SUMMARY

In order to enter an international circuit, a language must reach a certain level of informatization. This means the existence of some resources and programs specially made for the respective language that can be stored and processed.

NLP (Natural Language Processing) researchers have as object of study the use of the computational means for text or speech processing in a natural language. By the investigation of the linguistic phenomena and by automatic retrieval of the information about the language from very large corpuses, automatic processing programs can be trained to process this information in an optimal way, making translations, summaries, statistic researches, providing automatic answers to questions or to vocal commands.

Besides the linear natural language corpuses, the informatized languages created also syntactic, semantic or discourse tree structures, called *treebanks*. The rules, in which these structures and annotated categories are formed, are also common to many languages so that they can be aligned.

Recent European comparative studies showed that, besides the English language, all the European languages have not a sufficient extent of informatization. Concerning the lexical, the Romanian language has an average level of informatization due to the alignment of the Romanian wordnet with the other Balkanic and European languages, but concerning the existence of annotated corpus of linguistic samples accessible to researchers, it is among the last ones.

Our research comes in the effort of correcting this deficiency by creating an annotated corpus at the lexical, morphologic and syntactic level. The goal of the research, the building of a treebank for the Romanian language, answers to an important necessity in the process of the language informatization.

The present paper, that describes this corpus and the process of its building, is structured in 2 sections: theoretical bases (chapters 1 and 3) and applied process (chapters 2, 4, 5 and 6).

One of the challenges of the researchers dedicated to language syntax is the building of a data base that includes enough examples of syntactic analysis from which a program, capable of generating a language model, will get its source. We describe in this paper not only the building of the resource, whose actual developing is preceded by the establishing of an annotation methodology of the syntactic phenomena, but also the simultaneously developing of an automatic syntactic annotation tool, i.e. the first phases of the training of the parser.

We established to create a complex and useful treebank that contains texts from a diverse range of language styles and that can be used in the training, testing and assessing of a parser for the Romanian language. A parser like this is momentarily being developed in Iaşi in a partnership research between the Theoretical Informatics Institute of the Romanian Academy and the Faculty of Computer Science of "Alexandru Ioan Cuza" University.

The present thesis is structured in 6 chapters. In the first chapter we define the main notions that make the object of this thesis: linguistic resources, natural language processing, corpus, annotation, treebank. A treebank is a corpus of texts where each sentence is associated to a syntactic tree structure (thus the name of "treebank"). The syntactic structures consist in lexical units connected by dependency binary relations, asymmetric, between a head and a dependent. Due to the fact that this building of the resource means specially the syntactic annotation of some collections of texts, we dedicated a section to the presenting of some general considerations and syntax notions and also to some details about syntactic units, relations and functions.

Concerning the syntactic theories of the linguists, we took into consideration the transformational generative grammar and the case grammar (Fillmore, 1968), and, among the syntactic models of the computer scientists, we discussed about the model based on immediate constituents, HPSG (Head-Driven Phrase Structure Grammar, Sag & Pollard, 1994), as well as the dependency grammar model (Tesnière, 1959), with the main axiom:

In a line, all the elements, except one and only one, are subordinated to other elements. So, as a dependency tree has only one word as root, the noun phrase will be subordinated to the predicate.

In the second chapter we present the steps of building a Romanian treebank, starting with the acquisition of a collection of language samples in the form of sentences or phrases. We described methods and techniques of the acquisition of these samples, as well as the principles followed in the selection, so that they illustrate a wider range of syntactic phenomena of the natural language or specific to Romanian language.

The lexical sources used in the first phase of the building of the corpus are represented by different belletristic texts from a set of grammar analyses, texts from Wikipedia, from *Acquis Communitaire*, a part of the texts from the English *FrameNet*, texts from The *Thesaurus Dictionary of the Romanian Language*, and a part from George Orwell's novel, "1984", which is considered a special lexical resource different from the rest of the belletristic texts due to its frequently use in NLP, being annotated by experts and aligned with versions in many languages.

We described then methods and programs used to store and process these linguistic data. The linguistic samples have been first transformed in a format that allows (non-arborescent) automated linear pre-annotation, assigning morphologic and syntactic categories to the constituents (lexemes and punctuation signs).

