Design and implementation of the online ILSP Greek Corpus

Nick Hatzigeorgiu, Maria Gavrilidou, Stelios Piperidis, George Carayannis, Anastasia Papakostopoulou, Athanassia Spiliotopoulou, Anna Vacalopoulou, Penny Labropoulou, Elena Mantzari, Harris Papageorgiou, Iason Demiros

Institute for Language and Speech Processing
Epidavrou & Artemidos 6, 151 25 Maroussi, Greece
{nikos, maria, spip, gcara, natassa, aspil, avacalop, penny, elena, xaris, iason}@ilsp.gr

Abstract

This paper presents the Hellenic National (HNC), which is the corpus of Modern Greek developed by the Institute for Language and Speech Processing (ILSP). The presentation describes all stages of the creation of the corpus: collection of the material, tagging and tokenizing, construction of the database and the online implementation which aims at rendering the corpus accessible over Internet to the research community.

1. Introduction

The ILSP HNC is the first widely available corpus of the Modern Greek language. It consists of a database of written Greek texts containing over 24 million words, classified and marked-up according to the PAROLE standards. The texts belong to the categories of Books (15.75%), Newspapers (69.01%), Periodicals (6.97%) and Miscellaneous (8.27%). As regards classification, the PAROLE text-typology schema (i.e. classification by Medium, Genre and Topic) has been further elaborated by introducing open-ended sets of categories specifying detailed Genre and Topic for each text (Gavrilidou et al., 1998).

SGML Mark-up has been used for the bibliographical and the structural annotation of the texts at sentence level. The ILSP tokenizer has been used to mark tokens such as foreign words (i.e. non-Greek), abbreviations, digits and list elements. The tokenized texts are then used to populate the database.

The database used for this corpus is a typical relational database and not a text-tools-based database commonly used for the construction of electronic corpora (Futrelle and Zhang, 1994; Loeffen, 1994). The user interface consists of a group of HTML web pages which are created dynamically and present a graphical front-end for the corpus database.

The present paper describes the methodology and implementation for all the steps involved in the development of the online system for the HNC. This includes description of the ILSP tools used for the processing of the texts, the construction and population of the relational database, the creation of the Web application and interface and the services available to the user.

2. Collection of the material

2.1. Collection criteria

The HNC Corpus developed by ILSP is being compiled since 1992, when the process of its construction started as a part of the Parole LE2-4017/10369 project. It consists of more than 24 million tokens (running words), which are classified and annotated according to the PAROLE Corpus Encoding Standard. It consists of more than 24 million tokens (running words), covering a large variety of written Greek texts, published from 1976 on.

Monolingual general language corpora aim at the reflection of the actual use of a language (Johansson and Stenstrom, 1991; Sinclair, 1987; Zampolli, 1990). HNC intends to provide evidence for the current use of Modern Greek. Thus, the following design criteria were used in terms of text selection for HNC:

- texts originally written in Greek
- degree of readership (the selected newspapers are amongst the ones with the highest circulation; most of the selected books have been on best-selling lists)
- register (the selected texts are derived from various media and cover various genres and topics)
- exclusion of texts rich in idiomatic or dialectic forms

2.2. Sampling method

As regards books, the texts which are included in the corpus are samples of the original text supplied by the sources. This policy was imposed by reasons related to the copyright status of the texts, as sampling provides a certain guarantee to the copyright holders against any commercial exploitation of the content of the text. The selection of the text sample is random. Texts belonging to all other Medium types (see below) are included in their full length.

2.3. Text classification

Classification of texts adheres to the PAROLE standards (PAROLE, 1995) which follow the TEI and EAGLES guidelines (Sperberg-McQueen and Burnard, 1994;

EAGLES, 1994). Texts are classified as regards Medium, Genre and Topic. As far as Medium is concerned, texts are classified into the following categories, according to their source:

- Book
- Newspaper
- Periodical
- Miscellaneous (including correspondence, electronic texts, ephemera, hand-written material and typed material)

As regards Genre, the PAROLE categories (see Appendix) have been refined; open-ended sets of categories have been introduced (Detailed Genre), providing the user with a more accurate classification of the text. A similar procedure has been followed during the categorization according to Topic (see Appendix). A further subclassification was adopted to indicate the text's Detailed Topic (e.g. some Detailed Topics corresponding to the topic "Health" are: "Medicine" "Dentistry" and "Psychology"). Besides the text classification, for each text bibliographical information (e.g. title, author, date) is coded.

