DIADORIM – A Lexical Database for Brazilian Portuguese

Juliana Galvani Greghi*†, Ronaldo Teixeira Martins†, Maria das Graças Volpe Nunes*†

* Universidade de São Paulo – Instituto de Ciências Matemáticas e de Computação
Av. Trabalhador São-Carlense 400, 13560-970, São Carlos, São Paulo, Brazil
{jggreghi,mdgvnune}@icmc.sc.usp.br

† Núcleo Interinstitucional de Lingüística Computacional (NILC)
Av. Trabalhador São-Carlense 400, 13560-970, São Carlos, São Paulo, Brazil
ronaldo@nilc.icmc.sc.usp.br

Abstract

This paper aims at providing a general description for DIADORIM, a lexical database for Brazilian Portuguese. DIADORIM is said to successively merge two very different previous application-oriented dictionaries, increasing their user-friendliness, the reusability of their entries and their capability of incorporating new features. Besides improving the structure of the previous databases, DIADORIM also preserves their performance and functionality, as indicated in the real use and simulation tests carried out during its evaluation.

.

1. Introduction

How can one merge two very different application-oriented lexical databases? Is it possible to keep their distinctive properties without excessively increasing their complexity? Or should one take some system entropy for granted? Is it cost-effective to bring together morphemes, words, compound words, complex words, idioms and arbitrary strings of characters standing for recursive patterns in natural language generation? Is it suitable to generalize over semantic and syntactic features bound to very specific application-dependent routines? These are some of the issues addressed in the development of DIADORIM, a lexical database for Brazilian Portuguese. In what follows, we present and discuss some theoretical and practical decisions taken for the implementation of such a central repository of lexical data, which is meant to enhance the reuse of existing lexical resources and to guarantee coherence and consistency within and across dictionaries. The reader will find here a presentation of the motivations for DIADORIM (Section 2), a general description of its architecture (Section 3), its underlying knowledge and linguistic model (Sections 4 and 5), the history of its implementation (Section 6), and some conclusions on the process of merging two very different application-oriented lexical databases (Section 7).

2. Motivation

DIADORIM aims at unifying two very different lexical databases, used by two very different NLP tools carried out by NILC (Interinstitutional Nucleus for Computational Linguistics): ReGra and the UNL-Portuguese Server. The former is a Brazilian Portuguese grammar and style checker (Martins et al., 1998); the latter is a multilingual interlingua-based machine translation system (Nunes et al., 2001). Their goals and structure are completely different, and so are the dictionaries they use.

In the dictionary used by ReGra, the entries stand for isolated words, i.e., any string of characters surrounded by blank spaces appearing in Brazilian Portuguese texts. No morphological analysis is carried out, and inflected and derived forms are represented as single autonomous words, even if they constitute undetachable part of collocations and other complex fixed expressions. This is due both to the compacting algorithm, which cannot handle blank spaces inside dictionary entries, and to the lexical matching strategies, which cannot cope with juxtaposition and requires hence a blank space as a lexical border-marker. Both operational constraints - necessary for the efficiency of the system - hindered any possibility of representing analyzed or compound structures.

Furthermore, the input to the system is supposed to bring spelling and grammatical problems. Accordingly, ReGra's dictionary cannot be robust as to accept every entry; otherwise it would be impossible to prevent word-formation processes that, albeit regular and predictable, are forbidden by grammars and dictionaries. In this sense, ReGra's dictionary must comprise only authorized words, i.e, those explicitly appearing in Brazilian Portuguese official dictionaries.

Finally, ReGra - as a grammar and style checker - does not carry out any semantic analysis, and demands only morphological and syntactic information in the entries.

An example of a dictionary entry for ReGra is presented below:

cantar=<V.[BI.INT.TD.][FUT-SUBJ.ELE.FUT-SUBJ.EU.