The morphological information mark-up on the corpus has been made automatically, first, with the RACAI webservice (racai.ro), then later, with the Iaşi NLP-tools webservice (Simionescu, 2011). These tools introduce the following information: the word isolated by tokenization, then the lemma, i. e. the base form of the word, followed by a line of letters that represent a code for the part of speech, its type, genre, number, case, determination category, person, tense of the verb, punctuation and the phrase limit.

In chapter 3 we detailed the principles and rules followed in the tree based hierarchic annotation of the sentence and phrases previously annotated in a linear way. The chapter contains a set of syntactic structures that make up an annotation guide. An annotation methodology of the corpus at the syntactic level represents a set of instructions that allows a consistent annotation with a linguistic theory.

The method in which we described the syntactic structures of the sentences from the natural language is one of D-trees; i. e. trees resulted after the syntactic analysis of a dependency grammar (Mel'čuk, 1987).

The dependency grammars represent the sentence structures like a set of dependency relations. The sentence is not built-up by syntactic groups, categories, but by words connected between them through dependency relations, stressing on the detailed specification of the connection between any 2 elements that are in a dependency relation (phobos.ro).

The ways of determining the dependency structure, which helps to the establishing of the dependency types, had as a quide the norms of the Academy Grammar, but there were some deviations from these norms. For example, according to the Academy Grammar, the adverb followed by the preposition “de” (“atât de”, “destul de”) has the syntactic function of quantitative adverb (“destul de frumos”)

“Destul de” became in our annotation a comparative element for “frumos”. We named the *comp*. relation for the arch between the preposition and the determined word. This convention was established in order to point out the particular character of these structures, because, in this present phase of the research, we decided to annotate in the same way all the mood adverbs, ignoring its sub-types.

Chapter 4 describes the most important step, the actual tree-base syntactic annotation. In this chapter we gave details of the way we used the interactive tool (TreeAnnotator) with which the annotation was done, pointing out different problems met during the process and the ways of solving them. It’s worth mentioning the principles of solving them according to the conventions of the axiomatic system of the dependency grammars which we adopted, establishing a solution that respects also the linguistic interpretations as they are deduced from the last referential academic works.

In chapter 5 we have presented the usage of the corpus in training, testing and assessing the output data after the processing of the texts with an automatic syntactic analyzer (parser). We have tested and assessed 2 of these types of analyzers, assessing and comparing the output. We’ve discussed the problems we met and their solutions.

After the entire process of annotation, manual and also automatic, we reached a number of 4 467 sentences or phrases annotated with TreeAnnotator, summing a total of approximately 105 000 words.

The 4 467 phrases which include a number of over 40 000 dependency relation are only a beginning, a starting point for the training process of one or more syntactic parsers for the Romanian language. We will continue with the syntactic annotation (this time automatic annotations which will be then corrected manually and reentered after the correction) till the parser will reach a satisfying level of accuracy. The more the program will be trained on a bigger number of texts (i. e. it will have a richer data base), the better its results will be.

A parser is a program capable of proposing a structure of the input text. Thus the program divides the line of linguistic signs in its compound parts, offering a classification of the syntactic function and relation of each part.

The FDG parser (*Functional Dependency Grammar*) discriminates between the rules of dependency and the rules of the surface ordering, following the Tesnière model of non-projectivity and adopting the concept of nuclei, primitive elements of the dependency structures possibly built of more lexemes.

In order to make a structure clear, choosing an interpretation out of many possibilities and in order to create relations between nodes, FDG uses certain strategies. A parsing is correct, in a context, when the desambiguization and the relationing can be done simultaneously (Popa, 2010).

The representations used in the syntactic analysis based on dependencies are made of lexical nodes connected by dependency arches that are annotated with types of dependency.

ROMParser, the first parser used in the building of our treebank, has at its base a similar version of the Nivre algorithm, who used non-deterministic procedures which guided the parser through a classifier trained on annotated texts, by linguists, with syntactic structures.

The final dependency structure represents the output of the *oracle* predictions. The oracle is a classifier that applies a non-determinist process[1] based on the machine learning model SVM (*support vector machines*). This classifier is previously trained on the tree corpus manually annotated in order to predict parsing actions using *a vector of characteristics*. The characteristics can be divided into 2 categories: *static* and *dynamic*.

The static characteristics remain constant during the parsing of a phrase. In this properties there are included the parts of speech of the words involved in a certain sequence of the parsing and its lemmas.

