or texts.
>
> Semantically oriented approaches: The systems perform fairly idiosyncratic syntactic analysis, devoting most of their efforts to the detailed semantic analysis of their input.

INGRIA decided to consider only the syntactically oriented approaches. While information might be shared, with varying degrees of success between the syntactically oriented systems, there is less likelihood of sharing information: (a) between syntactically and semantically oriented systems, and (b) between different semantically oriented systems.

Furthermore, the information required by semantically oriented systems does not very often relate in a direct or obvious way to any particular linguistic theory or place specific requirements on processing configurations and lexicon content and structure.

SMALL's theory of word-expert parsing, for instance, places a heavy demand on the lexicon (see SMALL, 1981): the lexical entries (word experts) are complicated programs with routine structures, the specification of which requires detailed knowledge of the architecture of the parser, judgement of what constitutes relevant information and how to translate that procedurally, and readiness to bring in an arbitrary amount of more general, common-world, knowledge (see BOGURAEV, 1987).

Obviously, we are very far from a generalization that allows the inclusion of this type of information in a general-purpose and extensive LDB.

Several NLP systems, falling within the large class of 'knowledge-based systems', require a significant amount of structured knowledge about the real world, or at least about a particular domain of discourse. The ways in which knowledge-based systems organize and maintain their knowledge bases differ widely. There is no firm consensus on what kinds of structure are best suited for capturing the knowledge useful for NLP. Nonetheless, it is possible to observe a common theme in a large number of NLP systems.

A part of their knowledge is often represented using a scheme based on the general notions of frame-like concepts with slot-like descriptions, organized into an inheritance hierarchy.

Large, hierarchically structured networks of concepts are certainly a very useful source of data for the construction of the knowledge of these systems, whether they represent semantic knowledge in terms of decomposition into markers taken from a set of primitives, formulae constructed from semantic primitives, frame-based structures, or other information.

We have described above (Section 3.) how we use the definitions in our present machine-readable dictionaries as an aid for the construction of various semantic relations of this kind in the ILDB.

We summarize here the classification schemata of syntactical information suggested by INGRIA (which largely coincides with the schemata of CUMMING), because the next step of our project will probably broadly follow these schemata and will require choices among different possible competing systems of representation.

4.2. Types of information

INGRIA and CUMMING consider the following types of information as generally present in the lexicons of the computational systems revised:

a) Syntactic categories: Most lexicons agree in their assignment of lexical entries to the major categories (N, V, Adj, Adv, Prep), though they may differ as to the exact names of the categories.[2]

[2] For example, *each* is coded as ART, DET, ADJ, QUANT, DETERMINER respectively, in the systems examined by Ingria.

However, the treatment of other categories, and even of the subcategories of the major categories, differs from one lexicon to the other. A very interesting example is represented by the difference in the treatment of quantifiers. The problem, however, is limited, because those categories constitute a "close" subset of the lexicon, and a normalization may easily be reached through manual intervention.

b) Contextual features: This type of information may be defined in terms of the contexts in which a given lexical entry may occur. Following CHOMSKY (1965), they may be devided into two types:

Subcategorization: This specifies the complement structures, e.g., transitive verbs that occur with an object NP.
Selectional restrictions: This specifies the nature of the items that can appear in complementary or in subject position (e.g., transitive verbs that require a direct object to be animate). Some systems regard selectional information as more syntactic in nature, others as more semantic.

c) Inherent features: These cannot easily be reduced to a contextual definition, e.g.: countable/non-countable; abstract, animate, human, etc.
Some are treated as semantic (animate, human), others as syntactic (e.g., non-count).

4.3. Types of representations

The authors also pointed out a set of diversities in the representations adopted by the reviewed systems:

a) Syntactic categories: Simple symbols: each configuration of categories and subcategories is represented by a single code (e.g., the KUNO system (1965) has 133 different syntactic codes).
Complex symbols: each category and subcategory is represented by an independent code. Each lexical entry is cross-classified with respect to an array of categories and subcategories.

b) Contextual features:
Subcategorization
Two main types are recognized:

i) using features which assign the entry to a specific class, whose syntactic behaviour is described elsewhere in the system;
ii) specifying the number of slots and the types of elements that may appear as complements in each slot.