INF-PESS.ELE.INF-PESS.EU]N.[a][cantar]0.>

For further information on the structure of the dictionary see (Nunes et al., *op cit.*).

In the UNL-Portuguese Server Dictionary, in turn, generation procedures lead to the representation of minimum-length recursive strings of characters, most of which are either smaller than morphemes (in order to avoid allomorphs, i.e., alternate spelling shapes of a morpheme) or larger than compound and complex words (in order to cope with collocations and recursive

phrases). As a consequence, the UNL-Portuguese Server Dictionary includes many non-existing words, parts of words, clusters of words, etc. Contrarily to ReGra, the input to the UNL-Portuguese Server is supposed to be correct, and the system is meant to be as robust as possible to analyze and generate any sequence of characters provided by a Brazilian Portuguese user.

The UNL-Portuguese Server - as any machine translation system - requires also richer dictionaries, with semantic and deeper morphological information (as inflection procedures and morphotactic flags).

Examples of UNL-Portuguese Server entries are presented below:

[cant] {} cant "sing" (ver,P05,stm,vd1,act)<P,0,0>;

For further information on the structure of the dictionary see (Dias da Silva et al., 1998).

The outstanding differences between both databases have often implied duplication of information and double, high-cost, very time-consuming maintenance efforts. DIADORIM is an attempt of overcoming most of these divergences. It is claimed to bring together both ReGra' s and the UNŁPortuguese Server dictionaries without impairing systems performance.

3. DIADORIM

DIADORIM was conceived as a general database formed by two very different structures: a knowledge-base (KB) and a language-base (LB). The former is a net-like structure representing semantic relations between lexical items; the latter is a tree-like structure assigning linguistic properties for the same items. The lexical entry, as a node of the KB structure and, at the same time, as a root of the LB structure, works as a bridge entity between both information sets. It brings the database as flexible as necessary to represent all required features. Figure 1 depicts the general architecture of DIADORIM:

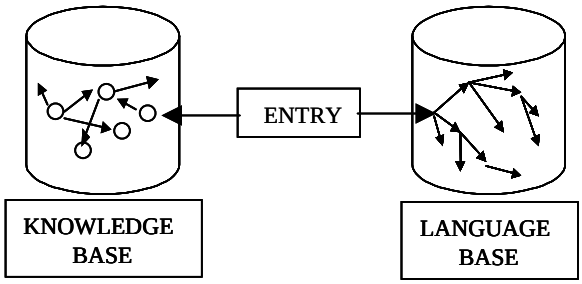

Figure 1 – General architecture of DIADORIM

4. The Knowledge-Base Structure

The knowledge-base (KB) structure constitutes a network comprising nodes and hyper-nodes standing, respectively, for concepts and conceptual structures. These nodes are linked to each other by means of arcs standing for semantic relations between concepts. Every non-linguistic information (as the fact, for instance, that rose is a kind of flower, that roses can be red or white, that they have thorns, that they can be used as gifts, etc.) would be assigned at this level. The KB structure plays then the role of an encyclopedia, and it is necessary for the semantic processing required by the UNL Project.

As to the set of arcs between nodes, we have adopted the set of relations predicted by the UNL Project (Uchida et al., 1999). They correspond to semantic-value relations linking concept-like information in a way very close to that intended by Fillmore (1968). Following the tradition in machine translation research, UNL adopts a comparatively large set of semantic cases. The inventory of relations can be divided into three large subgroups, according to functional aspects of relations between concepts, namely, the ontological, psychological, and logical subgroups.

Ontological relations are used as constraints in reducing lexical granularity or avoiding ambiguity. They are the guidelines for the thesaurus and can be compared to the concept of associative (paradigmatic) relation (Saussure, 1916), in contrast to the idea of syntagmatic relation conveyed by the psychological labels discussed below. The current version of DIADORIM states four ontological relations: synonymy (equ), antonymy (ant), meronymy (pof) and hyponymy (icl). Most of these relations have been automatically extracted from an electronic thesaurus for Brazilian Portuguese (TeP) (Dias-da-Silva et al, 2000).

