Extraction of Author's Definitions Using Indexed Reference Identification

Marc Bertin, Iana Atanassova and Jean-Pierre Descles
Paris-Sorbonne University
Maison de la Recherche
28 rue Serpente
75006 Paris
{marc.bertin / iana.atanassova / jean-pierre.descles}@*paris-sorbonne.fr*

Abstract

In this paper we present the implementation of definition extraction from multilingual corpora of scientific articles. We establish relations between the definitions and authors by using indexed references in the text. Our method is based on a linguistic ontology designed for this purpose. We propose two evaluations of the annotations.

Keywords

Semantic annotation, definition extraction, indexed references

1 Introduction

The use of definitions plays an important role in a number of scientific disciplines. The complexity of some domains raises the necessity to develop tools for the automatic annotation of information relevant to definitions. The growth in scientific literature production leads us to propose new tools for text navigation and quick access to the textual information.

In this paper we explore a new way to extract definitions from scientific text corpora by establishing a relation between the usage of a definition and a cited author.

In section 2, we describe the elaboration of a linguistic ontology, based on the analysis of multilingual corpora. Then, the identification of indexed references is used to establish the relations between authors.

In section 3, we explain our implementation. The goal of our system is to provide to the user the possibility to clarify a notion and its usage in a given context from a terminological or conceptual viewpoint. This means that we need to maintain the link between the extracted definitions and their contexts, in order to provide access to the argumentation in the text. The user can thus visualise the context in which the term in question has been defined. Section 4 shows the results produced by the application. In section 5 we discuss the problem of the evaluation of the semantic annotations and propose two types of evaluations: one by the precision/recall measures and another by the Cohen's weighted Kappa coefficient.

Finally, we conclude by a discussion of the perspectives for the utilisation of this tool.

2 Methodology

We propose a method for the identification of definitions and also for the identification of relations between authors. This approach allows us to associate a definition to an author and to establish a link with other texts that could interest the user. The system allows a fully automated text processing, which comprises several stages.

2.1 Protocol

Our protocol is as follows: first we carry out the identification of the sentences containing indexed references, by using regular expressions. Then, we annotate the definitions in the sentences identified in the previous stage. Finally, we extract the definitions and create indexes for the information retrieval. The results are stored in a database. Different types of visualizations and information retrieval are provided by our web-based interface.

2.2 Multilingual Corpora

We have constructed multilingual corpora, in order to create our linguistic resources organized in a linguistic ontology. The corpora comprise mainly scientific texts and articles available online. The French corpus consists of texts from several scientific reviews (Intellectica, ALSIC, TALN, IRISA) and six PhD theses from the domains of Linguistics and Computer science. The articles in English corpus are from Nature, Journal of Cell Science, Biophysical Journal, Proceedings of the National Academy of Sciences, The Journal of Cell Biology, and others.

Corpus	Texts	Sentences
French	205	119410
English	116	38378
Total	321	158788

Table 1: *Corpora*

In table 1 we present the sizes of the corpora. In order to ensure compatibility with the tools of segmentation and annotation, the corpora have been converted into text files. The sentence counts are obtained after the segmentation, which will be detailed later in the section 3.2.2.

From a legal point of view, texts can be cited freely, even if under copyright[1], provided that the following three criteria are respected. Firstly, citations must be short: our interface provides output in the form of text segments corresponding to sentences. Secondly, the purpose of extraction must be infromative, such as in the case of information retrieval. Finally, the source must be mentioned.

Moreover, we establish a relation between the definition and the cited document or author through the bibliography. This stage is important for the creation of an author network.

2.3 Definition Ontology

This section describes our linguistic approach and the construction of an ontology for the annotation of definitions. The method we present is based on enunciative discourse considerations and a corpus analysis, through which we construct an ontology by abduction.

2.3.1 Linguistic Analysis

We can examine a definition sentence by studying the relation between the *definiendum*, what is to be defined, and the *definiens*, what defines it. This linguistic study of our corpus has led us to a better understanding of the distinction between a *definition* and a *definatory characteristic*, which has been taken in consideration for the construction of our linguistic resources. We define a definatory characteristic as a sentence that gives only some essential properties of the defined object. We have distinguished three categories of definatory characteristics: *identification, determined categorization* and *pseudo-definition*. We have also considered two sub-categories of the definition: general definitions and axiomatic definitions. The full ontology that we have created contains some further sub-categorizations that are presented on figure 1.