3. Tokenization and Sentence splitting

Recognizing and labeling surface phenomena in the text isis a prerequisite for most Natural Language Processing systems. In order to be able to make full use of the corpus, texts should be rendered in an appropriate form. In this first stage of processing parallel texts are normalized and handled.

Normalization and text handling can be seen as a sophisticated interface between input text streams and various text manipulation modules. At the stage of analysis, the text handler has the responsibility of transforming a text from the original form in which it is found into a form suitable for the manipulation required by the application. The main operations usually associated with the text handler include:

- format analysis of the input text's physical appearance (as evidenced by the word-processing and/or typesetting commands, such as bold and italic characters, indentation, etc.) and mapping of these into a standardized markup language or a canonical form recognized by the application. The texts contained in the HNC corpus are in simple unstructured ASCII format so that word-processing and/or typesetting information has been excluded.
- identification of textual units at the level of paragraphs, sentences and tokens.
- identification of extra-linguistic elements, such as dates, abbreviations, acronyms, list enumerators, digits, etc.

Though rarely discussed and sometimes quickly dismissed, tokenization in a text processing system poses a number of thorny questions, few of which have any perfect answers. During the last decade, however, work on tokenization and sentence segmentation has been driven by the development of systems aiming at automatically exploiting real - life texts. (Grefenstette and Tapanainen, 1994) apply regular expression grammars, equip the system with abbreviation lists and improve sentence recognition by adding increasing levels of linguistic sophistication. (Palmer and Hearst, 1994) implement an efficient, trainable algorithm that uses a lexicon with part-of-speech probabilities and a feed-forward neural network. (Chanod and Tapanainen, 1996) propose a finite-state automaton for simple tokens and a lexical transducer that encodes a wide variety of multiword expressions. (Reynar and Ratnaparkhi, 1997) propose a solution based on a maximum entropy model which requires a pre-annotated corpus and a few hints about what information to use.

In the framework of the presented corpus application, basic text handling is performed with the use of a Multext-like tokenizer (Di Christo et al., 1995) developed by ILSP. This includes identifying word boundaries, sentence boundaries, dates, abbreviations, etc. Identifying word and sentence boundaries involves resolving ambiguity in punctuation use since structurally recognizable tokens may contain ambiguous punctuation; this may be the case for digits, alphanumeric references, dates, acronyms and abbreviations. Following common practice, the tokenizer makes use of a regular expression definition of words, coupled with downstream precompiled lists for the Greek language and simple heuristics for distinguishing between these abbreviations or other evident abbreviations and final stops. This proves to be sufficient for recognizing sentences and words effectively.

4. Construction of the database

The relational database is the core of the HNC. It contains all the text material, the tags, the identifiers and the indexes used by the HNC. No additional material is saved in external text files. The alternative solution, i.e. to use a relational database only to hold the indexes, and to store the rest of the information in text files was not adopted, given that relational databases today can hold vast amounts of textual data without any major performance penalties. All the information required by the corpus front-end can be extracted from the database using standard SQL queries.

There are three types of objects in the corpus database: database tables, indexes and stored procedures. Database tables contain the data. Indexes are used to speed up the queries. Stored procedures are basically some preformatted SQL queries used by the front-end.

The database tables used in this application can be further divided into three distinct categories: main tables, auxiliary tables and combined tables. The main tables contain information related to the documents that constitute the corpus (including the PAROLE text classification), the sentences that the documents are made of and the words that those sentences contain. The tokenized texts

are used by a corpus loader to fill those tables. The tables also give information on the distance between words in the sentences, which is used in some of the queries. The auxiliary tables contain the morphological lexicon used in queries performing search by lemma. The combined tables are a combination of main and auxiliary tables and are used only to assist the query execution speed, i.e. they facilitate certain types of complicated queries to execute faster they would if formulated using only the main and auxiliary tables. In a way, the combined tables are a form of (large and complicated) indexes.

5. The Morphological lexicon

The Morphological lexicon of ILSP contains 65,500 lemmas, which produce 1,650,000 inflected forms. The data organization is the broadly used structure that connects lemmas to inflectional paradigms, through which the inflected forms are produced. Lemmas bear information on stress position and dieresis diacritic position, while inflectional paradigms link the lemmas to inflectional endings' indexes and to stress paradigms' indexes. Stress paradigms code stress movement (Greek inflection is characterized by the possibility of movement of the stress diacritic on the different inflected forms' syllables. Thus, the nominative case of a noun might have the stress diacritic on the antepenultimate syllable, whereas the genitive case "moves" it to the penultimate).

