Automatic Extraction of Definitions in Portuguese: A Rule-Based Approach

Rosa Del Gaudio and António Branco

University of Lisbon
Faculdade de Ciências, Departamento de Informática
NLX - Natural Language and Speech Group
Campo Grande, 1749-016 Lisbon, Portugal
rosa@di.fc.ul.pt, antonio.branco@di.fc.ul.pt

Abstract. In this paper we present a rule-based system for automatic extraction of definitions from Portuguese texts. As input, this system takes text that is previously annotated with morpho-syntactic information, namely on POS and inflection features. It handles three types of definitions, whose connector between *definiendum* and *definiens* is the so-called copula verb "to be", a verb other that one, or punctuation marks. The primary goal of this system is to act as a tool for supporting glossary construction in e-learning management systems. It was tested using a collection of texts that can be taken as learning objects, in three different domains: information society, computer science for non experts, and e-learning. For each one of these domains and for each type of definition typology, evaluation results are presented. On average, we obtain 14% for precision, 86% for recall and 0.33 for F_2 score.

1 Introduction

The aim of this paper is to present a rule-based system for the automatic extraction of definitions from Portuguese texts, and the result of its evaluation against test data made of texts belonging to the domains of computer science, information society and e-learning.

In this work, a *definition* is assumed to be a sentence containing an expression (the *definiendum*) and its definition (the *definiens*). In line with the Aristotelic characterization, there are two types of definitions that typically can be considered, the formal and the semi-formal ones [1]. Formal definitions follow the schema $X = Y + C$, where X is the *definiendum*, " = " is the equivalence relation expressed by some connector, Y is the *Genus*, the class of which X is a subclass, and C represents the characteristics that turn X distinguishable from other subclasses of Y. Semi-formal definitions present a list of characteristics without the *Genus*.

In both types, in case the equivalence relation is expressed by the verb "to be", such definition is classified as a copula definition, as exemplified below:

J. Neves, M. Santos, and J. Machado (Eds.): EPIA 2007, LNAI 4874, pp. 659–670, 2007.

FTP é um protocolo que possibilita a transfêrencia de arquivos de um local para outro pela Internet.
FTP is a protocol that allows the transfer of archives from a place to another through the Internet.

Definitions are not limited to this pattern [2, 3]. It is possible to find definitions expressed by:

- punctuation clues:
  TCP/IP: protocolos utilizados na troca de informações entre computadores.
  TCP/IP: protocols used in the transfer of information between computers.

- linguistic expressions other than the copular verb:
  Uma ontologia pode ser descrita como uma definição formal de objectos.
  An ontology can be described as a formal definition of objects.

- complex syntactic patterns such as apposition, *inter alia*:
  Os Browsers, Navegadores da Web, podem executar som.
  Browsers, tools for navigating the Web, can also reproduce sound.

The definitions taken into account in the present work are not limited to copula definitions. The system is aimed at identifying definitory contexts based on verbs other than "to be" and punctuation marks that act as connectors between the two terms of a definition. Here, we will be calling verb definition to all those definitions that are introduced by a verb other than "to be", and punctuation definitions to the ones introduced by punctuation marks.

The research presented here was carried out within the LT4eL project[1] funded by European Union (FP6) whose main goal is to improve e-learning systems by using multilingual language technology tools and semantic web techniques. In particular, a Learning Management System (LMS) is being improved with new functionalities such us an automatic key-words extractor [4] and a glossary candidate extractor. In this paper, we will focus on the module to extract definition from Portuguese documents.

The reminder of the paper is organized as follows. In Sect. 2 we present the corpus collected in order to develop and test our system. In Sect. 3 the grammars developed to extract definition are described.

In Sect. 4, the results of the evaluation of the grammar, in terms of recall, precision and F2-score, are presented and discussed.