The categorization in this linguistic ontology is based on an analysis of the types of relations. Here we will describe briefly the differences between some of the categories that we have retained.

Firstly, it must be noted that in definition sentences, apart from the relation between the definiendum and the definiens, there exists a second relation, which is between this first relation and the agent who established the definition. The presence of this agent is not always manifested in discourse and sometimes there is no actual trace. In the case when the agent is present in the text, we can speak of a *contextualised definition*, because it is often marked in the context by a deictic, which is limited to a domain or to a period in time, or else introduced by a passive construction or 'on' in French.

Secondly, we have *axiomatic definitions*, which are utterances expressing a primary truth.

Finally, there are cases where the author uses a reported definition. In these cases the enunciator can choose whether to attest the definition or not, in order to use it in the elaboration of a demonstration, or to introduce a new notion. This type of definitions takes part in the text evolution by means of modalities and we speak of *committed definition*.

The objective that we have fixed is to extract definition sentences, in which the definition is explicitly attributed to an author or another work, cited in the text. We will also call them *signed definitions*, which correspond to the category of Reported Definitions in our ontology.

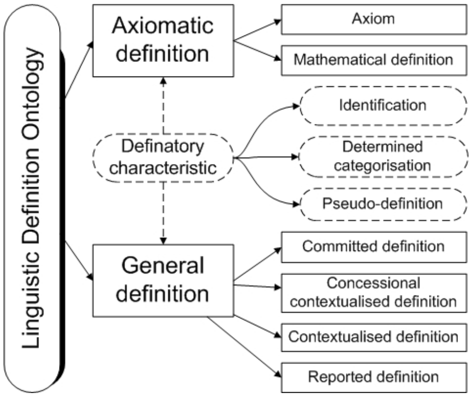

Linguistic Definition Ontology
Axiomatic definition
Axiom
Mathematical definition
Definatory characteristic
Identification
Determined categorisation
Pseudo-definition
General definition
Committed definition
Concessional contextualised definition
Contextualised definition
Reported definition

Fig. 1: *Linguistic ontology of the definition*

2.4 Indexed Reference Identification

The method that we propose is based on the indexed references in the text which point to the bibliography as define in [1]. More precisely, the indexed references allow us, in the case when we identify a definition in the research scope determined by the segmentation, to link this definition to the author cited in the text. The theoretical framework as well as the experimental procedures for the indexed reference identification are described below.

2.4.1 Bibliographic International Standards

We have considered several norms for bibliographic references, namely the norms ISO-690 and ISO 690-2, which are the international standards from the International Organization for Standardization, as well as the French norms AFNOR NF Z 44-005 and AFNOR NF Z 44-005-2.

In practice the norms are not rigorously applied by authors of scientific texts. For this reason, a method based only on the norms described above is not sufficient to carry out the text processing on a large scale. That is why, although the identification of the indexed references may seem trivial at first glance, a large number of morphological and syntactic variations must be taken into account. To illustrate this complexity, here is a list of forms that we have extracted from our corpus: *(Hoc, 1990a), (Thom, 1970), (Dingwall et al., 1995; Hartmann and Görlich, 1995), [24], Pickett-Heaps et al. (1990), (like other authors e.g. Raven, 1983), (Cwuc and SPRAGUE 1989), (18, 53, 56).*

[1] cf. CPI art L. 122-5

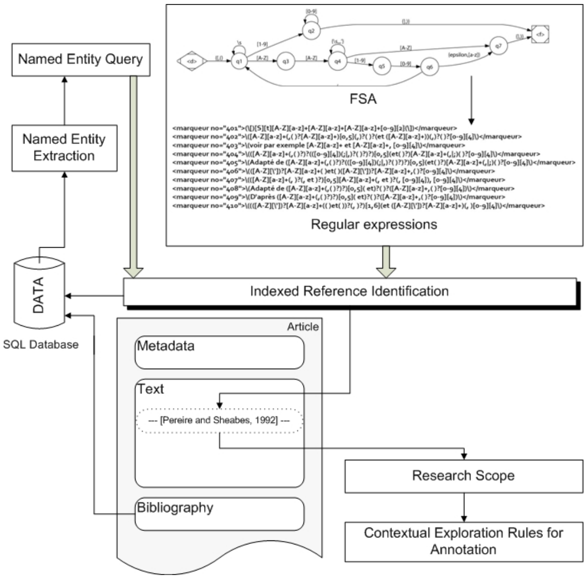
2.4.2 Finite State Automata

Although the identification of indexed references has been approached by Citeseer, we have developed our own module. In fact, at the beginning of our work such modules were not available[2]. The specificity of our module is its capacity to identify also the author names which can appear in the forms. The classification that we use has been published in [2].