The second type has some operational advantages:

1) All kinds of subcategorizations and selectional restrictions which need to be stated as properties of particular lexical items can be easily handled without any special mechanism. Only the allowed patterns are listed in the lexicon. Any combination of complement types may be represented without having to decide beforehand on a particular inventory of possibilities.
2) All kinds of idioms and collocational restrictions can be potentially handled by specifying the exact wording of the lexical phrase.

3) An indefinitely large syntactic range may be "simulated" by treating as idioms syntactic constructions that may not be generated by the grammar.

The principle may be extended to the point where the lexicon "takes over" most of the grammar. If the principle is brought to its utmost effect, the grammatical patterns tend to be represented only in the specification of the lexical items to which they apply.

There are also some disadvantages:

- the lexicon becomes larger;
- fewer phenomena are treated in a general way;
- updating and additions present problems of length and difficulty;
- properties of lexical items that may in fact be predictable (on the basis of other lexical properties) must be specified anyway.

In some ways, the differences between the two systems may be reduced by automatic procedures. For example, a case frame can be mapped onto a feature representation, in which a given feature corresponds to a particular case pattern, or vice versa. E.g., the feature "transitive" can be mapped onto a case frame representation containing a direct object slot. Explicit representation of the case frame seems to allow more freedom. On the other hand, since features can be thought of as an indication of the inclusion into classes of lexical items, a single lexical feature may efficiently encode a range of possible case frames that tend to co-occur with particular types of words.

In other words, all the subcategorizational possibilities of a particular sense of a verb are taken to be predictable from a single feature representing its word-class membership.

Of course, in order to be able to take advantage of this type of generalization, one must have a detailed theory of the word classes of a language; and it is clear that a reasonably complete grammar must make reference to a very large set of such word classes.

This observation is, in a certain sense, the starting point for our feasibility study, described below.

Selectional restrictions

Two principal ways of representation are recognized:

a) Semantic restrictions are explicitly associated to each slot of a lexical entry:

b) The restrictions are not represented directly in the lexicon but are captured in another part of the system. E.g., in lexicons organized as semantic networks, hierarchically ordered concepts can be related one to the other by relations that specify the semantic roles of a given concept as well as the relations with other concepts that represent possible fillers of each role.

4.4. A "neutral" scheme of classification

Encouraged by the results obtained by B. Ingria and S. Cumming and also by the discussions which followed the workshop held in New York July 1986, we have promoted a working group which will involve outstanding representatives of the major current "linguistic schools". The group will investigate in detail the possibility of representing the linguistic information frequently used in parsers and generators (e. g., the major syntactic categories, subcategorization and complementation, verb classes, nominal taxonomies, etc.) in such a way that they can be reutilized in the following theoretical frameworks: government and binding; generalized phrase structure grammar; lexical functional grammar; relational grammar; systemic grammar; categorial grammar. This group will work on various languages. We shall start by examining in detail the treatment which the foregoing theories will assign to a representative sample of English and Italian verbs.

Let us suppose we are describing the Italian verbs by using the criteria, tests, and formal apparatus of a given theory. At the end, we shall subdivide the Italian verbs in classes, regrouping in a class all the verbs with the same description. We consider as members of the same class those verbs which have received the same description. The intuition we wish to prove is that the aforementioned theories will classify the Italian verbs substantially in the same way; in any case, the different theories will identify the same number of classes having the same members. Each theory will of course describe the syntactic behaviour of a class using its own formal and explicative apparatus. If this is true, it would be possible to label the verbs of the ILDB by distributing them into classes. The interface between the ILDB and a given theory and its relevant computational systems would thus contain the description of the syntactic behaviour of the different classes according to that theory.

This is, of course, only an abstract scheme, and we should envisage a number of strategies for its correct application. However, we feel that this intuition is the same as stating that the properties taken into account by the different theories are in large part the same, although differently described and explained.

In this framework, it should be possible to reutilize the partial descriptions of the Italian verbs so far produced by the different schools. In particular, the descriptions performed following the model of M. Gross (see Elia, 1984) should prove very useful.