Psychological relations can also be used to constrain concepts. However, its main function concerns the setting of general relations between co-occurring concepts. A UNL sentence is portrayed as a scene, depicting specific phenomena (actions, activities, events, processes, states, properties), each of which assign specific roles to be played by their participants and to which time and spatial boundaries, as well as other general modifiers, can be assigned. Psychological relations convey information on characters, space, time, plot and external causes of the described scene. Characters are considered to be any animate or inanimate participant playing any role in events. There can be up to eight characters in a scene, signaled by the following relations: agent (agt), co-agent (cag), object (obj), co-object (cob), beneficiary (ben), partner (ptn), instrument (ins) and affected place (opl). They are supposed to (help to) practice or to (help to) suffer state modifications during the scene. Place information can be assigned to the whole scene or to the characters in a scene. The scene can be referred to by its general physical (plc) or metaphysical (scn) place, as well as by its starting (plf) or finishing place (plt). Characters can be referred to by their origin (frm), their path (via) or their destination (to). Time is solely assigned to the entire scene and general reference to time (tim) can be detailed by reference to the starting (tmf) and finishing (tmt) times of the scene, its range (fmt), and duration (dur). A general plot structure can be thus adopted for the scene structure: from the initial state (src), characters are conducted to a final state (gol) through an intermediary state (via), by means of some specific method (met) or manner (man). Finally, the set of

psychological relations also comprises information on the initial (rsn) and final (pur) causes of the scene.

Logical relations are often isomorphic to first-order logic operators, and they are used to coordinate, subordinate or predicate characters inside a single scene, or scenes inside a more complex structure. The coordinate relations are associated with conjunction (and), disjunction (or) and co-occurrence (coo) of characters and whole scenes. The subordinate relations can be used to indicate comparison (bas), proportion (per), condition (con) and sequence (seq). Modification (mod), attribution (aoj), co-attribution (cao), quantification (qua), possession (pos), naming (nam) and content (cnt) are taken as predicate relations, in that they intervene in the status of the characters.

Figure 2 below depicts an excerpt from the general architecture of the KB structure.

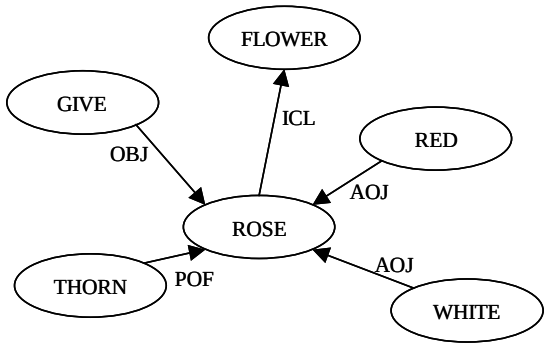

Figure 2 – Excerpt from the general architecture of KB structure

For further information see (Martins et al., 2000).

5. The Language-Base Structure

The language-base (LB) structure aims at representing a set of relations between lexical items. It is a multilayered structure comprising all levels of language description (phonetics, phonology, morphology, syntax, semantics and pragmatics) (Figure 3).

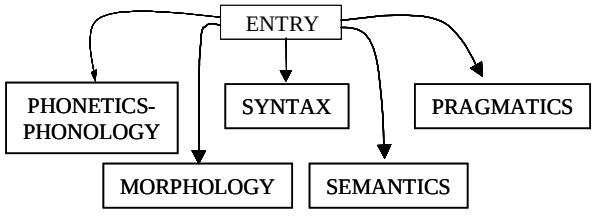

Figure 3 – General architecture of the Language-Base structure

At the phonetic-phonologic level only information about personal pronouns are represented, but we hope to complete this set of data as soon as possible.

