ANNE ABEILLÉ, *Treebanks. Building and Using Parsed Corpora*, Kluwer Academic Publishers, 2003[1]

The book edited by Anne Abeillé is a collection of 21 papers on building and using parsed corpora, most of them formerly presented at workshops and conferences (ATALA, LINC, LREC, EACL).

The objective of the book, as stated in the Introduction, is to present an overview of the work being done in the field of treebanks, the results achieved so far and the open questions. The addressees are linguists, computational linguists, psycholinguists, and sociolinguists.

The book is organized in two parts: *Building treebanks* (15 chapters, pp. 1-277) and *Using treebanks* (6 chapters, pp. 279-389), each of them having subparts. It also contains a preface (pp. xi), an introduction (pp. xiii-xxvi), a list of contributing authors and their affiliation (pp. 391-397), and an index of topics (pp. 399-405).

The organization of the Introduction (written by Anne Abeillé) mimics the structure of the whole book, namely it has two parts, entitled *Building treebanks* and *Using treebanks*, respectively. After making the terminological distinction between tagged corpora and parsed corpora (or treebanks), the editor emphasizes the reasons for the need of the existence of treebanks and makes a general presentation of the topics to be covered by the papers in the volume, stressing the fact that the problems encountered for each language are, at great extent, the same, thus a certain redundancy in the papers collected in this volume.

The chapters of the first part, *Building Treebanks*, are grouped according to the language or language families for which the approaches to building treebanks are presented: the first four chapters are dedicated to English treebanks, the next two to German ones; there are two papers on Slavic treebanks, four on Romance parsed corpora and the last three chapters of the first part address to treebanks for other languages (Sinica, Japanese, Turkish).

In Chapter 1, *The Penn Treebank: an Overview,* Ann Taylor, Mitchell Marcus, and Beatrice Santorini present the annotation schemes and the methodology used during the 8-year treebank project. The part-of-speech (POS) tagset is based on that of the Brown Corpus, but adjusted to serve the stochastic orientation of Penn Treebank and its concern with sparse data, and reduced to eliminate lexical and syntactic redundancies. More than one tag can be assigned to a word, thus avoiding arbitrary decisions. POS tags also contain syntactic functions, so they serve as basis for syntactic bracketing, which was modified during the project from a skeletal context free bracketing with limited empty categories and no indication of non-contiguous structures and dependencies to a style of annotation which aimed at clearly distinguishing between arguments and adjuncts of a predicate, recovering the structure of discontiguous constituents, and making use of null elements and coindexing to deal with wh-movement, passive, subjects of infinitival constructions. The first objective was not always easy to achieve via structural differences, that is why a set of easily

[1] This review was published in a previous version on LinguistList: http://linguistlist.org/pubs/reviews/get-review.cfm?SubID=26735&RequestTimeout=500.

identifiable roles were defined, although sometimes these ones proved difficult to apply, too. The Penn Treebank (PTB) project also produced disfluency annotation of transcribed conversations, labeling complete and incomplete utterances, non-sentence elements (filters, explicit ending terms, discourse markers, coordinating conjunctions) and restarts (with or without repair).

For all these three types of annotations a two-step methodology was adopted: an automatic step (represented by PARTS and Brill taggers for POS tagging, the Fidditch deterministic parser for syntactic bracketing, and a mere Perl script identifying common non-sentential elements) followed by manual correction.

In Chapter 2, *Thoughts on Two Decades of Drawing Trees,* Geoffrey Sampson enlarges on the idea that the annotation of both written and (transcribed) oral (real-life) corpora makes obvious the deficiencies of theoretical linguistics, may even contradict some widely accepted conventional linguistic wisdom (for instance, sentences of the form subject-intransitive verb are rather infrequent in English corpus, contrary to what can be found in some linguistics textbooks), and may yield findings about human language unsuspected before the existence of such resources.

