Ontology Acquisition Process: A Framework for Experimenting with different NLP Techniques

Ricardo Gacitua[1], Pete Sawyer[1], Scott Piao[2], Paul Rayson[1]

[1]Computing Department, Lancaster University, UK

[2]School of Computer Science, Manchester University, UK

[r.gacitua, sawyer, p.rayson] @lancs.ac.uk, scott.piao@manchester.ac.uk

Abstract

Since the manual construction of ontologies is time-consuming and expensive, an increasing number of initiatives to ease the construction by automatic or semi-automatic means have been published. Most initiatives combine a certain level of NLP techniques with machine learning approaches to find concepts and relationships. However, a challenging issue is to quantitatively evaluate the usefulness or accuracy of the techniques and combinations of techniques when applied to ontology learning. We are developing a framework for acquiring an ontology from a large collection of domain texts. This framework provides support for evaluating different NLP and machine learning techniques when they are applied to ontology learning. Our initial experiment supports our assumptions on the usefulness of our approach.

1. Introduction

The rapid development of the Internet and computer technology make us live in an “information overload” world. Appropriate access to and digestion of information is therefore essential. In most domains knowledge about domain entities and their properties and relationships is embodied in documents with varying degrees of explicitness and precision.

Because knowledge-objects of a given domain are expressed and conveyed in texts using domain-specific terminology, it is reasonable to think that the mining and extracting of terminology will lead to a certain domain representation model such as an ontology [1]. Despite the various tools in existence for encoding information in ontology languages [2], human domain experts have to do the work of deriving the classifications and relationships to be encoded. However, such manual work is tedious, time consuming and error-prone, even with the assistance of computers. In this context, a number of Natural language Processing (NLP) and text mining techniques have shown potential for partially dealing with a synthesis process. For example, Cimiano et al. [3] use statistical analysis to extract terms and taxonomy. Likewise, Reinberg et al. (2004)[4] use shallow linguistic parsing for concept formation and relation extraction. However, an ongoing challenge [5] is to evaluate the accuracy and efficiency of the techniques used to support large scale ontology extraction for real-world applications.

We propose that, in order to evaluate their effectiveness, it is necessary to determine the techniques providing optimal performances for the ontology process. However, it is not a trivial task to evaluate the efficiency of the techniques for ontology learning. Reinberg and Spyns [6] point out that “*To our knowledge no comparative study has been published yet on the efficiency and effectiveness of the various techniques applied to ontology learning*”. In addition, Aussenac-Gilles[7] indicate that: “*A listing of existing techniques, their properties and possible combinations would be a useful guideline to progress toward tool or technique combination into specific processes. This is one of the research challenges of the Semantic Web for the years to come*”.

Our work focuses on integrating a number of NLP and machine learning techniques to determine the best combination for the semi-automatic extraction of domain concepts and their encoding in the OWL ontology language (semi-automatic ontology generation). For this purpose, we are developing a framework and an integrated tool-suite based on an architecture that integrates existing linguistic tools developed at Lancaster University. This will provide a workbench for information extraction, which is integrated into an existing open source ontology editor, supplying ontology engineers with a coordinated tool for knowledge objects extraction and ontology modelling, as well as testing different techniques.

We aim to exploit NLP tools and techniques which have been deployed by the Computing Department at Lancaster University to assist ontology engineering. In particular, we use WMatrix [8]. It is a software application for corpus analysis and comparison. This tool provides a Web interface for syntactic and semantic corpus annotation tools, and implements standard corpus linguistic methodologies such as frequency lists and concordances.

Our research project addresses the important challenges of ontology engineering, covering the issue of validating innovative NLP and machine learning approaches as a scientific means to capture knowledge-objects contained in domain-specific texts and rapidly organises them into domain ontologies to be used in third-party applications.

In this paper we present the results achieved so far:

(i) The definition of a framework to support the semi-automatic ontology acquisition process.
(ii) A prototype workbench.
(iii) A preliminary experiment.

This work is part of a larger project to build ontologies semi-automatically by processing a collection of domain texts. Future projects include both semi-automatic construction of a concepts hierarchy as a first step to ontology learning, and integration with an ontology editor. The aim of the project is not to develop new NLP techniques. Rather, the work involves the innovative adaptation, integration and application of existing NLP techniques in order to test them and validate their utility.