We identify automatically the indexed references by the use of Finite State Automata (FSA). For this we have to take into consideration the norms established on the one hand by the practices proper to authors and on the other hand by the different domains. That is why in order to create robust FSA, different corrections had to be made to take into consideration the different customs in writing indexed references. The annotation platform we have chosen takes as input rules based on lists of regular expressions. Therefore, for the implementation of this methodology, we have converted the FSA into regular expressions.

2.4.3 Identification of Known Named Entities

The identification of an indexed reference can become difficult because of the presence of named entities in the reference. The named entities are the more complex part of the indexed references and introduce considerable complications in the FSA due the various name morphologies in different languages.



```
<marqueur no="401">(\[)[S][t][A-Z][a-z]+[A-Z][a-z]+[A-Z][a-z]+[0-9][2](\])</marqueur>
<marqueur no="402">\(([A-Z][a-z]+(,( )?[A-Z][a-z]+){0,5}(,)?( )?(et ([A-Z][a-z]+))(,)?( )?[0-9]{4}\)</marqueur>
<marqueur no="403">\(voir par exemple [A-Z][a-z]+ et [A-Z][a-z]+, [0-9]{4}\)</marqueur>
<marqueur no="404">\(([A-Z][a-z]+(,( )?)?(([0-9]{4})(;|,)?( )?)?){0,5}(et( )?)[A-Z][a-z]+(,|;)( )?[0-9]{4}\)</marqueur>
<marqueur no="405">\(Adapté de ([A-Z][a-z]+(,( )?)?(([0-9]{4})(;|,)?( )?)?){0,5}(et( )?)[A-Z][a-z]+(,|;)( )?[0-9]{4}\)</marqueur>
<marqueur no="406">\(([A-Z][\'])?[A-Z][a-z]+( )et( )([A-Z][\'])?[A-Z][a-z]+,( )?[0-9]{4}\)</marqueur>
<marqueur no="407">\((([A-Z][a-z]+(, )?(, et )?){0,5}[A-Z][a-z]+(, et )?(, [0-9]{4}), [0-9]{4}\)</marqueur>
<marqueur no="408">\(Adapté de ([A-Z][a-z]+(,( )?)?){0,5}( et)?( )?([A-Z][a-z]+,( )?[0-9]{4})\)</marqueur>
<marqueur no="409">\(D'après ([A-Z][a-z]+(,( )?)?){0,5}( et)?( )?([A-Z][a-z]+,( )?[0-9]{4})\)</marqueur>
<marqueur no="410">\((([A-Z][\'])?[A-Z][a-z]+(( )et( ))?(, )?){1,6}(et ([A-Z][\'])?[A-Z][a-z]+)(, )[0-9]{4}\)</marqueur>
```

Fig. 2: *Indexed reference identification with Named Entity extraction*

Figure 2 describes the implementation of a solution which improves the system performance, through the utilization of author names, that have been already identified by the system, as part of the regular expressions. In fact, by using some data already existing in the bibliographic databases, we can generate certain forms and limit the noise in the more complex forms. Moreover, this approach permits some extensions of the method: we can consider new sentences for the signed definition extraction through matching not only indexed references, but also author names in the text that can be cited without bibliographic links.

3 Semantic Annotation

3.1 Annotation Tools

Definition identification is traditionally based on pattern matching, as for example in [11]. These approaches are used for the development of platforms such as TerminoWeb[3] of the National Research Council Canada.