The aim of Chapter 3, *Bank of English and beyond*, by Timo Järvinen is twofold. On the one hand, the author describes the four modules (pre-processing – i.e. segmentation and tokenization –, POS assignment, POS tagging, functional analysis) of the English Constraint Grammar (ENGCG) system (chosen for its morphological accuracy) used for annotating corpora for compiling the second edition of the Collins COBUILD Dictionary of English, and also the methodology adopted taking into consideration the huge amount of data that was to be dealt with; thus, manual inspection was possible only for some random fragments of the data and automatic methods were created for monitoring them.

On the other hand, Järvinen pleas for a Functional Dependency Grammar (FDG) parser. In spite of the morphological accuracy, in the CG system syntactic ambiguity was too high. The FDG parser better deals with long-distance dependencies, ellipses and other complex phenomena. He points out the need for a deep parsing, instead of the shallow one, his reason being, besides the lower ambiguity, the practical orientation of the former.

Sean Wallis entitled Chapter 4 *Completing Parsed Corpora*. A more challenging title for this paper could have been: "Do we need linguists for constructing treebanks?" For answering this question, Wallis starts by giving us a brief overview of the phases of the annotation employed on International Corpus of English – British Component (ICE-GB) and by pointing out the fact that the use of two parsers (i.e., TOSCA and Survey parser) increased the number of inconsistencies in the corpus, thus the necessity of a post-correction. He provides two arguments against Sinclair (Sinclair, J. (1992) The automatic analysis of corpora. In J. Svartvik (Ed.) Directions in Corpus Linguistics. *Proceeedings of Nobel Symposium 82*. Berlin: Mouton de Gruyter, pp. 379-397), who found human annotators a source of errors in the treebank.

In order to ensure the cleanness of the parsed corpus, one has two problems to solve: the decision (i.e. the correctness of the analysis) and the consistency (of the analysis throughout the corpus) ones. S. Wallis draws a distinction between longitudinal (that is, working through a corpus sentence-by-sentence, until it is completed) and transverse (i.e., working through a corpus construction-by-construction) correction, bringing arguments in favor of the latter: less time-consuming, control of the accuracy of the analysis and of its consistency. The price paid is difficulty in implementation and in managing the process. But once the tool for grammatical queries search facility (Fuzzy tree Fragment) is created, it can also be used not only for correction, but also for searching and browsing the corpus for linguistic queries, so a post-project use of the tool.

As clearly stated in the Critique section of Wallis's paper, the question formulated by us above receives an affirmative answer if the final aim of the corpus is not a study of the parser performance, but of language variation.

Chapter 5, *Syntactic Annotation of a German Newspaper Corpus*, by Thorsten Brants, Wojciech Skut, and Hans Uszkoreit, is a presentation of the syntactic annotation of the German NEGRA newspaper corpus. Language-specific reasons (free word order, among others), corpus structure (frequent elliptical constructions) and the characteristics of the formalism contributed to the choosing of Dependency Grammar for the annotation. However, it was modified so that to take advantage of phrase-structure grammar, too: flat structures, no empty categories, treatment of the head as a grammatical function expressed by labeling, not by the syntactic structure, allowance of crossing branches (which give rise to a large number of errors), a more explicit annotation of grammatical functions, encoding of predicate-argument information.

A characteristic of this project is the interactive annotation process which makes use of the TnT statistical tagger and second order Markov models for POS tagging. Syntactic structure is built incrementally, using cascaded Markov models. A graphical user interface allows for manual tree manipulation and runs taggers and parsers in the background. Human annotators need to concentrate only on the problematic cases, which are assigned different probabilities by statistical tagger and parser. Accuracy is ensured by annotating the same set of sentences by two different annotators. Differences are discussed and after agreeing on them, modifications are applied to the annotation.

The design of the corpus and the annotation scheme make it usable for different linguistic investigations and also for training taggers and chunkers.