5. Bilingual Components

A new direction towards which computational lexicography and lexicology are moving is the organization of bilingual lexical data base systems. We are now working on a project which uses bilingual machine-readable dictionaries as a "bridge" connecting two otherwise independent monolingual lexical data bases (see Calzolari/Picchi, 1986). One of the objectives is to integrate the different types of information traditionally contained in monolingual and bilingual dictionaries, so as to expand the informational content of the single components in the new integrated system.

Bilingual dictionaries contain more information about usage and fixed expressions, or idioms. This kind of information can obviously be well integrated in the monolingual dictionary and also made easy of access. This can be done automatically by means of pointers going from each full-word to the expression or the example in which it appears, with no redundancy in the storage of the data. The entries of the new system should therefore be of a composite nature, perhaps organized at different levels according to the different possibilities of access. We can envisage the original monolingual lexical entries, augmented with the different types of information coming from the corresponding bilingual entry: different sense discriminations, other examples, syntactic information, collocations, idioms, etc.

We can also reverse the perspective and examine bilingual entries provided with the information traditionally contained in monolingual entries, mostly definitions.

One of the two different viewpoints, both virtually present in the integrated bilingual system, will be simply activated and made available to the user by the first manner of access to the on-line bilingual lexical data base.

We would like to maintain in a unique structure both the independent features of the monolingual and bilingual dictionary sources and the integration of the two with different views on the data.

Moreover, we would like to introduce within the integrated system:

- the possibility of a standard look-up of the information given in natural language in traditional dictionaries;
- more sophisticated searching procedures for information retrieval operations on the data of the mono- and bilingual lexical data bases where the "natural language" data have been, where possible, transformed into "formalized" or "coded" data (features or relations). For example, the information which appears in the form of examples can also appear in a coded form giving the surface syntactic structure.

Some operations which can be performed automatically or semi-automatically on a machine-readable bilingual dictionary are the following:

- checking the reversibility of the two sides of the dictionary: it appears that the two sides actually present many differences in quantity, quality, and display of the information;
- fitting together the two sides and unifying them into an integrated whole. (In this case not everything can be automated.) This operation can be logically subdivided into two complementary steps: a) elimination of redundancies; and b) addition of new links from and to all the relevant lexical entries in both sides. A problem to take into account here is the fact that there are cases in which the lexicographer does not want to reverse his entries.

A monolingual lexical data base organized also as a thesaurus can prove to be very useful in extending the information provided by a bilingual dictionary. In the COLLINS ENGLISH–ITALIAN DICTIONARY this information falls into the category of what are called "semantic indicators". These may be field labels, synonyms, hyponyms, or contextual indicators such as typical subjects or objects of verbs, typical nouns of which an adjective can be predicated, etc.

The monolingual lexical data base can be used to expand the information, which is provided as a single word to the whole set of words to which it actually refers.

For example, the entry **pettinare** has different translations according to the contextual indicators referring to the object (in brackets):

pettinare ... *(capelli) to comb*
(tessuto) to comb, tease

In a certain sense the generic semantic restrictions on the possible object can be taken as a semantic feature and can be procedurally expanded by the monolingual thesaurus to all the possible hyponyms (to be generated at the time of the query) so that the appropriate translation can be chosen in any context where a specific name of *tessuto* ('material') is found. This is already possible in our ILDB.

Also with regard to bilingual dictionaries, the method we are adopting consists of reusing available data in machine-readable form by analyzing and transforming the information already contained in common dictionaries.

After the first processing phases that we have envisaged on the bilingual dictionary data, it will make no difference which of the two languages is taken as a starting point. In a certain sense, we would no longer have a source language and a target language, since the look-up and access procedures are independent and neutral with respect to direction, the dictionary being bidirectional. Bidirectional cross-references will also be automatically generated for the information contained at each sense level as semantic indicators, i.e., synonyms/hyperonyms or contextual indicators.

It is important when using a bilingual dictionary to be able to start from "groups" of words and to correlate them with a corresponding "group" of words in the other language. The information that serves to discriminate among different word senses and that can be formalized at the semantic level in a monolingual dictionary should in principle be of the same type that is given in bilingual dictionaries in the form of "semantic indicators" or "selective conditions" to constrain the choice of a particular translation.