An errors analysis and a discussion on possible alternative methods to evaluate our system are provided in Sec. 5

In Sect. 6, we discuss some related work with special attention to their evaluation results, and finally in Sect. 7 conclusions are presented as well as possible ways to improve the system in future work.

[1] www.lt4el.eu

2 The Corpus

The corpus collected in order to develop and test our system is composed by 33 documents covering three different domains: Information Technology for non experts, e-Learning, and Information Society.

Table 1. Corpus domain composition

Domain	tokens
Information SocietyS	92825
Information Technology	90688
e-Learning	91225
Total	274000

Table 1 shows the composition of the corpus.

The documents were preprocessed in order to convert them into a common XML format, conforming to a DTD derived from the XCES DTD for linguistically annotated corpora [5].

The corpus was then automatically annotated with morpho-syntactic information using the LX-Suite [6]. This is a set of tools for the shallow processing of Portuguese with state of the art performance. This pipeline of modules comprises several tools, namely a sentence chunker (99.94% F-score), a tokenizer (99.72%), a POS tagger (98.52%), and nominal and verbal featurizers (99.18%) and lemmatizers (98.73%).

The last step was the manual annotation of definitions. To each definitory context was assigned the information about the type of definition. The definition typology is made of four different classes whose members were tagged with `is_def`, for copula definitions, `verb_def`, for verbal non copula definitions, `punct_def`, for definitions whose connector is a punctuation mark, and finally `other_def`, for all the remaining definitions. Table 2 displays the distribution of the different types of definitions in the corpus.

The domains of Information Society and Information Technology present a higher number of definitions, in particular of copula definitions. The reason could be that this domain is composed by documents conceived to serve as tutorials for non experts, and have thus a more didactic style. In Sect. 4, we will see how this difference can affect the performance of the system.

Table 2. The distribution of types of definitions in the corpus

Type	Information Society	Information Technology	e-Learning	Total
is_def	80	62	24	166
verb_def	85	93	92	270
punct_def	4	84	18	106
other_def	30	54	23	107
total	199	295	157	651

```
- <s id="s204">
  - <definingText def="m106" def_type1="is_def" id="d01">
      <tok base="o" class="word" ctag="DA" id="t4097" msd="ms" sp="y">O</tok>
    - <markedTerm dt="y" id="m106" kw="y">
        <tok base="tcp" class="word" ctag="PNM" id="t4098" sp="y">TCP</tok>
      </markedTerm>
      <tok base="ser" class="word" ctag="V" id="t4099" msd="pi-3s" sp="y">é</tok>
      <tok base="o" class="word" ctag="DA" id="t4100" msd="ms" sp="y">o</tok>
      <tok base="protocolo" class="word" ctag="CN" id="t4101" msd="ms" sp="y">protocolo</tok>
      <tok base="que" class="word" ctag="CJ" id="t4102" sp="y">que</tok>
      <tok base="dividir" class="word" ctag="V" id="t4103" msd="pi-3s" sp="y">divide</tok>
      <tok base="a" class="word" ctag="DA" id="t4104" msd="fs" sp="y">a</tok>
      <tok base="informação" class="word" ctag="CN" id="t4105" msd="fs" sp="y">informação</tok>
      <tok base="em" class="word" ctag="PREP" id="t4106" sp="y">em</tok>
      <tok base="pacote" class="word" ctag="CN" id="t4107" msd="mp" sp="y">pacotes</tok>
    </definingText>
  </s>
```

Fig. 1. The sentence `O TCP é um protocolo que divide a informação em pacotes (The TCP is a protocol that splits information into packets)` in final XML format

In Fig. 1, we present a sample of the final result. Of particular interest for the development of our grammars are the attribute `base`, containing the lemma of each word, the attribute `ctag`, containing the POS information, and the `msd` with the morpho-syntactic information on inflection.

3 The Grammars

The grammars we developed are regular grammars based on the tools lxtransduce, a component of the LTXML2 tool set developed at the University of Edinburgh[2]. It is a transducer which adds or rewrites XML markup on the basis of the rules provided.