Chapter 6, *Annotation of Error Types for German Newsgroup Corpus*, by Markus Becker, Andrew Bredenkamp, Berthold Crysmann, Judith Klein, is a presentation of the applications used for the development of controlled language and grammar checking applications for German.

The corpus in the FLAG project consisted of email messages (as they present the characteristics needed: high error density, accessibility, electronic availability). Their annotation was 3-phased: developing of a typology of grammatical errors in the target language (German), manual annotation on paper, and annotation by means of computer tools.

The first phase relied on traditional grammar books and its outcome was a type hierarchy of possible errors, also containing error domains (i.e. it tries to define the relations between the affected words) useful in guiding the detection of errors. Although the hierarchy was a fine-grained one, in the annotation process only a pool of 16 error types were to be detected and classified.

After being manually annotated, the same set of sentences were annotated in turn by means of two tools: Annotate and DiET. The annotation with the former one has a tree-format: the nodes are the error types, and the edges are descriptive information on these types; thus, a rich representation of the structure of errors in terms of relations. However, this representation is built bottom-up, the error-type being added last. DiET offers a better method for configuring an annotation schema; that is why the annotation was performed with this latter tool.

The overwhelming type of errors were the orthographical ones (83%), followed, at huge distance, by grammatical ones (16%).

In Chapter 7 Alena Böhmová, Jan Hajič, Eva Hajičová, Barbora Hladká present *The Prague Dependency Treebank*. For the annotation of the Czech newspaper corpus, a 3-level structure was used. At the morphological level, the automatic analyzer produces for each token in the input data the lemma and the associated MTag. Whenever more than one lemma and/or an MTag are produced, manual disambiguation is needed. For the analytical (syntactic) level of annotation the dependency structure was used. It is based on a dependency/determination relation. Solutions were found for problematic structures, as coordination, ellipses, ambiguity, and apposition. Two modes of annotation were employed: first, manual annotation, then the Collins parser was trained on such annotated data and further used to generate the structure, while syntactic functions went on being manually assigned. The separately produced morphological and analytical syntactic annotations are then merged together, all possible discrepancies being manually solved. The third level of annotation, the tectogramatical one, describes the meaning of the sentences in terms of tectogrammatical functions and the information structure of sentences. Analytic trees are transduced to tectogrammatical ones in two phases: an automatic one (which makes the necessary changes to syntactic trees, as merging the auxiliary nodes with verbs) and a manual one.

Chapter 8, *An HPSG-Annotated Test Suite for Polish*, Małgorzata Marciniak, Agnieszka Mykowiecka, Adam Przepiórkowski, Anna Kupść.

The aim of the paper is to present the construction of a test-suite for Polish, consisting of written sentences, both correct and incorrect ones, the latter being manually annotated with correctness markers. Each of these two types are further classified into three subgroups, according to their complexity. Moreover, each sentence is hand annotated with the list of linguistic phenomena they display, choosing from nine groups of hierarchies of such phenomena. Sentences are annotated with attribute-value matrices (AVMs), whose content is restricted by an HPSG signature. The result is a database of sentences, the correct ones augmented with their HPSG structures, and a database of wordforms. The aim of the former database is to evaluate computational grammars for Polish.

In Chapter 9, *Developing a Syntactic Annotation Scheme and Tools for a Spanish Treebank*, Antonio Moreno, Susana López, Fernando Sánchez, Ralph Grishman report on building an annotated Spanish corpora, based on newspaper articles. Problems specific to Spanish are presented: dealing with multiword constituents and with amalgams or portmanteau words, with null subjects and ellipses, "se"-constructions, etc. There are three levels of annotations: syntactic categories, syntactic functions, morpho-syntactic features and some semantic features. The annotation and debugging tools are also presented in the paper, alongside with some error statistics, current state of the Spanish treebank and future development.