The remainder of our paper is organized as follows: - we begin by introducing related works; then, we characterize the main parts of the framework and, we present a brief snapshot of our workbench; next, we present experiments using a set of linguistic techniques; finally, we discuss the experiments results of our experiments and present the conclusions.

2. Related Work

In recent years, a number of frameworks that support ontology learning processes have been reported. They implement several techniques from different fields such as knowledge acquisition, machine learning, information retrieval, natural language processing, artificial intelligence reasoning and database management, as shown in the following works:

- ASIUM [9, 10] learns verb frames and taxonomic knowledge, based on statistical analysis of syntactic parsing of French texts,
- KAON-TextToOnto [11, 12] learns concepts and relations from unstructured, semi-structured, and structured data, using a multi-strategy method, a combination of association rules, formal concept analysis and clustering,
- Ontolearn [5, 13] learns by interpretation of compounds,
- OntoLT [14] learns concepts by term extraction using statistical methods and definition of linguistic patterns, as well as mapping to ontological structures. OntoLT includes a statistical analysis functionality to lexically constrain a mapping rule towards linguistic entities that are relevant for the domain. It computes a relevance score for each linguistic entity by comparison of its frequency in a domain corpus with that of its frequency in a reference corpus. Linguistic entities that are more specific for the domain corpus will receive a higher score.
- DODDLE II [15] learns taxonomic and non-taxonomic relations using co-ocurrence analysis, exploiting a machine readable dictionary (WordNet) and domain-specific text.
- WEB->KB [16, 17] combines Bayesian learning and FOL rule learning methods to learn instances and rules for instance extraction from World Wide Web documents.

All of them combine some linguistic analysis methods with machine learning algorithms in order to find potentially interesting concepts and relations between them. However, none provides any mechanism for carrying out experiments with a combination of the techniques or for including a new one.

There has been a considerable diversity of theories about the usefulness of NLP components and the information to be provided as the input to the ontology acquisition process. This diversity leads to researchers in this area building their own systems or using their own terminology to define specific aspects of a topic. On the other hand, tools and infrastructures for ontology acquisition, NLP and Knowledge Management have mostly remained independent of each other, although in fact they share a number of components. There is little reported work on tool integration.

3. The Ontology Framework

In this section, we focus on describing our ontology acquisition framework for the semi-automatic ontology acquisition process.

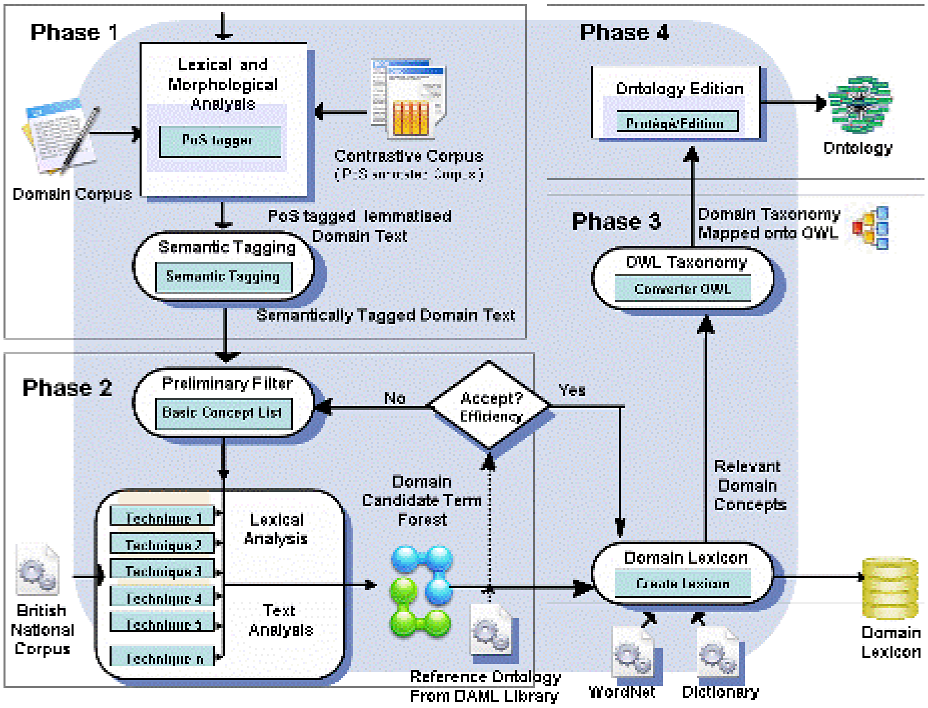


Figure 1: The Ontology Acquisition Framework.