The dynamic characteristics are the one about the history of the parsing and the context of the target nodes that are to be parsed. In this category can be included the properties of the words next to the current nodes (like in the case of the concordance), the properties of the possible nodes that have been subordinated to these or the possible heads. The classifier uses the properties of the head nodes or dependent of the current node, i. e. the properties of the (sub) tree "branches" of which the target node belongs to.

For the evaluation of the system the following measurements has been used:

LAS – *Labeled attachment score*; represents the percentage of nodes for which there has been found the correct head and the correct dependency relation (57,89%);

UAS – *Unlabeled attachment score*; represents the percentage of nodes for which there has been found the correct head (66,30%);

LA – *Label accuracy*; represents the percentage of nodes for which there has been found the correct dependency relation (64,01%);

GHN – *Good head number*; represents the total number of trees from the tested set of texts for which the score LAS is 100% (4,04%).

The evaluation of the second parser had the results: *Label precision* – represents the percentage of nodes for which there has been found the correct dependency relation (62,75%); *Head precision* – represents the percentage of nodes for which there has been found the correct head (69,21%); *Both precision* - represents the percentage of nodes for which there has been found the correct head and the correct dependency relation (59,12%).

The parser behaves well taking into consideration the reduced size of the corpus used for training, succeeding, in spite of this obstacle, in achieving an accuracy not far from the one for the other languages, having in mind also the big difference between the sizes of the corpuses (for English the corpus had 2 500 000 words; for Czech – 1 500 000 words and for Romanian – 105 000 words).

Through the alternation of the training process with the automatic parsing of some new linguistic samples, there followed more parsing processes in order to achieve a greater number of syntactic annotated sentence/phrases, momentarily reaching a corpus of trees of 4 467 entries.

The thesis closes with a conclusion chapter and critical considerations on the original contributions of the research and its impact which the present study can have on the computational linguistics researches dedicated to the Romanian language, but also on the linguists' compared syntax researches.

[1] The process is non-deterministic because one can apply more transitions to a configuration.

The present work answers to 2 important directions of the informatization of the language: both the necessity of the building of complex annotated corpuses and the need of developing of some work tools; we described here the first steps not only in building a corpus with a complex annotation, at the syntactic level, in the form of D-trees, but also the way in which the recent created corpus can serve to the adapting and the training of a very useful and complex tool, the parser for the Romanian language.

We chose to build the treebank starting from the dependency grammars theory because this type of annotating conventions is at the base of the most tree type corpuses of the international languages, so the alignment in the future of the Romanian treebank with these will be possible.

We want to reach a size of the corpus of at least 20 000 linguistic samples or 500 000 annotated words, but this thing can be achieved only in a long period of time and a great effort.

The utility of the corpus will grow only if there will be applied consistently the same annotating conventions and they will be put in a common ground with the conventions used by other similar corpuses, like those from the project *Universal Dependencies* (Rosa & al., 2014).

The utility of this treebank can be an interdisciplinary one. For example, in the psycholinguistic and sociolinguistic field, the corpuses can be used in the evaluation of the predictions on the frequency of certain types of syntactic constructions. This fact can trace more or less certain defining traits of a person or a collectivity.

With the help of the treebanks, the linguists can search examples for a certain hypothesis or theory. Treebanks represent an important source of data for the testing of linguistic theories and hypotheses.

Once created, treebanks can stay at the base of the developing of other types of annotations (like the level of speech, semantic level or pragmatic one). After the lexical-morphologic and syntactic annotation, we can now make the semantic annotation. This would be of a great utility for some applications like: text classification, word sense desambiguization, multilingual texts alignment, questioning and answering systems, text inference recognition systems and others.

By the present paper, *Linguistic Resources for the Natural Language Processing*, we created a tree type syntactic corpus. The syntactic annotation conventions and the dependency relations used in our research started from the first trying of building a treebank for the Romanian language (Hristea & Popescu, 2003). We discussed the conventions used by these authors and we adopted sometimes different solutions, more consistent with the dependency grammars theory and with a greater range of covering over the special syntactic phenomena met in the corpus of annotated sentences and phrases of the Romanian language.

The next step is that the computer scientists should create programs that will transpose electronically the trees from that corpus in the format used by us, in order to include, of course with the authors' agreement, the old treebank in the new one, thus providing the Romanian language with a very large linguistic resource.

Such a research is useful not only to Romanian linguists and researchers from the NLP field, but also to other non-native researchers who want to study the specific linguistic phenomena of the Romanian language.