Chapter 10, *Building a Treebank for French*, is authored by Anne Abeillé, Lionel Clément, François Toussenel. A newspaper corpus, representative of contemporary written French, was subject to automatic tagging (segmentation with special attention to compounds, tagging relying on trigram method, and retagging making use of contextual information) and parsing (surface and shallow annotation, theory-neutral, with the aim of identifying sentence boundaries and limited embedding). Each annotation with morphosyntax, lemmas (based on lexical rules), compounds and sentence boundaries was followed by manual validation. The resulting treebank was used for evaluating lemmatizers and for training taggers.

Chapter 11, *Building the Italian Syntactic-Semantic Treebank*, by Simonetta Montemagni, Francesco Barsotti, Marco Battista, Nicoletta Calzolari, Ornella Corazzari, Alessandro Lenci, Antonio Zampolli, Francesca Fanciulli, Maria Massetani, Remo Raffaelli, Roberto Basili, Maria Teresa Pazienza, Dario Saracino, Fabio Zanzotto, Nadia Mana, Fabio Pianesi, Rodolfo Delmonte, presents the syntactic-semantic annotation of a balanced corpus and of a specialized one. Four levels of annotations were adopted: morpho-syntactic annotation (POS, lemma, morpho-syntactic features), syntactic annotation made up of constituency annotation (identification of phrase boundaries and labeling of constituents) and functional annotation (with functional relations), lexico-semantic annotation (distinguishing among single lexical items, semantically complex units and title sense units; specification of senses for each word – relying on ItalWordNet – along with other lexico-semantic information, such as figurative usage, idiomatic expressions, etc.). The first two types of annotations were performed semi-automatically, while the other two were performed manually. There are two innovations brought about by this treebank: sense tagging (which resembles a semantic annotation of the corpus) and two distinct layers of syntactic annotation, the constituency and the functional ones, grounded by language specific phenomena (such as free constituent order and pro-drop property) and by further usages of the obtained treebank which is compatible with different approaches to syntax. In the second part of the article the annotation tool, GesTALt, is presented: its consisting applications and the architecture of the tool. In the end the usages of the obtained data are presented: improvement of a translation system, enrichment of dictionaries, improvement at the level of analysis.

Chapter 12 is called *Automated Creation of a Medieval Portuguese Partial Treebank* and authored by Vitor Rocio, Mário Amado Alves, J. Gabriel Lopes, Maria Francisca Xavier, Gracia Vicente. The novelty of the approach presented in this paper arises from the use of tools and resources developed for Contemporary Portuguese to the annotation of a corpus of Medieval Portuguese. The differences between these two phases of the language are presented. The neural-network based POS tagger was trained on a set of words manually tagged for each of the texts in the Medieval Portuguese corpus. It was then used to extract a dictionary and to tag the rest of the texts. Manual correction followed. For the lexical analysis, a morphocentric lexical knowledge-base (LKB) was used. The lexical analyzer uses as input the output from the POS tagger and applies to it the knowledge in the LKB. Its output serves as input for the syntactic analyzer. The authors present the resources used and the adaptations required to deal with the corpus. A similar method for dealing with corpora of other Romance languages is envisaged.

In Chapter 13, *Sinica Treebank*, Keh-Jiann Chen, Chi-Ching Lou, Ming-Chung Chang, Feng-Yi Chen, Chao-Jan Chen, Chu-Ren Huang, Zhao-Ming Gao report on the construction of a treebank for Mandarin Chinese, relying on Sinica Corpus, already annotated at the moment of starting the treebank, so its resources could be used for the latter. The authors provide reasons for their choosing of the grammar formalism used for the representation of lexico-grammatical information, namely Information-based Case Grammar. They also present the concepts they work with: the principles of inheritance, the phrasal categories, etc. Sinica treebank is not a mere syntactically annotated corpora, but also a semantically annotated one, containing thematic information. The automatic annotation process was followed by a manual checking, as in most cases. The language-specific phenomena (for instance, constructions with nominal predicates) are given a short presentation, along with the solution adopted in the annotation process. The treebank aims at being used as a reliable resource by (theoretical) linguists, but not only by them, so tools for extracting information from it were developed.