3.1. Phases of the Ontology Framework

The workflow of our ontology framework proceeds through the stages of semi-automatic abstraction and classification of domain concepts, encoding them in the OWL ontology language [20], and editing them using an enhanced version of an existing editor - Protégé[2]. This set of tools provides ontology engineers with a coordinated and integrated workbench for extracting terms and modelling ontology. In addition, our framework uses external resources available from the Internet such as WordNet, dictionaries etc. There are four main phases of process, as shown in Fig.1. Below we provide detailed descriptions of these phases.

Phase 1: Part-of-Speech (POS) and Semantic annotation of corpus

Domain texts are tagged morpho-syntactically and semantically using Wmatrix. The system assigns a semantic category to each word employing a comprehensive semantic category scheme called The UCREL Semantic Analysis System (USAS). This scheme has a hierarchical semantic taxonomy containing 21 major discourse fields and 232 fine-grained semantic fields. Also, USAS combines several resources including the CLAWS POS tagger [21] which is used to assign POS tags to words.

Phase 2: Extraction of concepts

The domain terminology is extracted from the tagged domain corpus by identifying a list of domain candidate terms (Domain Candidate Term Forest). In this phase the system provides a set of statistical and linguistic techniques which an ontology engineer can combine for identifying candidate terms with high precision. Where a domain ontology exists in The DARPA Agent Markup Language (DAML) Library, it can be used as a reference and to calculate precision and recall. We initially plan to apply the framework and workbench to a set of domain documents for which domain ontology already exists.

Phase 3: Domain Ontology Construction

Concepts extracted during the previous phase are then added to a bootstrap ontology. We assume that a hierarchical classification of terms, rather than a strict OWL-like ontology, will be sufficient for the first stage of our project. In this phase, a domain lexicon is built. Definitions for each concept are extracted from several on-line sources automatically, such as WordNet and online dictionaries. In the case of a concept definition not being found, domain experts can supply one.

Phase 4: Domain Ontology Edition

In the final phase, the bootstrap ontology is turned into OWL. Then it is processed using an ontology editor to manage the versioning of the domain ontology and modify or improve it. For the editor, we will use Protégé which is open source, knowledge-based, standalone software with an extensible architecture.

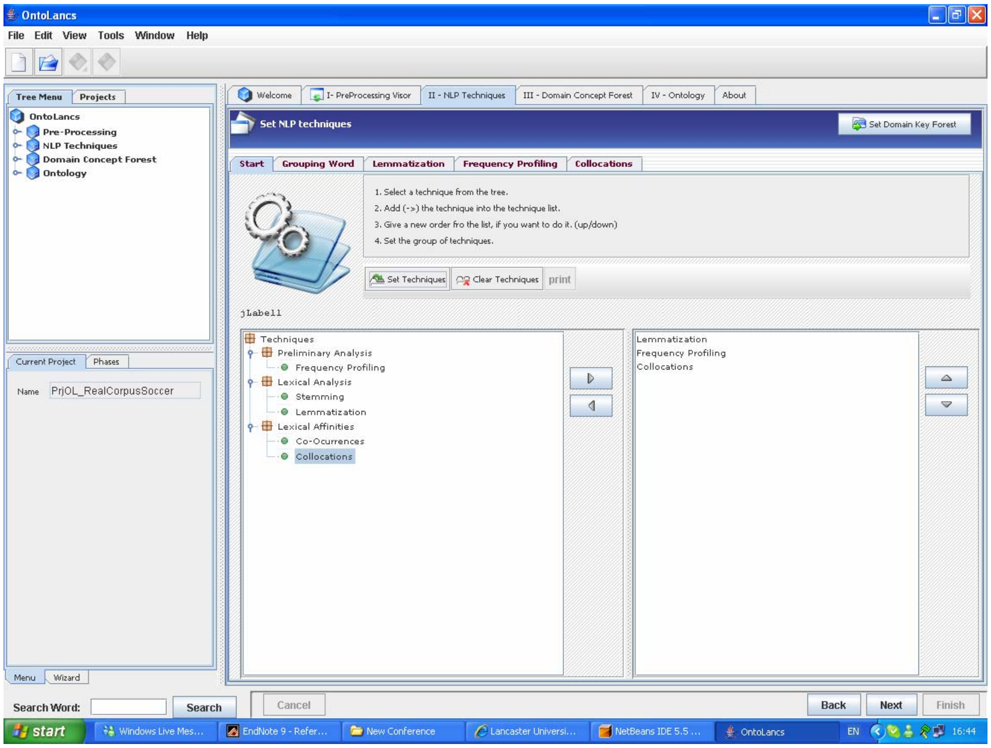
Our framework provides new functionalities in comparison with other similar work. Primarily it facilitates experiments with different NLP techniques in order to assess their efficiency and effectiveness, including the performance of various combinations of NLP and machine learning techniques. All such functions are being built into a prototype workbench to evaluate and refine existing techniques using a range of domain document corpora.