Lxtansduce allows the development of grammars containing a set of rules, each of which may match part of the input. Grammars are XML documents conforming to a DTD (lxtransduce.dtd). The XPath-based rules are matched against the input document. These rules may contain simple regular-expression, or they may contain references to other rules in sequences or in disjunctions, hence making it possible to write complex procedures on the basis of simple rules.

All the grammars we developed present a similar structure and can be divided in 4 parts. The first part is composed by simple rules for capturing nouns, adjectives, prepositions, etc. The second part by rules that match verbs. The third part is composed by rules for matching nouns and prepositional phrases. The last part consist of complex rules that combines the previous ones in order to match the *definiens* and the *definiendum*.

[2] http://www.ltg.ed.ac.uk/~richard/ltxml2/lxtransduce-manual.html

A development corpus, consisting of 75% of the whole 274 000 token corpus, was inspected in order to obtain generalizations helping to concisely delimit lexical and syntactic patterns entering in definitory contexts. This sub-corpus was used also for testing the successive development versions of each grammar.

The held out 25% of the corpus was thus reserved for testing the system.

Three grammars were developed, one for each of the three major types of definitions, namely copula, other verbs, and punctuation definitions.

A sub-grammar for copula definition. Initially we developed a baseline grammar for this type of definition. This grammar marked as definition all that sentences containing the verb "to be" as the main verb of the sentence. In order to improve this grammar with syntactic information, copula definitions manually marked in the developing corpus were gathered. All information was removed except for the information on part-of-speech in order to discover the relevant patterns. Patterns occurring more than three times in the development corpus were implemented in this sub-grammar. Finally the syntactic patterns of all the sentence erroneously marked as definition by the baseline grammar were extracted and analyzed, in order to discover patterns that were common to good and bad definition. We decide not to implement in our grammar patterns whose occurrence was higher in the erroneously marked definitions that in the manually marked ones. We ended up with a grammar composed by 56 rules, 37 simple rules (capturing nouns, adjectives, prepositions, etc), 5 rules to capture the verb and 9 to capture noun and prepositional phrases and 2 rule for capturing the definitory context.

The following rule is a sample of the rules in the copula sub-grammar.

```
<rule name="copula1">
<seq>
<ref name="SERdef"/>
<best> <seq>
<ref name="Art"/>
<ref name="adj|adv|prep|" mult="*"/>
<ref name="Noun" mult="+"/> </seq>
<ref name="tok" mult="*"/>
<end/> </seq> </rule>
```

This is a complex rule that make use of other rules, defined previously in the grammar. This rule matches a sequence composed by the verb "to be" followed by an article and one or more nouns. Between the article and the noun an adjective or an adverb or a preposition can occur. The rule named `SERdef` matches the verb "to be" only if it occurs in the third person singular or plural of the present or future past or in gerundive or infinitive form.

A sub-grammar for other verbs definition. In order to develop a grammar for this kind of definitions we start immediately to extract lexico-syntactic patterns. In fact it is hard to figure out how a baseline grammar could be implemented. We decided to follow the same methodology used for copula definition.

In a first phase we extracted all the definitions whose connector was a verb other than "to be", and collected all such verbs appearing in the developing corpus, obtaining a list of definitory verbs. This list was improved by adding synonyms. We decided to exclude some verbs initially collected from the final list because their occurrence in the corpus is very high, but their occurrence in definitions is very low. Their introduction in the final list would not improve recall and would have a detrimental effect on the precision score.

In a second phase we divided all the verbs obtained in three different classes: verbs that appear in active form, verbs that appear in passive form and verb that appear in reflexive form. For each class a syntectic rule was wrote. This information was listed in a separate file called lexicon.