In Chapter 14, *Building a Japanese Parsed Corpus*, Sadao Kurohashi, Makoto Nagao preset the morphological and syntactic annotation of a Japanese newspaper corpus. It was developed in parallel with the improvement of the morphological analyzer JUMAN and of the dependency structure analyzer KNP (chosen in accordance with the characteristics of Japanese). The dependency relation is defined on bunsetsu, the traditional Japanese linguistic unit. The free word order of Japanese raised a problem which remained unsolved: predicate-argument relation in embedded sentences.

The aims of realizing the Turkish treebank are to be representative and to contain all the relevant information for its potential users. In Chapter 15, *Building a Turkish Treebank*, Kemal Oflazer, Bilge Say, Dilek Zeynep Hakkani-Tür, Gökhan Tür present its two levels of annotation: morphological and syntactical ones. Both take into consideration the characteristics of Turkish, especially its rich inflectional and derivational morphology. Thus, each word is annotated for each of its morphemes, as this information may be necessary for syntax. The free word order and the discontinuities favor the usage of the dependency framework. Its typical problems (pro-drop phenomenon, verb ellipsis, etc.) are given the solution adopted in the annotation process.

With Chapter 16, *Encoding Syntactic Annotation*, by Nancy Ide, Laurent Romary, we enter the second part of the book, *Using Treebanks*. The emerge of treebanks, alongside with the proliferation of annotation schemes, triggered the need for a general framework to accommodate these annotation schemes and the different theoretical and practical approaches. The general framework presented in this paper is an abstract model, theory and tagset independent, that can be instantiated in different ways, according to the annotator's approach and goal. This abstract model uses two knowledge sources: Data Category Registry (an inventory of data categories for syntactic annotation) and a meta-model (a domain-dependent abstract structural framework for syntactic annotation). Two other sources are used for the project-specific formats of the annotation scheme: Data Category Specification (DCS) (the description of the set of data categories used within a certain annotation scheme) and Dialect Specification (defining the project-specific format for syntactic annotation). Combining the meta-model with the DCS, a virtual annotation markup language (AML) can be defined for comparing annotations, for merging them or for designing tools for visualization, editing, extraction, etc. A concrete AML results from the combination of a virtual AML and Dialect Specification. The abstract model ensures the coherence and consistency of the annotation schemes.

The emergence of syntactic parsers triggered the need for methods evaluating them. In fact, this has become a real branch in the field of NLP research. In Chapter 17, *Parser Evaluation*, John Carroll, Guido Minnen, Ted Briscoe present a corpus annotation scheme that can be used for the evaluation of syntactic parsers. The scheme makes use of a grammatical relation hierarchy, containing types of syntactic dependencies between heads and dependents. Based on EAGLES lexicon/syntax standards (Barnett *et al.* 1996. EAGLES Recommemdations on Subcategorisation. Report of the EAGLES Working Group on Computational Lexicons, ftp://ftp.ilc.pi.cnr.it/pub/eagles/lexicons/

synlex.ps.gz.), this hierarchy aims at being language – and application – independent. Carroll *et al*. present a 10,000 words corpus semi-automatically marked up. For its evaluation three measures are calculated: precision (the number of bracketing matches with respect to the total number of bracketings returned by the parser), recall (the number of bracketing matches with respect to the number of bracketings in the corpus) and F-score (this is a measure combining the previous two measures: (2 x precision x recall)/(precision + recall)). This last measure can be used to illustrate the parser accuracy. The evaluation of grammatical relations provides information about levels of precision and recall for groups or single relations. Thus, they are useful for indicating the areas where more effort should be concentrated for bettering.