3.2. An Integrated Ontology Workbench

This section provides a brief description of the implementation of the first phase in the prototype workbench. Our framework is designed to include a set of NLP and machine learning techniques to enable its enhancement by including new techniques in the future (see figure 2). Each of them can be selected or left out to make a combination of techniques. Like a pipeline, the output of one technique will be the input of another technique.

3.2.1. Phase 1 - Part-of-Speech (POS) and Semantic annotation of corpus: In our own case, we used a Java API library (Jmatrix) to connect our workbench to Wmatrix in order to get POS tags and semantic tags for each word. The integration between Wmatrix and the ontology workbench provides a platform for dealing with the scalability problem. Running in a powerful server, Wmatrix is capable of processing a large volume of corpora. Furthermore, the workbench has pre-loaded the BNC corpus - a balanced synchronic text corpus containing 100 million words with morphosyntactic annotation. In order to identify a preliminary set of concepts the workbench provides functions to analyze the corpus and filter the candidates using POS tags and absolute frequency as a preliminary filter. Figure 2 shows the GUI of the workbench.

Phase 2 - Extraction of Candidate terms: In this case, a first linguistic technique is provided. It comprises 2 basic filters: (a) **Filter - *Group by POS, which provides an option to select a set of POS,*** tag categories and filter the list of terms. (b) **Filter - Absolute frequency**, which provides an option to filter the list of terms by frequency ranges. Since the experiments are the preliminary baseline, they do not consider human intervention, although we claim the necessity of human supervision to improve the efficiency of the ontology acquisition



Name PrjOL_RealCorpusSoccer

1. Select a technique from the tree.
2. Add (->) the technique into the technique list.
3. Give a new order fro the list, if you want to do it. (up/down)
4. Set the group of techniques.

Figure 2: Combining Techniques – OntoLancs Workbench

process. So the results are gathered from the workbench automatically.

4. Experiments

This section describes a first set of experiments. Our preliminary set of experiments consists in applying the first linguistic technique - Group & Filter by POS on the set of candidate terms. For this, we built a football corpus which comprises **102.543 words**.

As a first step we formed the collected information into 10 groups: culture, formation, glossary, help, game law, main topics, positions, some explanation, tactics and history. Each group was turned into one text file, thus our corpus comprises 10 files. All documents were gathered by running a Google query “Football Game”. Then we selected those written by FIFA (Fédération Internationale de Football Association) and published in football web sites.

In order to evaluate our extraction process we selected an ontology Soccer[1] from the DAML Library which has **199 classes.** That ontology is used to annotate videos in order to produce personalized summaries of soccer matches. Although, we cannot ensure the conceptual correctness of the DAML reference ontology and the match with the application context of our domain corpus, we assumed as a preliminary premise that the DAML reference ontology is the right one to assess our concept extraction process.

First, we excluded a pre-defined list of stop words, which are not useful for identifying concepts, and then we grouped the initial list of candidate terms using the different categories:

- **Grouped by POS tags**. In this case, we used 3 sorts of word grouping:
    - (a) *Using specific POS Tags.* For instance: **Kick _VV0** (base form or lexical verb) is considered different from **Kick_VVI** (infinitive),
    - (b) *Using a generic POS Tag*. In this case, we used a generic POS Tag. For instance **Kick _VV0** and **Kick** ***VVI*** *are turned into* ***Kick*”verb”,**
    - (c) *Not using a POS Tag*. In this case, we used only a word with a generic category: “any”. For instance, **Kick_noun** and **Kick_verb** are turned into **Kick_any.**

Finally, we checked how many of those candidates terms appear in the DAML reference ontology.