The following rule is a sample of how verbs are listed in the lexicon.

```
<lex word="significar"> <cat>act</cat> </lex>
```

In this example the verb *significar* ("to mean") is listed, in his infinitive form that corresponds to the attribute `base` in the corpus. The tag `cat` allows to indicate a category for the lexical item. In our grammar, `act` indicates that the verb occurs in definitions in the active form. A rule was written to match this kind of verbs:

```
<rule name="ActExpr">
<query match="tok[mylex(@base) and (@msd[starts-with(.,'fi-3')]
or @msd[starts-with(.,'pi-3')])]"constraint="mylex(@base)/cat='act'"/>
<ref name="Adv" mult="?"/>
</rule>
```

This rule matches a verb in present and future past (third person singular and plural), but only if the base form is listed in the lexicon and the category is equal to `act`. Similar rules were developed for verbs that occur in passive and reflexive form.

A sub-grammar for punctuation definition. In this sub-grammar, we take into consideration only those definitions introduced by a colon mark since it is the more frequent pattern in our data. The following rule characterizes this grammar. It marks up sentences that start with a noun phrase followed by a colon.

```
<rule name="punct_def">
<seq> <start/>
<ref name="CompmylexSN" mult="+"/>
<query match="tok[.~'^:']"/>
<ref name="tok" mult="+"/>
<end/> </seq> </rule>
```

4 Results

In this section we report on the results of the three grammar. Further more the results of a fourth grammar are presented. This grammar was obtained by

combing the previous three in order to obtain the general performance of our system.

Scores for Recall, Precision and F_2-measure, for developing corpus (dv) and for test (ts) corpus are indicated.

These scores were calculated at the sentence level: a sentence (manually or automatic annotated) is considered a true positive of a definition if it contains a part of a definition. Recall is the proportion of the sentences correctly classified by the system with respect to the sentences (manually annotated) containing a definition. Precision is the proportion of the sentences correctly classified by the system with respect to the sentences automatically annotated.

We opted for an F_2-measure instead of an F_1 one because of the context in which this system is expected to operate. Since the goal is to help the user in the construction of a glossary, it is important that the system retrieves as many definition candidates as possible. The final implementation will allow user to quickly delete or modificate bad definitions.

Table 4 displays the results of the copula grammar. These results can be put in contrast with those obtained with a grammar that provides the performance of the baseline grammar for copula definitions. An improvement of 0.18 in the F_2-measure was obtained.

Table 3. Baseline for copula grammar

	Precision		Recall		F_2	
	dv	ts	dv	ts	dv	ts
IS	0.11	0.12	1.00	0.96	0.27	0.29
IT	0.09	0.26	1.00	0.97	0.22	0.51
e-L	0.04	0.50	0.82	0.83	0.12	0.14
Total	0.09	0.13	0.98	0.95	0.22	0.31

Table 4. Results for copula grammar

	Precision		Recall		F_2	
	dv	ts	dv	ts	dv	ts
IS	0.40	0.33	0.80	0.60	0.60	0.47
IT	0.26	0.51	0.56	0.67	0.40	0.61
e-L	0.13	0.16	0.54	0.75	0.26	0.34
Total	0.30	0.32	0.69	0.66	0.48	0.49

The results obtained with the grammar for other verbs are not as satisfactory as the ones obtained with the copula grammar. This is probably due to the larger diversity of patterns and meaning for each such verb.

As can be seen in Table 2, only 4 definitions of this type occur in the documents of the IS domain and 18 in the e-learning domain. Consequently, this grammar for punctuation definitions ended up by scoring very badly in these documents. Nevertheless, the global evaluation result for this sub-grammar is better than the results obtained with the grammar for other verb definitions.

Finally, Table 7 presents the results obtained by a grammar that combines all the other three sub-grammars described in this work. This table gives the overall performance of the system based on the grammars developed so far, that is, this result represents the performance the end user will face when he will be using the glossary candidate detector.

To obtain the precision and recall score for this grammar, it is not necessary anymore to take into account the type of definition. Any sentence that is correctly tagged as a definitory context (no matter which definition type it receives) will be brought on board.