Given the above, three types of data are included in the Lexicon:

- lexical data: stems, endings and uninflected words
- inflectional data: inflectional paradigms
- features: these consist of Part of Speech information and a set of features that are used to characterise each word-form of the Greek language at the morphosyntactic level

Inflectional paradigms

Inflectional paradigms for nominal inflection, for instance, include the information shown in Table 1.

Inflectional Paradigm
Grammatical category (PoS)
Gender (Nouns, Adjectives, Pronouns, Participles)
Index: group of endings
Index: stress category

Table 1: ILSP Lexicon - Inflectional paradigm

An example of an inflectional paradigm is given in Table 2.

The Morphological Lexicon in this application is used for the lexicon look-up process, in the case of a lemma-based query or a query based on morphosyntactic tags (see section 6).

6. Web interface and search capabilities

During the last few years we have witnessed an explosion of Web usage, as well as an explosion of web technologies and tools. Many of those technologies have matured considerably, and building complete database front-ends based almost entirely on Web technologies is possible. This web-based approach in the construction of the front end for the HNC was considered preferable, since this makes the corpus available to the widest possible audience. Thus, the corpus database and the front-end for the corpus database constitute a complete Internet application, namely a three-tiered web database application.

The first tier of this web application is the database that contains all the data. The second tier is the web server that communicates with the database, performing database lookups depending on criteria set by the user and formatting the results that will be presented to the user. The third tier is a set of dynamically created HTML pages that acts as the user interface. The user can access the ILSP HNC using any web browser on any client platform (e.g. PC, Mac or Unix). The only requirement is access to Internet; no additional software installation or disk space in the client machine is needed. Also, as the database grows there is no need for the user to change anything on his/her machine. Of course, since this is a Modern Greek corpus, it is assumed that the user has some minimal Greek capabilities in his/her computer (i.e. a Greek font installed).

The absence of any client-site installation requirements is one of the major advantages of the Internet application approach. It makes all software updates and upgrades and possible bug fixes transparent to the user. The only disadvantage of this approach is the Internet latency that can be quite annoying in some cases. However, this is a temporary disadvantage, since the Internet bandwidth increases continuously and our application is a text based application that does not require the transfer of images, sound or video through the Internet, i.e. the amount of data transferred through the Internet is low compared to other applications.

Furthermore, we tried to make the user interface as user-friendly as possible. The user who accesses the dynamically created HTML pages can perform all the necessary work using an entirely graphical interface, with standard web page elements (e.g. drop-down lists and checkboxes). The user doesn't have to be familiar with any special tags, special characters, symbols or syntax. This graphical user-friendly web interface is quite rich, enabling the user to access almost the full capabilities of the corpus. The queries can use the whole corpus as search space or be confined in a user-defined subset of the corpus. These corpus subsets are completely dynamical and can be defined using specific types of constraints, based on the refined classification of texts. All the results are returned to the user as a set of sentences. If the user wants to, s/he can copy these sentences to a simple text editor (like Notepad in Microsoft Windows) and save them as text files in his or her local hard disk for future reference and use.

Index of inflectional paradigm	PoS	Gender	Index of group of endings	Index of stress category
16	Noun	Masculine	49	1

Table 2: ILSP Lexicon - Inflectional paradigm example

The starting web page for the HNC can be found in the following web address: http://www.xanthi.ilsp.gr/corpus/. For a user to access the corpus using the web interface, the first step is to login into the system providing a username and a password. If the user doesn't specify an existing username and a password then the system accepts him/her as a "guest user" with limited searching capabilities. In the second (optional) step the user can specify a search subcorpus. If the user wants to specify a subcorpus, s/he can do so using three web pages that guide the user to specify some or all of the eight different types of constraints available: medium, genre, topic, detailed genre, detailed topic, publisher, author and date. For registered users the definition of subcorpus is saved for future use. Also, if the user wants to use a previously saved subcorpus, s/he can do that after the first step or at any time afterwards. There is also a web page for the management of those pre-defined subcorpora, where the user can delete a previously defined subcorpus.