In order to evaluate quantitatively the results of this process we used the precision and recall measures previously defined to measure either information retrieval results, or information extraction results. In our case, we applied those measures to the tagged set of candidate terms with regards to the classes in the DAML reference ontology. Hence, we obtained the two following adapted measures:

- **Precision** measures the number of classes of the ontology, which were matched by a candidate term, divided by the number of the candidate terms.
- **Recall** measures the number of classes of the ontology, which were matched by a candidate term divided by the number of ontology classes.

The results of the first evaluation, after applying grouping by POS, show low values of recall and precision. This is a consequence of the fact that we used an unsupervised method and applied a limited number of techniques to identify domain concepts.

In the above experiments, we observed the results on the word grouping table (see Table 1). Although we applied one linguistic technique only, we collected a reasonable number of matched ontology classes. In addition, the same result was obtained as the baseline by using only 3 POS categories: nouns, verbs and adjectives. We can conclude that other POS categories are not useful in

TABLE I

PRECISION AND RECALL – WORD GROUPING

	SPECIFIC POS		CATEGORY POS		ANY	
	Recall	Precision	Recall	Precision	Recall	Precision
No Filter	*47.8*	*1.4*	*47.8*	*1.4*	*47.8*	*1.4*
Nouns	42.7	2.3	42.7	2.3	47.8	1.4
Nouns + Verbs	46.6	1.7	47.8	1.4	47.8	1.4
Nouns + Adjectives	44.5	1.9	47.8	1.4	47.8	1.4
Nouns+ Adjectives+ Verbs	47.8	1.5	47.8	1.5	47.8	1.4

Table1. Word Grouping – recall and precision values obtained after applying the word grouping technique on the football corpus, and matched against the soccer ontology from DAML Library.

[1] (Hhttp://www.lgi2p.ema.fr/~ranwezs/ontologies/soccerV2.0.damlH),

identifying domain concepts by using nouns, adjectives and verbs.

It is obvious that the metrics precision and recall on these experiments are low for several reasons. For instance: (1) a real ontology has classes defined by multiword. Filter by POS tags does not consider multiwords, (2) the preliminary list of concepts tagged by POS has several words that are general concepts and those should not be considered as domain concepts, for instance: adjectives. Thus, values of recall and precision will become higher whether new techniques to identify multiwords are included, and also human supervision to filter general concepts is provided.

5. Conclusions and Further Work

In this paper we have described an early project which proposes a new ontology framework. This framework allows for experimentation with individual and/or combinations of techniques for the ontology acquisition process. An ontology engineer can decide what techniques or combinations of them will be used to extract concepts and turn them into an ontology. Our research project addresses an important challenge of ontology research, i.e., how to validate NLP and machine learning for the purpose of capturing knowledge-objects contained in domain-specific texts. Then, the rapid organization of the candidate objects into a domain ontology. Our initial experiment supports our assumption about the usefulness of our approach that is evaluating the effectiveness of the techniques for ontology learning acquisition. The availability of linguistic tools integrated into a practical ontology engineering process can potentially aid the rapid development of domain ontologies.

6. References

[1] A. Maedche and S. Staab, “Mining ontologies from text,” *Knowledge Engineering and Knowledge Management, Proceedings*, vol. 1937, pp. 189-202, 2000.

[2] N. F. Noy, R. W. Fergerson, and M. A. Musen, “The knowledge model of Protege-2000: Combining interoperability and flexibility,” presented at 12th International Conference on Knowledge Engineering and Knowledge Management (EKAW’2000), Juan-les-Pins, France, 2000.

[3] P. Cimiano, A. Hotho, and S. Staab, “Learning concept hierarchies from text corpora using formal concept analysis” *Journal of Artificial Intelligence research*, vol. 24, pp. 305-339, 2005.

[4] M. Reinberger, P. Spyns, A. Pretorius, and W. Daelemans, “ Automatic initiation of an ontology,,” in *On the Move to Meaningful Internet Systems*, Z. T. e. a. e. in R. Meersman, Ed.: Springer, LNC 3290, 2004, pp. 600-617.