At the morphological level, information classifies lexical items in one of four categories: morphemes, words, compound words or complex words. Morphemes are root and affixes (prefix and suffix) which, isolated, do not constitute valid lexical items. Words would involve combinations of roots and affixes allowed by Portuguese and attested by dictionaries. Compound words involve combinations of more than one root and are classified according to the specific morphological process (juxtaposition or agglutination). Complex words would correspond to expressions of language.

At the syntactic level, entries are classified according to one of the following categories: [+N,+V], [-N,-V], [+N,-V], [-V,+N], where [N] stands for a noun-like feature and [V] for a verb-like feature, as suggested by Chomsky (1970). Each of these categories is subclassified according to normal part-of-speech information and its corresponding morphosyntactic attributes. Government information and subcategorization procedures are also predicted for each entry. Figure 4 below shows the general architecture for the syntactic level.

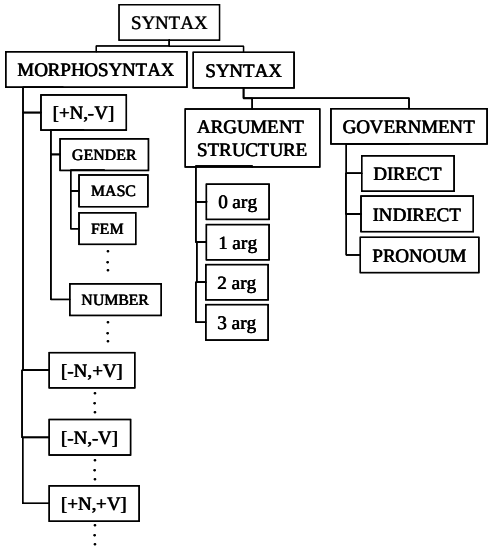

Figure 4 – General architecture for the syntactic level

Structures related to the semantic and pragmatic level have not been defined yet. They must be developed with results obtained by an ongoing project. They differ from the knowledge-base structure in the sense they should convey linguistic semantics information.

6. The Implementation

The second phase of the development of DIADORIM was to decide how to implement the database. In this work, we considered, basically, two possible ways to do that: using text files and special markup languages or using a database management system (DBMS).

The use of text files increases the portability of the resource, but requires powerful tools to process information and makes updating tasks more difficult. Another problem could be manipulation and information retrieval: the effort expended in a simple

query could be reduced by using a specific tool for data storage and manipulation, like a DBMS.

DBMSs might hinder the exchange of data between different research groups but would favor updating and recovering information tasks.

In this work, we have decided for using a DBMS based on the following:

- the need for organizing and representing information available at NILC under a same standard, aiming at increasing reusability.
- the large amount of data to be stored, extracted from the dictionary used in the ReGra Project, the Portuguese-UNL dictionary and the body of information provided by TeP.
- security

The next step was to decide which data model should be used to describe the structure of the database. For this work, we considered two models: Relational Model and Feature-Based Model (IDE ET AL., 1993).

The Relational model has been used in the recent past in the database application world and there are many commercial relational DBMSs. In this model, databases are represented as a collection of relations. Each relation resembles a table or a simple file. In a table, each row represents a set of related data values, describing a real-world entity or relationship.

Although this model has some disadvantages, it was capable of supporting a lexical database as proposed in this work.

Feature-based model relies on *feature structures*, which have been used in computational linguistics and natural language processing to encode linguistic information (see, for instance, Kaplan & Bresnan, 1982). The implementation of this model presents difficulties because there are no DBMSs based on feature structures. To enrich the expressiveness and flexibility required, an object-oriented DBMS could be used.

From the analysis of data sources and considering the real requirements of projects developed at NILC, we have opted to use Relational Model to implement the database. Reasons can be resumed as:

- high effort to reformulate the original set of data to adequate it to *Feature-based model*
- use of database to extract information exactly in the same format of original sources to support the existing projects

An Entity-Relationship (ER) Diagram was carried out for the lexical database corresponding to the structure proposed. This diagram preserved all previous information represented in the existing databases, avoiding overlaps between homographs and crossings between identical or very similar entries. An ER-to-Relational mapping guided the implementation of a database through a DBMS.