In Chapter 18, *Dependency-based Evaluation of MINIPAR*, Dekang Lin presents a dependency-based method for evaluating parsers performance. To represent a dependency tree he makes use of a set of tuples for each node in the tree, specifying the word, its grammatical category, its head (if the case, and also its position with respect to this head) and its relationship with the head (again, if the case). To perform the evaluation, for the parser generated trees (called here answers) and the manually constructed trees (called keys) dependency trees are generated and compared on a word-by-word basis. Very important, a selective evaluation is also possible: one can measure the parser performance with respect to a certain type of dependency relation or even to a certain word. Two scores are calculated: recall and precision. The author goes on with the presentation of MINIPAR, a principle-based broad coverage English parser (Berwick *et al*. 1991, *Principle-Based Parsing: Computation and Psycholinguistics*, Kluwer Academic Publishers). The dependency-based method presented above is used for evaluating this parser. One interesting outcome of this evaluation is that the parser performs better on longer sentences than on shorter ones. This may be the outcome of having trained the parser on press reportage, with long sentences, while the shorter sentences are found in fiction, the genre against which the parser is tested.

The assumption constituting the basis of Chapter 19, *Extracting Stochastic Grammars from Treebanks*, by Rens Bod is that "human language perception and production processes may very well work with representations of concrete past language experiences, and that language processing models could emulate this behavior if they analyzed new input by combining fragments of representations from annotated corpus". So, the idea is to use an already annotated corpus as a stochastic grammar. The idea is not new, but the aim of the article is to answer the question: in what measure can constraints be imposed on the used subtrees without decreasing the performance of the parser? The results reported here were obtained using a data-oriented parsing (DOP) model (presented in section 2 of the paper) which was applied to two corpora of phrase structure trees: Air Travel Information System (ATIS) and the Wall Street Journal (WSJ) part from PTB. The conclusion drawn from the experiments is that almost all constraints decrease the performance of the model: the most probable parse (which takes into consideration overlapping subtrees) gives better results than the most probable derivation (which does not takes it into consideration); the larger the subtrees, the better predictions (as larger subtrees capture more dependencies than small ones); the larger the lexical context (up to a certain depth, which seems to be corpus-specific), the better accuracy (as more lexical dependencies are taken into account); the low frequency subtrees have an important contribution to the parse accuracy (as they seem to be larger, thus to contain more lexical/structural context useful for further parsing); the use of subtrees with non-headwords have a good impact on the performance of the model (as they contain syntactic relations for those non-headwords, which cannot be found in other subtrees).

As stated in the title of Chapter 20, *A Uniform Method for Automatically Extracting Stochastic Lexicalized Tree Grammars from Treebanks and HPSG*, Günter Neumann presents a uniform method for automatically extraction of stochastic lexicalized tree grammars (SLTG) from treebanks (allowing corpus-based analysis of grammars) and HPSG (allowing extraction of domain-independent and phenomena-oriented subgrammars), with the future aim at merging the two SLTGs to improve the coverage of treebank grammars on unseen data and to ease adaptation of treebanks to new domains. The major operation in the extraction of SLTG is the recursive top-down tree decomposition according to the head principle, thus each extracted tree is automatically lexically anchored. The path from the lexical anchor to the root of the tree is called a head-chain. There are two more additional operations involved: each subtree of the head-chain is copied and the copied tree is processed individually by the decomposition operation, thus allowing a phrase to occur both in head and in non-head positions; for each SLTG-tree having a modifier phrase attached, a new tree is created with the modifier unattached, thus using the extracted grammar for recognizing sentences with less or no modifiers than the seen ones. There results a SLTG which is processed by a two-phase stochastic parser. The rest of the paper describes the extraction of SLTG from PTB and from NEGRA treebank, on the one hand, and from a set of parse trees with an English HPSG, on the other, and some experiments results of the use of an extracted SLTG.