Different approaches are possible for the semantic annotation. Among the tools that we have considered we can cite the GATE[4] platform [12] based on machine learning algorithms, generally used with JAPE [4], and the work of Xerox Concept-matching, based on XIP [10], a morphosyntactic analyser

In our work we have used the Excom platform [6], which implements the Contextual Exploration method [5]. This is a decision-making procedure, presented in the form of a set of rules and linguistic markers that trigger the application of the rules. They are applied to the segments containing indicators. The indicators are linguistic units that carry the semantic meaning of the categories for annotation. After the initial identification of the indicators in the text, the rules carry out the localisation of complementary linguistic clues which are co-present in the context of the indicators. After the verification of the presence or absence of the linguistic clues, the rules attribute a semantic annotation to the segment.

In our approach we consider as a working hypothesis the fact that in a scientific article the information related to signed definition can be found in the textual space close to an indexed reference, and more specifically in the same sentence. Our aim is to limit as much as possible the noise in the annotations, to be able to obtain foolproof matching between authors and definitions.

As we need to be able to disambiguate the linguistic forms according to the context, in order to limit the noise as much as possible and to deal with polysemy, we have chosen the Contextual Exploration framework as more adapted to our approach. For this reason, we have used the Excom annotation system[5].

3.2 System Overview

Here we describe in detail the main stages in the text processing, that we have divided into a four-stage process. The overall system pipeline is presented on figure 3.

[2] The source code of the Citeseer module has been recently published on http://sourceforge.net/projects/citeseerx/. The identification module is based on a Perl module from CPAN, which parses documents using regular expressions.

[3] http://termino.iit.nrc.ca/

[4] http://gate.ac.uk

[5] http://www.excom.fr

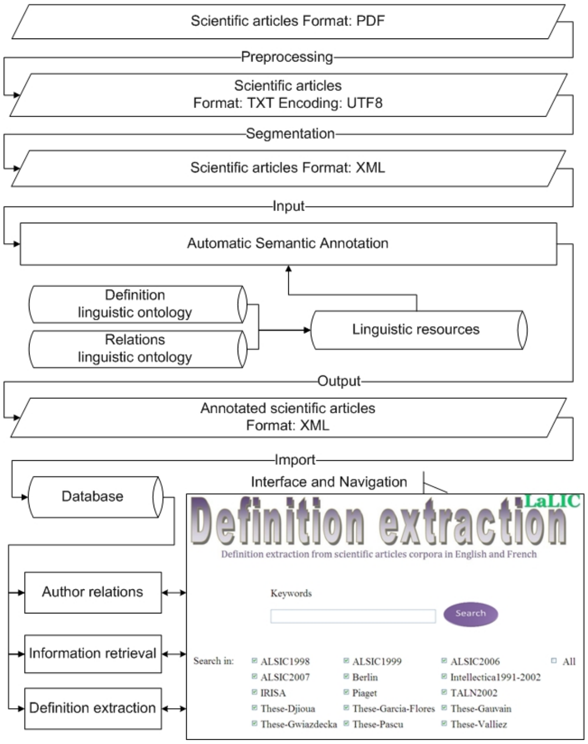


Fig. 3: *Stages in the automatic processing*

3.2.1 Preprocessing

The initial corpora being in PDF format, in the preprocessing stage the files are converted into text format. This is necessary because the next stages in our processing, namely the segmentation and annotation, need full text access to the corpora. The converted files are in the UTF8 encoding which permits the processing of different natural languages.

3.2.2 Segmentation

In the second stage, we segment the corpora into paragraphs and sentences, in order to prepare the input for the annotation module. The quality of the segmentation is important for the overall system performance, as the segmentation provides the text elements to be annotated. The segmentation is carried out by the SegaTex module [7] which we have chosen for the reliability of its results and its capacity to process texts in both English and French. This module takes as input text files and returns the segmented files in the Docbook XML format, with paragraph and sentence elements. This format is compatible with the annotation module.

3.2.3 Semantic Annotation

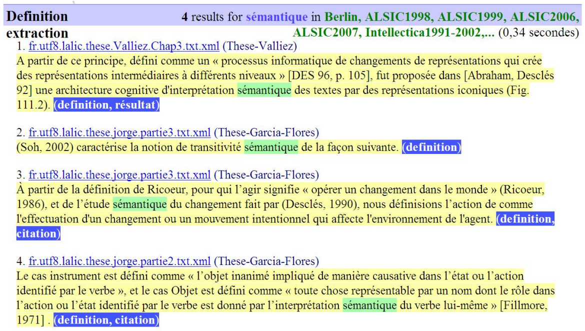
We will therefore analyze the discourse forms which are to be found in the text space close to the indexed reference. The semantic annotation is carried out by the Excom module which takes as input the segmented files as well as the linguistic resources that we have constructed. According to our protocol, we use two types of linguistic resources: regular expressions for the identification of the indexed references and contextual exploration rules for the annotation of definitions related to the ontology presented above. In the output, the identified indexed references are present as new elements and the annotations of the definitions are added as attributes to the relevant XML sentence elements in the corpora.