Table 5. Results for verb grammar

	Precision		Recall		F_2	
	dv	ts	dv	ts	dv	ts
IS	0.13	0.08	0.61	0.78	0.27	0.19
IT	0.13	0.22	0.63	0.66	0.28	0.39
e-L	0.12	0.13	1	0.59	0.28	0.27
Total	0.12	0.14	0.73	0.65	0.27	0.29

Table 6. Results for punctuation grammar

	Precision		Recall		F_2	
	dv	ts	dv	ts	dv	ts
IS	0.00	0.00	0.00	00	0.00	0.00
IT	0.48	0.43	.68	0.60	0.60	0.53
e-L	0.05	0.00	0.58	0.00	0.13	0.00
Total	0.19	0.28	0.64	0.47	0.35	0.38

Table 7. The combined result

	Precision		Recall		F_2	
	dv	ts	dv	ts	dv	ts
IS	0.14	0.14	0.79	0.86	0.31	0.32
IT	0.21	0.33	0.66	0.69	0.38	0.51
e-L	0.11	0.11	0.79	0.59	0.25	0.24
Total	0.15	0.14	0.73	0.86	0.32	0.33

As can be seen, the recall value remains quite high, 86%, while it is clear that for the precision value (14%), there is much room for improvement yet.

In many cases we obtained better results for the test corpus than for the developing one, and this represents an unsuspected result. In order to explain this outcome we analyzed the two sub-corpora separately. When we split the corpus in two parts we just toke in account the size of the two corpora end not the number of definitions in each one. As a matter of fact the developing corpus is characterized by 364 definitory contexts while the training corpus is characterized by 287 definitions. This means that the developing corpus has 55% of all definition instead of 75%, as a consequence the definition density is lower in this corpus than in the test corpus. This result supports our initial hypothesis that the style of a document influence the performance of the system.

5 Error Analysis

As expected, the results obtained with documents from the Information Society and Information Technology domains are better than the results obtained with documents from the e-Learning domain. This confirms our expectation drawn from the style and purpose of the material involved. Documents with a clear educational purpose, like those from IS and IT sub-corpora, are more formal in the structure and are more directed towards explaining concepts, many times via the presentation of the associated definitions. On the other hand, documents with a less educative purpose present less explicit definitions and for this reason it is more difficult to extract definitory contexts from them using basic patterns. More complex patterns and a grammar for deep linguistic processing are likely to be useful in dealing with such documents.

Also worth noting is the fact that though the linguistic annotation tools that were used score at the state of the art level, the above results can be

improved with the improvement of the annotation of the corpus. A few errors in the morpho-syntactic annotation were discovered during the development of the grammars that may affect the performance of the grammars.

On the evaluation methodology. Determining the performance of a definition extraction system is not a trivial issue. Many authors have pointed out that a quantitative evaluation as the one we carried out in this work may not be completely appropriate [7]. The first question that arises is about the definition of definitions. If different people are given the same document to mark with definitions, the result may be quite different. Some sentences will be marked by everybody, while others will not. As show in similar studies [8], when different people are asked to mark definitions in a document agreement may be quite low. This can in part explain the low precision we obtained for our system.

Also interesting to note is that a different, perhaps complementary, approach to evaluation is possible. The DEFINDER [9] system for automatic glossary construction was evaluated using a qualitative approach. The definitions extracted by the system were evaluated by end users along three different criteria: readability, usefulness and complexness, taking into account the knowledge of the end user in the domain. This method may be an important complement to the quantitative approach used here.

6 Related Work

In this section, we discuss studies that are related to the work reported here and whose results and methods may be put in contrast with ours.

Note however that comparing the results of different studies against ours is in most cases not easy. Many researches only report recall or precision, or don't specify how these values were obtained (e.g. token level vs. sentence level). The nature of the corpora used (size, composition, structure, etc.) is another sensible aspect that makes comparison more difficult. Also different systems are tuned to finding definitions with different purposes, for instance for relations extraction, or for question answering, etc.