The database was created with a main purpose: to create a central repository of lexical data to organize information available on projects at NILC and to support the development of further projects. To achieve this goal, we design Diadorim in a way that insertion of new sets of information would not invalidate the previous work.

The task of transferring data from dictionaries to Diadorim was divided in two steps: a) formatting existing data according to the proposed model; and b) inserting data on new format into defined tables. The former was performed with an auxiliary tool, specially developed to manipulate original text files, to extract available data and to convert them to the new format required by Diadorim. The latter was performed with help of internal features available on DBMS.

It is important to say that the DBMS used in this application applies the B-Tree indexing technique on created tables. As the dictionary of ReGra has more than 1.5 million entries and a set of transactions, inserting data in many tables would put the system in a undesired status. The internal feature copies data from original data source into defined tables and is appropriated to insert a large amount of data.

To facilitate access to all tables and information extracted from database we create interface modules that allow for queries about ReGra, Portuguese-UNL project and a database which supports an Electronic Thesaurus. This interface was developed as a web page and is divided in four modules: an open-access module to consult morphosyntactic information, an open-access module to consult thesaurus information, a restricted-access module to consult Portuguese-UNL dictionary information and a restricted-access module to update/delete/insert information. Besides the development of these access interfaces, it was implemented a tool that can extract specific lists of information from Diadorim.

Consults are triggered by entering a word in the search box. User can apply some constraints in focused query, like to choose a specific grammatical category that must be consulted. Wildcards can be used to present a list of words that satisfied the applied constraints.

The Portuguese-UNL access module and the edition module can be accessed only by authorized person. This characteristic permits more control on the existing data and tries to avoid incoherence and inconsistency across information.

The specific tool created to help in future projects allows the user to choose which type of information must be presented on the output file. All information is presented in a text file that can be used to support studies about characteristics of language. This tool is not available on the web page.

Two modules of interface were evaluated by potential users: the morphosyntactic module and the thesaurus module. To evaluate these modules we applied two different methods: H*euristic Evaluation* (Molich & Nielsen, 1990) (Nielsen & Molich, 1990) and *Think Aloud (*Dix et al, 2000).

Heuristic evaluation aims at evaluating the system as a whole. Users' evaluations are independent and the user is encouraged to criticize the system. Think Aloud is an observational method and the user must “think aloud” while making use of the system. Each method was applied to a different group of researchers: only the latter was carried out by computational linguists. Problems pointed by them have been considered and a new version of these modules has been created.

The database underwent simulated stress tests, in which several parallel connections requested identical sequences of assorted queries. In the worst case (30 connections requesting 130 queries each), the whole demand was executed in 8 min (eight minutes), on average, which was considered a very good result, beyond our expectations.

7. Conclusions

At first, integrating the source dictionaries seemed to be a very difficult, almost impossible, task, and the exact set of the relationships that should hold between entries was rather elusive, to say the least. During design, a central challenge was to keep the particular features of each source model while allowing for further extensions and updates. In this respect, DIADORIM has so far lived up to our expectations: the inclusion of the TeP Project data and model already represents an (early) extension to our original design that was carried out in a sound manner, as expected/required.

One of the most important achievements of Diadorim is simplicity in data manipulation. Using a database, the integration between sets of information can now be performed through elaborated queries, and so are retrieved any relationships. DIADORIM provides for very flexible access and can be the basis of a host of possible tools, of which our current interfaces are "mere" examples.

DIADORIM has succesfully been used in the development of new applications, and NLP researchers are currently using the acess interface giving positive feedback.

Acknowledgements

The Brazilian Research Agencies FAPESP and CNPq have supported this work.