Mapping between word senses in monolingual dictionaries and different translations in a bilingual dictionary is one of the most interesting of the problems concerning the connection of these different types of dictionaries. As one of the main problems in translation is the correct choice among the various meanings of lexically ambiguous words, we feel that it is absolutely necessary also for a Machine Translation or a Machine-Assisted Translation system to be linked to a linguistic data base, i.e., a source of lexical information organized in the form of a thesaurus by multi-dimensional taxonomies, where the possibility of disambiguating lexical items is at least semiautomated. One of the main uses of the system would be in machine-aided translation (MAT), as a powerful aid for translators. The end result may in fact be viewed as a 'translator workstation', where access is provided to many types of dictionaries and other lexical resources and where the power and the functions of lexical data bases and of textual data bases are exploited to the best advantage.

References

ALSHAWI, H.: Processing dictionary definitions with phrasal pattern hierarchies. Cambridge 1986.

ALSHAWI, H./B. BOGURAEV/T. BRISCOE: Towards a dictionary support environment for real-time parsing. In: Proceedings of the 2nd European Conference of the Association for Computational Linguistics. Geneva 1985.

AMSLER, R. A.: The Structure of the Merriam-Webster Pocket Dictionary. Ph. D. Thesis, Department of Computer Sciences, University of Texas. Austin 1980.

AMSLER, R. A.: Computational Lexicology: A Research Program. In: Proceedings of the American Federation for Information Processing Societies (AFIPS) 1982, 657–663.

AMSLER, R. A.: Machine-readable dictionaries. In: Annual Review of Information Science and Technology (ARIST). Ed. by M. E. WILLIAMS, ASIS, Knowledge Industry Publications, Vol. 19. 1984, 161–209.

AMSLER, R. A.: Deriving lexical knowledge base entries from existing machine-readable information sources. In: D. WALKER et al. (eds.) 1987.

ATKINS, B. T./J. KEGL/B. LEVIN: Explicit and Implicit Information in Dictionaries. In: Proceedings of the Conference on Advances in Lexicology. Waterloo 1986.

BOGURAEV, B. K.: Machine-readable Dictionaries in Computational Linguistics Research. In: D. WALKER et al. (eds.) 1987.

BRUSTKERN, J./K. D. HESS: Machine Readable German Dictionaries – From a Comparative Study to an Integration. In: Linguistica Computazionale 3. 1983 (Supplement), 77–93.

BUSA, R. (ed.): De Lexico Electronico Latino. In: Calcolo 5. 1968 (Suppl. 2.).

BYRD, R. J.: Word formation in natural language processing systems. In: Proceedings of the 8th International Joint Conference on Artificial Intelligence, Karlsruhe 1983, 704–706.

BYRD, R. J./N. CALZOLARI/M. S. CHODOROW/J. L. KLAVANS/M. S. NEFF/O. A. RIZK: Tools and Methods for Computational Lexicology. IBM T. J. Watson Research Center, unpublished 1986.

CALZOLARI, N.: Towards the organization of lexical definitions on a data base structure. In: COLING 82. Ed. by E. HAJICOVA. Prague: Charles University 1982, 61–64.

CALZOLARI, N.: Semantic links and the dictionary. In: Proceedings of the Sixth International Conference on Computers and the Humanities, Raleigh (North Carolina): Computer Science Press 1983a, 47–50.

CALZOLARI, N.: Lexical definitions in a computerized dictionary. In: Computers and Artificial Intelligence 2. 1983b, No. 3, 225–233.

CALZOLARI, N.: Detecting patterns in a lexical data base. In: Proceedings of Coling 84, Stanford University (Calif.): Association for Computational Linguistics, 1984, 170–173.

CALZOLARI, N.: Computer-aided lexicography: dictionaries and word databases. In: Computational Linguistics. Edited by I. S. BATORI, W. LENDERS, W. PUTSCHKE. Berlin: Walter de Gruyter, forthcoming.