Regarding the methods used for this task, the detection of morpho-syntactic patterns is the most used technique. Since the beginning of 90's Hearst [10] proposed a method to identify a set of lexico-syntactic patterns (e.g. such NP as NP, NP and other NP, NP especially NP, etc) to extract hyponym relations from large corpora and extend WordNet with them. This method was extended in recent years to cover other types of relations[11]. In particular, Malaise and colleagues [2] developed a system for the extraction of definitory expressions containing hyperonym and synonym relations from French corpora. They used a training corpus with documents from the domain of anthropology and a test corpus from the domain of dietetics. The evaluation of the system using a corpus of a different domain makes results more interesting as this puts the system under a more stressing performance. Nevertheless, it is not clear what is the nature and purpose of the documents making this corpora, namely if they are consumer-oriented, technical, scientific papers, etc. These authors used lexical-syntactic

markers and patterns to detect at the same time definitions and relations. For the two different, hyponym and synonym, relations, they obtained, respectively, 4% and 36% of recall, and 61% and 66% of precision.

Saggion [12] combines probabilistic technics with shallow linguistic analysis. In a first stage are collected 20 candidate texts with a probabilistic document retrieval system using as input the definition question expanded with a list of related terms. In a second stage the candidate sentences are analyzed in order to match the definition patterns. Regarding the results obtained at the TREC QA 2003 competition, he reports F_5 score, where the recall is 5 times more important than precision. His system obtained a F-score of 0.236, where the best score in the same competition was of 0.555 and the median was of 0.192. The main difference between this task and our work resides in the fact that we do not know beforehand the expressions that should receive definitions. This lack of information makes the task more difficult because it not possible to use the term as a clue for extracting its definitions.

DEFINDER [9], an automatic definition extraction system combines shallow natural language processing with deep grammatical analysis. Furthermore it makes use of cue-phrase and structural indicators that introduce definitions and the defined term. In terms of quantitative evaluation, this system presents 87% precision and 75% recall. This very high values are probably due to the nature of the corpus composed by consumer oriented medical jornal article.

More recently machine learning techniques were combined with patterns recognition in order to improve the general results. In particular [13] used a maximum entropy classifier to extract definition and syntactic features to distinguish actual definitions from other sentences.

Turning more specifically to the Portuguese language, there is only one publication in this area. Pinto and Oliveira [14] present a study on the extraction of definitions with a corpus from a medical domain. They first extract the relevant terms and then extract definitions for each term. The comparison of results is not feasible because they report results for each term. Recall and precision range between 0% and 100%.

By using the same methodology for Dutch as the one used here, Westerhout and Monachesi [8] obtained 0.26 of precision and 0.72 of recall, for copula definitions, and 0.44 of precision and a 0.56 of recall, for other verbs definition. This means that their system outperforms ours in precision though not in recall.

7 Conclusions and Future Work

In this work, we presented preliminary results of a rule-based system for the extraction of definitions from corpora. The practical objective of this system is to support the creation of glossaries in e-learning environments, and it is part of a larger project aiming at improving e-learning management systems with human language technology.

The better results were obtained with the system running over documents that are tutorials on information technology, where it scored a recall of 69%

and a precision of 33%. For less educational oriented documents, 59% and 11%, respectively, was obtained.

We also studied its performance on different types of definitions. The better results were obtained with copula definitions, with 67% of recall and 51% of precision, in the Information Technology domain.

Compared to work and results reported in other publications concerning related research, our results seem thus very promising. Nevertheless, further strategies should be explored to improve the performance of the grammar, in particular its precision.

In general, we will seek to take advantage of a module that allows deep linguistic analysis, able to deal with anaphora and apposition, for instance. At present, we know that a grammar for deep linguistic processing of Portuguese is being developed [15]. We plan to integrate this grammar in our system.