Following the optional specification of the search space, the interface guides the user through the formulation of the query to be performed (Fig. 1). This is the third and final step. It allows the user to search for sentences in the texts that contain up to three given word forms or lemmas or morphosyntactic tags or any combination thereof, with specific maximum distances between them, specified by the user. The search for specific word forms is performed by querying the corpus database, while the search for lemmas and morphosyntactic tags is performed via lexicon look-up (which identifies all the wordforms connected to the specific lemma) and subsequent querying of the database. The user can also specify the maximum number of sentences to be returned in the results (up to 1000 sentences), the number of sentences per page in the results and whether the currently chosen subcorpus, if any, will be used or not.

The results are presented to the user as a set of sentences characterized by a number identifying the sentence (Fig. 2). Each of the up to three word forms (or lemmas or morphosyntactic tags) is colored in the resulting sentences, so that the user can easily identify the reason a particular sentence was returned by the system. For each of those sentences the user can also choose to view all relevant information (document type, author etc) for the text that contains the particular sentence by clicking on the sentence identification number. If the user wants to save locally the sentences returned by the system, s/he can do so simply by copying the sentences to a local text file. It would be easy to present the user with the ability to view the complete document containing the particular sentence; however, due to copyright restrictions this feature has not been implemented.

As mentioned above, all the results are given to the user as a set of sentences. However, there is also another capability presented to the user, i.e. the execution of some statistical-type queries (frequency lists) for the words and lemmas. The user can view the list of most common words or lemmas of the HNC or can search for the frequency of occurrence of a particular word or lemma specified by the user. When the user requests the frequency of a lemma, the results include frequency information for every word form related to this lemma.

All these queries are performed using standard SQL queries of the database. Depending on the options chosen by the user, the interface sends the appropriate parameters to the server-side application that resides in the web server. This server side application formulates the SQL query and forwards it (using ODBC) to the relational database. The results are returned by the database to the server-side application that formats them as a series of HTML pages and returns them to the client.

From the programmer's viewpoint, implementation of the above interface is as easy (or as difficult!) as any other web database application. The most difficult part is to implement the rich search capabilities discussed above. Since there are up to three placeholders (for word forms or lemmas or morphosyntactic tags) and for each of those there are three possible types (word form, lemma or morphosyntactic tag) there are $3 + 9 + 27 = 39$ types of SQL queries. Actually, the number of possible queries is twice this number (i.e. $39 \times 2 = 78$) since we can have or not have a specified subcorpus. Since all the queries are transformed to SQL queries that are subsequently submitted to the relational database, we have to create 78 different types of queries from the information entered by the user. Of course, there are many similarities among those 78 types of queries. The major difficulty is to optimize each type because, as we found out, some types of queries are quite time consuming. This performance optimization process depends on the database structure but also on the formulation of each SQL query. Although some general guidelines do exist (Date, 1999) they depend on the particular DBMS system we are using and the performance optimization is non-trivial process.

7. Conclusion

We have described the Hellenic National Corpus (HNC), the corpus of Modern Greek language developed by the Institute for Language and Speech Processing (ILSP). The HNC has been constructed using the ILSP tokenizer, a tool developed by ILSP over the years. The core of the HNC is a relational database that contains all