CALZOLARI, N.: Structure and Access in an Automated Lexicon and Related Issues. In: D. WALKER et al. (eds.) 1987.

CALZOLARI, N./M. L. CECCOTTI: Organizing a Large Scale Lexical Database. In: Actes du Congres International Informatique et Sciences Humaines. Liege: L.A.S.L.A., 18–21 November 1981, 155–163.

CALZOLARI, N./E. PICCHI: The machine readable dictionary as a powerful tool for consulting large textual archives. In: Automatic Processing of Art History Data and Documents. Ed. by L. CORTI. Pisa: Scuola Normale Superiore 1983, 275–288.

CALZOLARI, N./E. PICCHI: A Project for a Bilingual Lexical Database System. In: Proceedings of the Conference on Advances in Lexicology. Waterloo 1986.

CHODOROV, M./R. BYRD/G. HEIDORN: Extracting semantic Hierarchies from a large on-line Dictionary. In: Proceedings of the 23rd Annual Meeting of the ACL. Chicago 1985, 299–304.

CHOMSKY, N.: Aspects of the Theory of Syntax. Cambridge/Massachusetts 1965.

COLLINS CONCISE ITALIAN–ENGLISH DICTIONARY. London: Collins 1985.

CUMMING, S.: The Lexicon in Text Generation. In: D. WALKER et al. (eds.) 1987.

ELIA, A.: Le verbe Italien. Fasano di Puglia 1984.

EVENS, M. W./B. E. LITOWITZ/J. A. MARKOWITZ/R. N. SMITH/O. WERNER: Lexical-Semantic Relations: a Comparative Survey. Edmonton/Alberta: Linguistic Research Inc. 1980.

GOETSCHALCKX, J./L. ROLLING: Lexicography in the Electronic Age. Amsterdam: North-Holland 1982.

GROSS, M.: Methodes en Syntax. Paris 1973.

GRUPPO DI PISA: Il Dizionario di Macchina dell'Italiano. In: Linguaggi e Formalizzazioni. Ed. by GAMBARARA, D., LO PIPARO, F., RUGGIERO, G. Roma: Bulzoni 1979, 683–707.

HARTMANN, R. R. K. (ed.): Lexicography: Principles and Practice. London. New York: Academic Press 1983.

HULTIN, N. C./H. M. LOGAN: The New Oxford English Dictionary Project at Waterloo: In: Dictionaries 6. 1984, 182–198.

INGRIA, R.: Lexical Information for parsing Systems: Points of Convergence and Divergence. In: D. WALKER et al. (eds.) 1987.

KAY, M.: The Dictionary of the Future and the Future of the Dictionary. In: ZAMPOLLI/CAPPELLI (eds.) 1983, 161–174.

KUNO, S.: The predictive Analyzer and a Path Elimination Technique. In: Communication for the Association of Computing Machinery 8. 1965, No. 7, 453–462.

MCNAUGHT, J.: Personal Communication. 1986.

MEIJS, W.: Lexical Organization from three different Angles. In: Journal of the ALLC 13. 1986, No. 1.

MICHIELS, A./J. NOEL: Approaches to thesaurus production. In: COLING 82: Proceedings of the Ninth International Conference on Computational Linguistics. Ed. by J. HORECKY. Amsterdam: North-Holland 1982, 227–232.

MOULIN, A./J. JANSEN/A. MICHIELS: Computer Exploitation of LDOCE's Grammatical Codes. London 1985. (Conference on Survey of English Usage.)

NAGAO, M./J. TSUJII/Y. UEDA/M. TAKIYAMA: An attempt to computerize dictionary data bases. In: GOETSCHALCKX/ROLLING (eds.) 1982, 51–73.

PICCHI, E./N. CALZOLARI: Textual perspectives through an automatized lexicon. In: Proceedings of the XII International ALLC Conference. Nice 1985.

SMALL, S.: Viewing word expert parsing as a linguistic theory. In: IJCAI 7. 1981, 70–76.

URDANG, L.: A lexicographer's adventures in computing. In: Dictionaries 6. 1985, 150–165.

WALKER, D. E./R. A. AMSLER: The Use of Machine-Readable Dictionaries in Sublanguage Analysis. In: Workshop on Sublanguage Analysis. Ed. by R. KITTREDGE. New York 1984.