[5] P. Velardi, R. Navigli, A. Cucchiarelli, and F. Neri, “Evaluation of OntoLearn, a methodology for Automatic Learning of Ontologies,” in *Ontology Learning from Text: Methods, Evaluation and Applications Series information for Frontiers in Artificial Intelligence and Applications, IOS Press*, P. C. P. Buitelaar, and B. Magnini, Eds., Ed.: IOS press, 2005.

[6] M. L. Reinberger and P. Spyns, “Unsupervised text Mining for the learning of DOGMA-inspired Ontologies.,” in *Ontologies Learning from Text: methods, Evaluation and Applications, Advances in Artificial Intelligence*, vol. vol. 24,, P. Buitelaar, Cimiano P., Magnini B. (eds.), Ed. Amsterdam: IOS Press, 2005, pp. pages 305-339.

[7] N. Aussenac-Gilles, “Supervised text analysis for ontology and terminology engineering,” presented at Proceedings of the Dagstuhl Seminar on Machine Learning for the Semantic Web, 2005.

[8] P. Sawyer, P. Rayson, and K. Cosh, “Shallow Knowledge as an Aid to Deep Understanding in Early-Phase Requirements Engineering,” *IEEE Transactions on Software Engineering* 2005.

[9] D. Faure and A. Nedellec, “Knowledge acquisition of predicate argument structures from technical texts using machine learning: The system ASIUM,” presented at Proceeding of the 11th European Workshop (EKW’99), 1999.

[10] D. Faure and T. Poibeau, “First Experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX,” presented at Proceeding of the Worskhop on Ontology Learning, 14th European Conference on Artificial Intelligence ECAI’00, Berlin, Germany, 2000.

[11] A. Maedche and S. Staab, “Ontology learning for the Semantic Web”, *IEEE Intelligent Systems & Their Applications*, vol. 16, pp. 72-79, 2001.

[12] A. Maedche and R. Volz, “The text-To-Onto Ontology Extraction and Maintenance Environment.,” presented at Proceeding of the ICDM Workshop on integrating data mining and knowledge management, San Jose, California, 2001.

[13] P. Velardi, R. Navigli, and M. Missikof, “Integrated Approach to Web Ontology Learning and Engineering,” *IEEE Computer* vol. 35, 2002.

[14] P. Buitelaar and M. Sintek, “OntoLT 1.0: Middleware for Ontology Extraction from Text,” presented at In: Proceeding. Of the Demo Session at the International Semantic Web Conference (ISWC), 2004.

[15] T. Yamaguchi, “Acquiring conceptual Relations from domain-specific Texts”, presented at Proceedings of the IJCAI

2001 Workshop on Ontology Learning (OL'2001), Seattle, USA., 2001.

[16] M. Craven, D. DiPasquo, F. D., A. McCallum, T. Mitchell, K. Nigam, and S. Slattery, "Learning to extract Symbolic Knowledge bases from the Wordl Wide Web," presented at AAAI'98, 1998.

[17] M. Craven, D. DiPasquo, F. D., A. McCallum, T. Mitchell, K. Nigam, and S. Slattery, "Learning to construct knowledges bases from the World Wide Web," *Artificial Intelligence*, vol. 118, pp. 69-113, 2000.

[18] D. Archer, P. Rayson, S. Piao, and T. McEnery, "Comparing the UCREL Semantic Annotation Scheme with Lexicographical Taxonomies.," presented at Proceedings of the 11th EURALEX (European Association for Lexicography) International Congress (Euralex 2004), , Lorient, France., 2004.

[19] P. Sawyer, P. Rayson, and K. Cosh, "Shallow Knowledge as an Aid to Deep Understanding in Early-Phase Requirements Engineering," *IEEE Transactions on Software Engineering*, 2005.

[20] M. Dean, D. Connolly, F. van Harmelen, J. Hendler, I. Horrocks, D. McGuinness, P. Pates-Schneider, and L. Stein, "OWL Web Ontology Language 1.0 Reference. W3C Working Draft.", 2002.

[21] R. Garside and N. Smith, "A hybrid grammatical tagger: CLAWS4 " in *Corpus Annotation: Linguistic Information from Computer Text Corpora*, R. in Garside, Leech, G., and McEnery, A., Ed. London: Longman, 1997, pp. pp. 102-121.