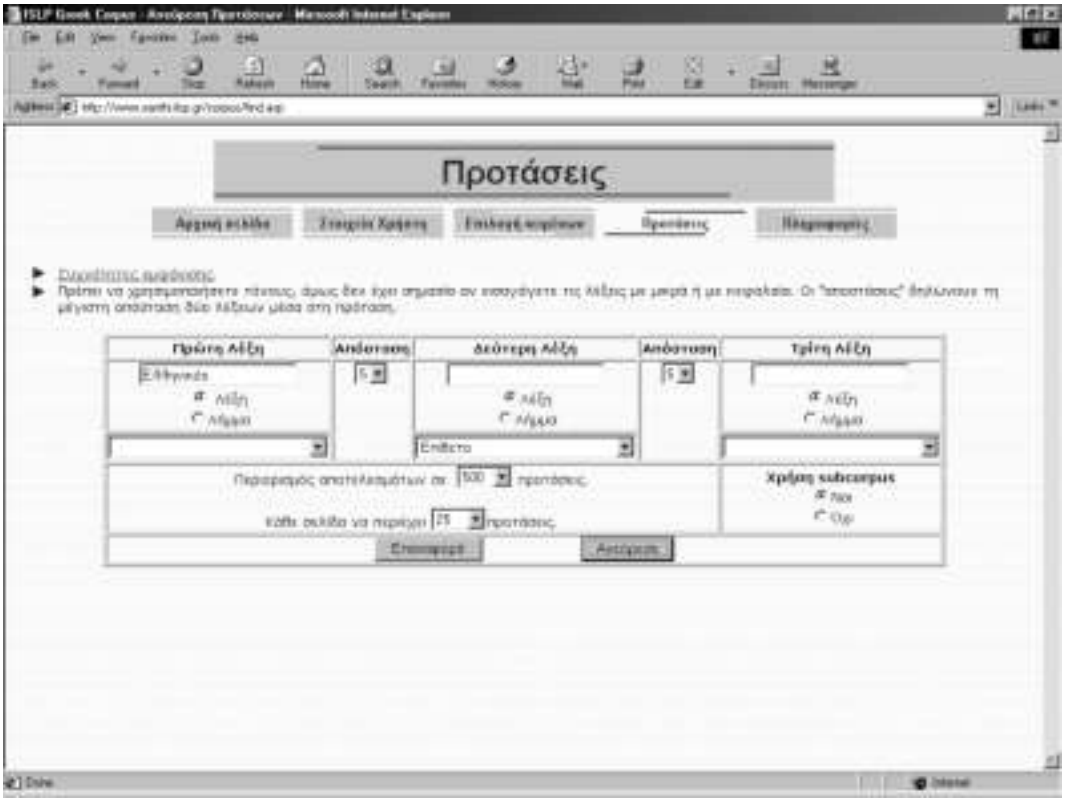

Figure 1: The HNC search page.

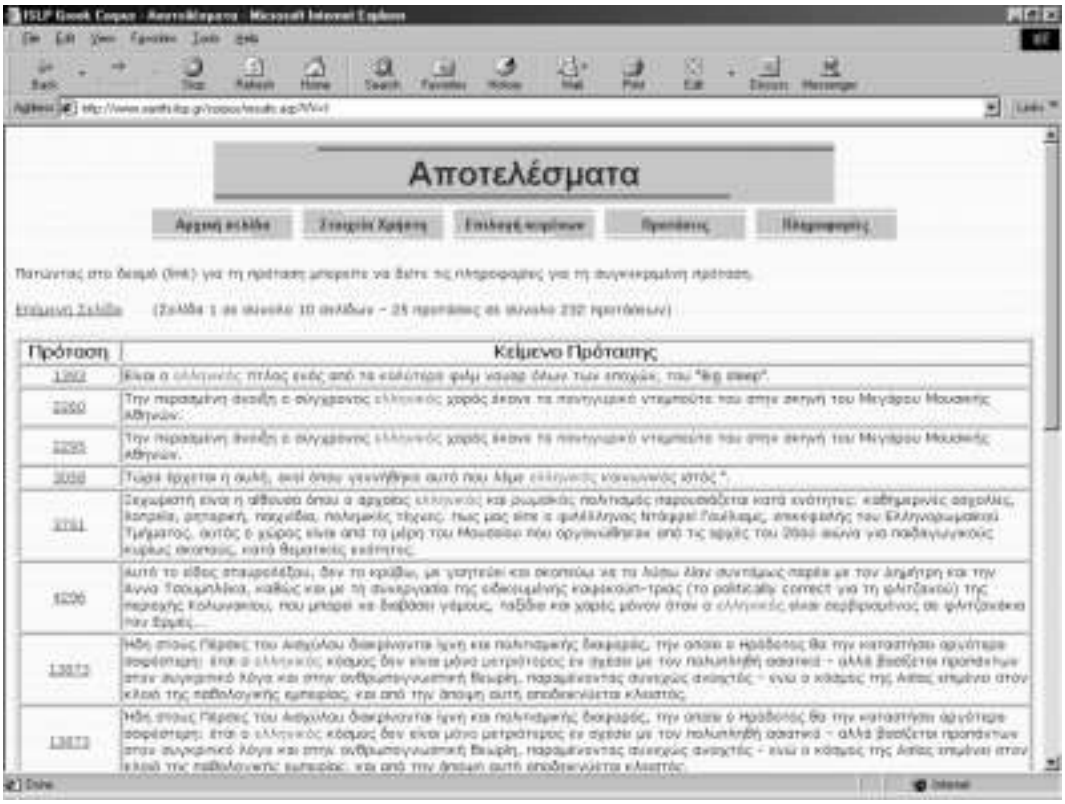

Figure 2: The HNC results page.

the processed data. The interface for the HNC is a Web database application, which can be accessed by anyone through the Internet. HNC contains over 24 million words but it is expected to grow as more data is entered into the database.