3.2.4 Interface and Navigation

We have developed a web-based graphical user interface, using the technology Apache/PHP/MySQL. The annotated corpora are imported into a database designed for this purpose, which contains the annotations and the text segments, as well as meta-data related to the files. The Definition Extraction Interface (DEI) permits the visualization of the information in the database, and different other functionalities that we will describe here.

The most important functionality of the DEI is the information retrieval among the annotated sentences. In the initial screen, shown on figure 3, the user can formulate a query by using keywords[6] and eventually restricting the search to a specific set of corpora. The results are presented in the form of a list of sentences, together with the annotations and links to the initial texts.

3.3 Results



Definition extraction **4** results for **sémantique** in **Berlin, ALSIC1998, ALSIC1999, ALSIC2006, ALSIC2007, Intellectica1991-2002,...** (0,34 secondes)

1. fr.utf8.lalic.these.Valliez.Chap3.txt.xml (These-Valliez)
A partir de ce principe, défini comme un « processus informatique de changements de représentations qui crée des représentations intermédiaires à différents niveaux » [DES 96, p. 105], fut proposée dans [Abraham, Desclés 92] une architecture cognitive d'interprétation sémantique des textes par des représentations iconiques (Fig. 111.2). **(definition, résultat)**

2. fr.utf8.lalic.these.jorge.partie3.txt.xml (These-Garcia-Flores)
(Soh, 2002) caractérise la notion de transitivité sémantique de la façon suivante. **(definition)**

3. fr.utf8.lalic.these.jorge.partie3.txt.xml (These-Garcia-Flores)
À partir de la définition de Ricoeur, pour qui l'agir signifie « opérer un changement dans le monde » (Ricoeur, 1986), et de l'étude sémantique du changement fait par (Desclés, 1990), nous définisions l'action de comme l'effectuation d'un changement ou un mouvement intentionnel qui affecte l'environnement de l'agent. **(definition, citation)**

4. fr.utf8.lalic.these.jorge.partie2.txt.xml (These-Garcia-Flores)
Le cas instrument est défini comme « l'objet inanimé impliqué de manière causative dans l'état ou l'action identifié par le verbe », et le cas Objet est défini comme « toute chose représentable par un nom dont le rôle dans l'action ou l'état identifié par le verbe est donné par l'interprétation sémantique du verbe lui-même » [Fillmore, 1971] . **(definition, citation)**

Fig. 4: *Search results*

Figure 4 presents the results from the French corpus for the keyword *"sémantique"*. The following excerpts were extracted from the English corpus:

1. Another homolog to RCCI has been identified in S. cerevisiae, called either SRMl (Cwuc and SPRAGUE 1989) or PRP20 (AEBI et al. 1990; FLEISCHMANN et al. 1991).

2. Silica polymerization occurs within an organelle called the silica deposition vesicle, bounded by a membrane called the silicalemma (18, 53, 56).

We can see that the first and the second examples are general reported definitions.

[6] Boolean expressions (AND, OR, NOT) with parentheses and quotation marks in queries are also implemented.

4 Evaluation and Discussion

4.1 Precision and Recall Measures

The first evaluation consists in measuring the accuracy of the retained indexed references, which have been identified automatically by the regular expressions. We have used the precision/recall measures [9] which determine the capacity of the system to correctly identify textual segments containing indexed references. Table 2 presents the number and the percentage of the sentences containing indexed references in each corpus. We can see that around 5% of the sentences have been extracted.

Corpus	Sentences	Annotated Sentences	Percentage
French	119410	5976	5,00 %
English	38378	1743	4,54 %
Total	157788	7719	4,89 %

Table 2: *Annotated sentences*

We have carried out the evaluation on a set of 500 sentences extracted randomly from our corpora. In table 3 we present the results obtained by this evaluation.