WALKER, D./A. ZAMPOLLI/N. CALZOLARI (eds.): Automating the Lexicon: Research and Practice in a Multilingual Environment. Ms. 1987 (in print).

ZAMPOLLI, A.: Lexicological and Lexicographical Activities at the Istituto di Linguistica Computazionale. In: ZAMPOLLI/CAPPELLI (eds.) 1983, 237–278.

ZAMPOLLI, A.: Project d'un dictionnaire de machine. In: R. BUSA (ed.) 1968, 109–126.

ZAMPOLLI, A./N. CALZOLARI: Computational Lexicography and Lexicology. AILA Bulletin 1985, 59–78.

ZAMPOLLI, A./A. CAPPELLI (eds.): The Possibilities and Limits of the Computer in producing and publishing Dictionaries. In: Linguistica Computazionale (Pisa) 3. 1983.

ZIMMERMANN, H. H.: Multifunctional Dictionaries. In: ZAMPOLLI/CAPPELLI (eds.) 1983, 279–288.

Dr. Nicoletta Calzolari, Prof. Antonio Zampolli, Dipartimento di Linguistica, Università di Pisa; Istituto di Linguistica Computazionale del CNR, Pisa.

Deutsche Neudrucke · Reihe Barock

Wir stellen zur Subskription:

Justus Georg Schottelius
Ausführliche Arbeit
Von der Teutschen HaubtSprache

1663
Herausgegeben von WOLFGANG HECHT

2., unveränderte Auflage (Nachdruck der Auflage 1967)
2 Bde. mit zus. X, [XXXII], 1466, [XXVIII], 28 Seiten. Gebd.*
Subskriptionspreis DM 298.– / ca. US-$ 180.–; nach Erscheinen DM 358.– / ca. US-$ 218.–
ISBN 3-484-16008-1 (Deutsche Neudrucke. Reihe Barock. Bände 11/12)

Nirgends sind die grammatischen und sprachtheoretischen Gedanken der Epoche zwischen Valentin Ickelsamer und Leibnitz umfassender dargestellt als im Hauptwerk des Justus Georg Schottelius (1612–1676), in der »Ausführlichen Arbeit Von der Teutschen HaubtSprache«.

Das große, in der hier wiedergegebenen Erstausgabe 1526 Seiten umfassende Werk ist die Zusammenfassung von Schottelius' philologischem Lebenswerk: Er verschaffte damit der deutschen Grammatik und Poetik die philosophische Begründung, er vereinte Sprachtheorie und Praxis, und er war bestrebt, der deutschen Sprache jenen Wert zuzuweisen, den man ihr bis in seine Zeit hinein immer noch streitig machte. Er ging von der Idee der ›Grundrichtigkeit‹ der Sprache, von dem Problem sprachlicher Gesetzmäßigkeit also, aus und baute darauf seine Stammwortlehre, seine Sprachauffassung und seine Ansichten über Wesen und Wert der deutschen Sprache auf.

Diese Leistung Schottelius' und die Wirkungsgeschichte seines Hauptwerkes werden im Nachwort des Herausgebers eingehend erläutert. Beigefügt sind eine übersichtliche Kurzbiographie, die Urkundenmaterial aus dem Privatbesitz der Familie erarbeitet, ferner eine Bibliographie der sprachtheoretischen Schriften des Schottelius und ein Literaturverzeichnis, in dem neben den Abhandlungen über Leben und Werk des barocken Sprachgelehrten vor allem die weit verstreuten Arbeiten über seine Sprachauffassung gesammelt sind.

Vorher ist als Wörterbuch-Nachdruck erschienen:

Ulrike Haß
Leonhard Schwartzenbachs »Synonyma«

Beschreibung und Nachdruck der Ausgabe Frankfurt 1564
Lexikographie und Textsortenzusammenhänge im Frühneuhochdeutschen

1986. X, 600 Seiten und 1 Abb. DM 218.– / ca. US-$ 132.–. ISBN 3-484-30911-3
(Lexicographica. Series Maior. Band 11)

Niemeyer