Chapter 21, *From Treebank Resources to LFG F-Structures*, by Anette Frank, Louisa Sadler, Josef van Genabith, Andy Way, presents two methods for automatic f-structure annotation. The first one consists in extracting a Context-Free Grammar (CFG) from a treebank, according to Charniak 1996 ("Tree-bank Grammars" in *AAAI-96. Proccedings of the Thirteenth national Conference of Artificial Intelligence*, pp. 1031-1036. MIT Press). A set of regular expression based annotation principles are then developed and applied to the CFG, resulting an annotated CFG. The annotated rules are rematched against the treebank trees, the result being f(unctional)-structures. The second method uses flat tree descriptions. Annotation principles define Φ-projection constraints which associate partial c(onstituent)-structures with their corresponding partial f-structures. When these principles are applied to flat set-based encoding of treebank trees, they induce the f-structure. The two methods are characterized by robustness, due to the following facts: principles are partial, underspecified and match unseen configurations, partial annotations are generated instead of failure, the constraint solver cope with conflicting information.

Although this was not the objective of the book, its first part can be used as a textbook for those venturing to construct a treebank. As the papers here focus on different types of languages, displaying grammatical phenomena and different ways of dealing with them, these papers can serve as a repository of solutions to various problems encountered when trying to design a corpus, to establish a certain annotation scheme to be used for a treebank, to develop annotation tools. The style in which the papers were written is helpful in this respect: they are clear, accessible and the information is introduced gradually.

The second part of the book has a more reduced group of addressees than the first one, due to its technical details involved by the presentation of different application in computer linguistics: lexicon induction (Järvinen), grammatical induction (Frank *et al.*, Bod) parser evaluation (Carroll *et al.*), checker evaluation (Becker *et al.*).

Verginica Barbu Mititelu
Romanian Academy, Research Institute for Artificial Intelligence
and
Romanian Academy, Institute of Linguistics

FLORENTINA HRISTEA, MARIUS POPESCU (eds), *Building Awareness in Language Technology*, Bucureşti, Editura Universităţii din Bucureşti, 2003

The University of Bucharest, through the Faculty of Mathematics and Computer Science and the Faculty of Letters, has been appointed, by the European Commission, Regional Information Centre for Human Language Technology (HLT), addressing the entire Balkan area (coordinator: Dr. Florentina Hristea), during September 1, 2001 – February 28, 2003.

The activity of this centre – **RORIC-LING** – was part of the broader project **BALRIC-LING (*BALkan Regional Information Centers for awareness and standardization of LINGuistic resources and tools for advanced HLT applications)***, funded by the European Commission (contract no. IST-2000-26454).

Within the framework of this research-development project, Romanian computer scientists coming from the Faculty of Mathematics and Computer Science had significant contributions concerning human language technologies. A tool for corpus annotation, which is based on the dependency grammar formalism, has been designed (Dr. Marius Popescu). The importance of the tool resides primarily in the fact that it is language independent. Linguists from the Faculty of Letters have tested this Dependency Grammar Annotator and have offered examples of its usage (annotated samples) in the case of the Romanian language (Dr. Cristian Moroianu). Algorithms for the semiautomatic generation of WordNet type synsets and clusters, which are equally language independent, have also been created (Dr. Florentina Hristea) and tested for Romanian. The main reason for making this scientific effort was the desire and the necessity to create a uniform ontological infrastructure across languages, which will simplify machine translation from a language to another (particularly from English to Romanian and vice-versa) and will facilitate the use of the same reasoning schemes and algorithms developed in conjunction with the American WordNet. Thus, Romanian specialists successfully address a topic of great scientific interest for the present-day international human language technology community. Finally, within the framework of this project, the premises of a Morphological Dictionary of Romanian have been set (Dr. Emil Ionescu). This represents a topic of great interest in the case of any language and a scientific project which, to our knowledge, is under continuation for Romanian.