Regarding punctuation definition, the pattern in the actual grammar can also be extended. At present, the pattern can recognize sentences composed by a simple noun followed by a colon plus the definition. Other rules with patterns involving brackets, quotation marks, and dashes will be integrated.

Finally, in this work we ignored an entire class of definitions that we called "other definition", which represents 16% of all definitions in our corpus. These definitions are introduced by lexical clues such as *that is*, *in other words*, etc. This class also contains definitions spanning over several sentences, where the terms to be defined appear in the first sentence, which is then characterized by a list of features, each one of them conveyed by expressions occurring in different sentences. These patterns need thus also to be taken into account in future efforts to improve the grammar and its results reported here.

References

[1] Meyer, I.: Extracting knowledge-rich contexts for terminography. In: Bourigault, D. (ed.) Recent Advances in Computational Terminology, pp. 279–302. John Benjamins, Amsterdam (2001)

[2] Malais, V., Zweigenbaum, P., Bachimont, B.: Detecting semantic relations between terms in definitions. In: CompuTerm 2004. CompuTerm Workshop at Coling 2004, 3rd edn., pp. 55–62 (2004)

[3] Alarcn, R., Sierra, G.: El rol de las predicaziones verbales en la extraccin automtica de conceptos. Estudios de Linguistica Aplicada 22(38), 129–144 (2003)

[4] Lemnitzer, L., Degórski, L.: Language technology for elearning – implementing a keyword extractor. In: EDEN Research Workshop Research into online distance education and eLearning. Making the Difference, Castelldefels, Spain (2006)

[5] NIKS: Xml, corpus encoding standard, document xces 0.2. Technical report, Department of Computer Science, Vassar College and Equipe Langue ed Dialogue, New York, USA and LORIA/CNRS, Vandouvre-les-Nancy, France (2002)

[6] Silva, J.R.: Shallow processing of Portuguese: From sentence chunking to nominal lemmatization. Master's thesis, Universidade de Lisboa, Faculdade de Ciências (2007)

[7] Przepiórkowski, A., adn Miroslav Spousta, L.D., Simov, K., Osenova, P., Lemnitzer, L., Kubon, V., Wójtowicz, B.: Towards the automatic extraction of definitions in Slavic. In: Piskorski, J., Pouliquen, B., Steinberger, R., Tanev, H. (eds.) Proceedings ofo the BSNLP workshop at ACL 2007, Prague (2007)
[8] Westerhout, E., Monachesi, P.: Extraction of Dutch definitory contexts for elearning purposes. In: CLIN proceedings 2007 (2007)
[9] Klavans, J., Muresan, S.: Evaluation of the DEFINDER system for fully automatic glossary construction. In: AMIA 2001. Proceedings of the American Medical Informatics Association Symposium (2001)
[10] Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th conference on Computational linguistics, Morristown, NJ, USA, Association for Computational Linguistics, pp. 539–545 (1992)
[11] Person, J.: The expression of definitions in specialised text: a corpus-based analysis. In: Gellerstam, M., Jaborg, J., Malgren, S.G., Noren, K., Rogstrom, L., Papmehl, C. (eds.) EURALEX 1996. 7th Internation Congress on Lexicography, Goteborg, Sweden, pp. 817–824 (1996)
[12] Saggion, H.: Identifying definitions in text collections for question answering. In: LREC 2004 (2004)
[13] Fahmi, I., Bouma, G.: Learning to identify definitions using syntactic feature. In: Basili, R., Moschitti, A. (eds.) Proceedings of the EACL workshop on Learning Structured Information in Natural Language Applications, Trento, Italy (2006)
[14] Pinto, A.S., Oliveira, D.: Extraçcão de definições no Corpógrafo. Technical report, Faculdade de Letras da Universidade do Porto (2004)
[15] Branco, A., Costa, F.: LXGRAM – deep linguistic processing of Portuguese with HSPG. Technical report, Departement of Informatics, University of Lisbon (2005)