8. Appendix

GENRE

- N/A: not applicable/mixed/unknown/not identified
- ADVertising
- DIScussion, debate and conversation, written as well as spoken, including i.e. interviews, parliamentary speeches, letters to the editor.
- FEAture article in newspaper etc. which does not belong to INF or another, more specific genre; reviews, radio/TV magazines, etc.
- FICtion, including faction and e.g. comic strips, entertainment, childrens and youth pages, jokes, and games; drama, including film manuscripts and e.g. tv-series; poems and song lyrics.
- INFormation: news article in newspaper texts, as well as similar programs in radio and television. Folders and leaflets from e.g. the authorities. Posters; signs.
- INStruction, including reference and text books, but also that kind of correspondence column in e.g. magazines, where readers questions are answered by specialists in tax, gardening, health etc.
- NON-fiction (other): biography, including obituaries and autobiographies; sermons; school and student essays; etc.
- OFFicial text, including laws, government circulars, official announcements, business correspondence
- PRIvate text, like diaries and private letters

TOPIC

- N/A: not applicable/mixed/unknown/not identified/miscellaneous
- BUSiness and economy, including advertising
- GEOgraphy and travel; anthropology and folklore
- HEAlth, including psychology
- HIStory, including biography
- HUManities and culture: art, literature, music, philosophy, religion, etc.
- LEIsure: sport, television, food, etc.
- SCIence and technology, including mathematics and environment
- SOCiety and politics, including, i.e. law, crime, and social services

9. References

Chanod, J. P. and P. Tapanainen, 1996. A non-deterministic tokenizer for finite-state parsing. In *Workshop on Extended Finite State Models of Language*. Budapest, Hungary.

Date, C. J., 1999. *An Introduction to Database Systems*, chapter 17. Addison-Wesley.

Di Christo, P. S. Harie, C. De Loupy, N. Ide, and J. Veronis, 1995. Set of programs for segmentation and lexical look up. Technical report, MULTEXT LRE 62-050. Project Deliverable 2.2.1.

EAGLES, 1994. Corpus encoding: Draft. Technical report, EAGLES. Document EAG-CSG/IR-T21.

Futrelle, Robert P. and Xiaolan Zhang, 1994. Large-scale persistent object systems for corpus linguistics and information retrieval. In *DL '94 Proceedings*.

Gavrilidou, M., P. Labropoulou, N. Papakostopoulou, S. Spiliotopoulou, and N. Nassos, 1998. Greek corpus documentation. Technical report, ILSP. Parole LE2-4017/10369, WP2.9-WP-ATH-1.

Grefenstette, G. and P. Tapanainen, 1994. What is a word, what is a sentence? problems of tokenization. In *3rd International Conference on Computational Lexicography*. Budapest, Hungary.

Johansson, S. and A.-B. Stenstrom (eds.), 1991. *English Computer Corpora : Selected Papers and Research Guide*. Mouton de Gruyter.

Loeffen, A., 1994. Text databases: A survey of text models and systems. *SIGMOD RECORD*, 23(1):97–106.

Palmer, D. D. and M. A. Hearst, 1994. Adaptive sentence boundary disambiguation. In *4th Conference on Applied Natural Language Processing*. Stuttgart, Germany.

PAROLE, 1995. Design and composition of reusable harmonised written language reference corpora for european languages. Technical report, PAROLE Consortium. MLAP: 63-386, WP 4 - Task 1.1.

Reynar, J. C. and A. R. Ratnaparkhi, 1997. A maximum entropy approach to identifying sentence boundaries. In *5th Conference on Applied Natural Language Processing*. Washington, D.C.

Sinclair, J. M. (ed.), 1987. *Looking up: An account of the COBUILD project*. Collins Publishers.

Sperberg-McQueen, C. M. and L. Burnard, 1994. Guidelines for electronic text encoding and interchange: Tei-p3. Technical report, Chicago and Oxford. ACH-ACL-ALLC Text Coding Initiative.

Zampolli, A., 1990. A survey of european corpus resources. *UK SALT Club*.