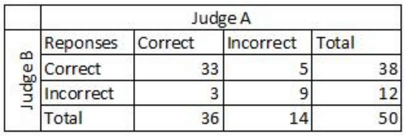
Recall	Precision	F-measure
0,911%	0,989%	0,9483

Table 3: *Evaluation of the Indexed References*

We consider that these results are satisfactory. It must be noted that there is very little noise which means that almost all of the identified indexed references are valid. On the other hand, the value of the recall is also very high. The several percents of indexed references not identified by the system are due to the various orthography rules for the names in different languages, as well as the presence of commentaries in the indexed reference itself.

4.2 Cohen's Weighted Kappa

The problem we have to consider is how to evaluate the semantic annotation which is by definition qualitative in nature. The test Kappa (K) proposed by Cohen[3] and developed by [8] provides a method to measure numerically the agreement between two or more observers or methods in the case when the judgments are qualitative in nature. We have adopted this method for the second stage of our evaluation.



		Judge A		
	Reponses	Correct	Incorrect	Total
Judge B	Correct	33	5	38
	Incorrect	3	9	12
	Total	36	14	50

Table 4: *Evaluation Results*

In order to carry out the test, we have constituted a base of annotated text segments and these segments have been evaluated independently by two human judges. The judges had to classify the segments into two categories: correct and incorrect. We have used a set of 50 sentences for this evaluation. Table 4 presents the results. For the Cohen's Kappa we obtain: $\kappa = 0,6515$, and therefore we have a subtantial agreement, according to the interpretation in [8].

5 Conclusion and Future Work

We note that according to the evaluation the system gives satisfactory results, which validates the linguistic resources and the definition ontology in our approach. Throughout the process of annotation and exploitation of the results we maintain the links between the extracted sentences and the original texts which makes possible the visualization of the context of each definition. The evaluations confirm the relevance of this application. However, we are not yet able to predict the result on a larger scale and on corpora in other domains. In the future we will extend this approach to the processing of bigger corpora in English and in French as well as other natural languages.

References

[1] M. Bertin. Categorizations and annotations of citation in research evaluation. In *FLAIRS 2008, Coconut Grove, Florida*, Coconut Grove, Florida, May 2008.

[2] M. Bertin, I. Atanassova, and J.-P. Desclés. Automatic analysis of author judgment in scientific articles based on semantic annotation. In *22nd International Florida Artificial Intelligence, Research Society Conference*, Sanibel Island, Florida, 19-21 mai 2009.

[3] J. Cohen. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.*, 20:27–46, 1960.

[4] H. Cunningham, D. Maynard, and V. Tablan. Jape–a java annotation patterns engine. *Advances in Text Processing, TIPSTER Program Phase II*, pages 185–189, 2000.

[5] J.-P. Desclés. Contextual exploration processing for discourse automatic annotations of texts. *FLAIRS 2006, Florida. Invited Speaker*, 2006.

[6] B. Djioua, F. J. Garcia, A. Blais, J.-P. Desclés, G. Guibert, A. Jackiewicz, F. L. Priol, L. Nait-Baha, and B. Sauzay. Excom: an automatic annotation engine for semantic information. *FLAIRS 2006, Florida*, pages 285–290, 2006.

[7] M. Ghassan. La segmentation de textes par exploration contextuelle automatique, présentation du module segatex. *ISLsp, Inscription Spatiale du Langage : structure et processus IRIT, Université Paul Sabatier, Toulouse*, 2002.

[8] J. Landis and G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, 1977.

[9] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, 1979.

[10] A. Sandor, A. Kaplan, and G. Rondeau. Discourse and citation analysis with concept-matching. *International Symposium : Discourse and document (ISDD), Caen, France*, 2006.

[11] G. Sierra, R. Alarcon, C. Aguilar, and C. Bach. Definitional verbal patterns for semantic relation extraction. *Terminology*, 14(1):74–98, 2008.

[12] V. Tablan, C. Ursu, K. Bontcheva, H. Cunningham, D. Maynard, O. Hamza, T. McEnery, P. Baker, and M. Leisher. A unicode-based environment for creation and use of language resources. In *Proceedings of 3rd Language Resources and Evaluation Conferenc*, 2002.