8. References

Chomsky, N. 1970. Remarks on nominalization. In Jacobs, R. A. & Rosenbaum, P. (eds.). Readings in English Transformational Grammar, Ginn and Companh, Waltham, Massachusetts.

Dias-da-Silva, B.C.; Sossolote C.; Zavaglia, C.; Montilha, G.; Rino, L.H.M.; Nunes, M.G.V.; Oliveira Jr., O.N.; Aluísio, S.M. 1998. The Design of the Brazilian Portuguese Machine Tractable Dictionary for an Interlingua Sentence Generator. Proceedings of III *Workshop on Writen and Spoken Portuguese Language Processing*, PROPOR' 98. XIV Brazilian Symposium on Artificial Intelligence. Porto Alegre. p.71-78. (http://nilc.icmc.sc.usp.br/download/Propor98DicUW.zip)

Dias da Silva, B.C.; Oliveira, M.F.; Moraes, H.R.; Paschoalino, C.; Hasegawa, R.; Amorim, D.; Nascimento, A.C. 2000.(*In Portuguese*) Building a Digital Thesaurus for Brasilian Portuguese. Proceedings of V *Workshop on Writen and Spoken Portuguese Language Processing*, PROPOR' 2000. XVI Brazilian Symposium on Artificial Intelligence. Atibaia. p. 1-11.

Dix, A.; Finlay, J.;Abowd, G.; Beale, R. 1999. *Human-Computer Interaction*, 2ed. Prentice Hall Europe.

Fillmore, Charles J. 1968. "The Case for Case". In Bach, E. and Harms, R.T. (eds.), *Universals in Linguistic Theory*. New York, Rinehart and Winston, 1-88.

Ide, N.; Le Maitre, J.; Véronis, J. 1993. Outline of a Model for Lexical Databases *Information Processing & Management*, 29(2), pp159-186.

Kaplan, R.; Bresnan, J. 1982. Lexical-functional grammar: A formal system for grammatical representation. *Mental Representation of Grammatical Relations.* Cambridge, Massachussets: MIT Press.

Martins, R.T.; Hasegawa, R.; Nunes, M.G.V.; Montilha, G.; Oliveira Jr., O.N. 1998. Linguistic issues in the development of ReGra: a Grammar Checker for Brazilian Portuguese. *Natural Language Engineering*. Volume 4 (Part 4 December 1998): p287-307;Cambridge University Press

Martins, R.T.; Rino, L.H.M.; Nunes, M.G.V.; Montilha, G.; Oliveira Jr., O.N. 2000. An interlingua aiming at communication on the Web: How language-independent can it be? Proceedings of the Workshop on Applied Interlinguas: Practical Applications of Interlingual Approaches to NLP, pp.24-33. *NAACL-ANLP 2000 Workshop*, April. Seattle, Washington, USA. (http://nilc.icmc.sc.usp.br/download/NILC-TR-00-1.zip)

Molich, R., Nielsen, J. 1990. Improving a human-computer dialogue, *Communications of the ACM* **33**, 3 (March), 338-348.

Nielsen, J., Molich, R. 1990.Heuristic evaluation of user interfaces. *Proc. ACM CHI'90* (Seattle, WA, 1-5 April), 249-256.

Nunes, M.G.V.; Martins, R.T.; Rino, L.H.M.; Oliveira Jr., O.N. The use of the Universal Networking Language for devising an automatic sentence generator for Brazilian Portuguese. Cadernos de Computação, 2(2), 2001, p.57-79, ICMC-USP.

Saussure, F de 1967. *Cours de Linguistique Générale*. Paris, Payot, 4 ed. Trad. bras. São Paulo, Cultrix, 1969.

Uchida, H., Zhu, M., & Della Senta, T. 1999. Universal Networking Language - A Gift for a Millenium. Institute of Advanced Studes, The United Nations University, Tokyo, Japan.