Although organized in order to offer scientific consulting related to these topics and others especially in the Balkans, the RORIC-LING information desk has been contacted by a great number of users from various countries, the scientific interest for this relatively new field going far beyond the original intended geographical area. Thus, RORIC-LING has registered a number of 291 subscribers, out of which 215 come from Romania and 76 from abroad. The foreign countries where RORIC-LING users exist are: Australia, Austria, Bulgaria, Canada, China, France, Germany, Greece, Great Britain, Italy, Norway, Turkey, and the United States.

At the end of the project duration, the RORIC-LING team has published "the project book" , "Building Awareness in Language Technology" (editors Florentina Hristea and Marius Popescu), Editura Universității din București, 2003, a volume which we would like to present in what follows, by means of the Foreword that the two editors wrote at the time:

EDITORS' FOREWORD[2]

This book encloses all papers authored by the members of the Romanian Regional Information Centre for Human Language Technologies (RORIC-LING), together with data samples and bulletins corresponding to the three virtual seminars that have been held every six months (2001-2003). It represents an attempt to help raise the awareness concerning some of the most advanced Human Language Technologies (HLT), as well as the possible scientific and industrial applications of the corresponding linguistic resources, both in Romania and in the entire Balkan region.

RORIC-LING is part of the BALRIC-LING project, funded by the European Commission (IST-2000-26454). The goal of RORIC-LING is that of building awareness in HLT primarily in Romania, a country still lacking true Language Engineering applications and HLT markets. The RORIC-LING information desk was open both to specialists and to nonspecialists, addressing subscribers coming from academic and research units, from software companies, as well as from other fields of activity.

Since HLT is a very broad field, RORIC-LING addresses only the following three topics:

– grammatical formalisms and their usage in the case of the Romanian language; corresponding tools for corpora annotation;

[2] Reprinted from "Building Awareness in Language Technology" (eds Florentina Hristea and Marius Popescu), Editura Universității din Bucureşti, 2003.

– semiautomatic generation of WordNet type Romanian synsets and clusters;

– a theoretical specification concerning a morphological model for Romanian.

The present book reflects the web site[3] that has been created within the framework of this project in order to facilitate communication with potential clients as well as quick dissemination of the information concerning all RORIC-LING topics. This web site contains all papers authored by the RORIC-LING specialists, accompanied by corresponding data samples and on-line demos, together with bulletins reflecting the virtual seminars that took place in connection with each of the RORIC-LING topics.

The project site itself is much richer than the contents of the present book. It includes comprehensive overviews, authored by specialists from ILSP (Greece) and Sheffield University (UK), which refer to the main BALRIC-LING topics. The same site is very rich in on-line demos that should be used by all visitors trying to get acquainted with a relatively new field.

Within this book the RORIC-LING topics are addressed bilingually, the same as in the project web page. The first part of the book is in English, while the second part represents the corresponding Romanian translation. Each of the two identical parts is organized according to the RORIC-LING topics. All materials have been taken directly from the RORIC-LING web site. As a result of a minimal editing effort, due to the time limits imposed by the project duration, and in order for the book to reflect the project web page as closely as possible, its style is not uniform. We ask the reader to accept our apologies for this inconvenience.

We hope this book will be of interest to all those involved in the field of HLT, but also to those who are not yet familiar with the field. The primary goal of this book is for it to act as an open invitation for its readers to access the project web page that, we hope, will have much more to offer.

The RORIC-LING team is greatly indebted to the European Commission for having encouraged this awareness effort. We would like to thank the European Commission for the importance it has attached to the RORIC-LING topics, as well as for having offered its full support. Special thanks are owed to Dr. Galia Angelova of the Bulgarian Academy of Sciences, Linguistic Modelling Department, the BALRIC-LING coordinator, for the continual support of our team within the framework of this project, as well as over the last years.

Florentina Hristea, Marius Popescu
The University of Bucharest – Regional Information
Centre for Human Language Technology

[3] http://phobos.cs.unibuc.ro/roric.