A MODULAR ARCHITECTURE FOR THE STORAGE AND MANAGEMENT OF HETEROGENEOUS LEXICAL INFORMATION

D. Assimakopoulos[+], *I. D. Koutsoubos*[+*], *M. Miatidis*[+*], *I. Tsakou*[+], *A. Ntoulas*[+*], *S. Stamou*[+*], *D.Christodoulakis*[+*]

* Computer Engineering & Informatics Department
+ Computer Technology Institute
University of Patras, Building B, 26 500 Rio, Patras, Hellas
{dassimak, ntoulas, stamou, tsakou, dxri}@cti.gr
{koutsoub, miatidis}@ceid.upatras.gr

ABSTRACT

In this paper, we present a three-layer modular architecture of a flexible system, providing the infrastructure for the storage and handling of large amounts of heterogeneous lexical information. Each of the three layers handles different types of linguistic data (morphological information, lexicographic information and language internal relations). In each layer the data is handled autonomously, while all three interact through a main entity (lemma - POS[1]), in order to exchange information. This architecture is implemented using a commercial RDBMS.

Keywords: morphological information, lexicographic information, language internal relations, wordnet.

1. INTRODUCTION

In this paper we present a modular architecture for a linguistic resource for the storage and handling of large amounts of lexical information. The motivation for designing such architecture was the need for a flexible system for storing and handling different kinds of lexical information. Such a system is necessary in order to integrate in one unified resource morphosyntactic and lexicographic information, as well as language internal / semantic relations. In this way, it is possible to access and exploit simultaneously different types of natural language information (e.g. morphological, semantic, stylistic, etc.) for the development of flexible and more complete linguistic applications.

As a result, we designed an architecture of three layers residing on a commercial RDBMS. Each one provides the infrastructure for storing and managing a different kind of lexical information: morphological, lexicographic and language internal relations. Each of these layers handles the respective data autonomously, while all three interact in order to exchange information.

Section 2 describes the structure and contents of the layer responsible for the storage and handling of morphological information. Section 3 describes the structure and contents of the lexicographic information layer, while Section 4 provides information on the structure and contents of the language internal relations layer. In Section 5 we give the outline of the interconnection of the three layers. Finally, Section 6 contains our future goals and some conclusions.

2. MORPHOLOGICAL LAYER

The first layer of the relational database (RDBMS) comprises and handles morphological information. The contents of this layer originated from Computer Technology Institute's (CTI) morphological lexicon. The information of CTI's lexicon is stored in files, imported into the system's morphological layer using filters developed for this purpose.

The morphological layer contains approximately 90.000 lemmas, along with their respective word forms reaching approximately one million. All word forms are fully morphosyntactically annotated according to specific rules (260 totally), describing the declination of all declinable Greek words [14]. The information provided for each lemma contains all the possible word forms of the lemma, each accompanied by a set of attributes, such as part of speech, gender, case, number etc. All the attributes attached to each word form depend on the lemma's part of speech [16], which is the basic entity around which all layers are organised. Consequently, entity relations prevent the user from attaching to a word form attributes which do not agree with its part of speech (eg. gender to a verb).

An example of a lemma and some of its word forms encoded in the morphological layer of the system follows:

{á-íãñ $OUSARSas3}
Ïíãñáò (OUSIASTIKO, ARCHAIO, ARSENIKO, ENIKOS, ONOMASTIKI)[2]
Ïíãñá (OUSIASTIKO, ARCHAIO, ARSENIKO, ENIKOS, GENIKI, AITIATIKI, KLITIKI)

[1] Lemma – Part of Speech

[2] i.e. Man (NOUN, ANCIENT, MALE, SINGULAR, NOMINATIVE)

Üíäñåò (OUSIASTIKO, ARCHAIO, ARSENIKO, PLITHINTIKOS, ONOMASTIKI, AITIATIKI, KLITIKI)
áíäñþ (OUSIASTIKO, ARCHAIO, ARSENIKO, PLITHINTIKOS, GENIKI)

The morphological layer enables users to trace the lemma – part of speech from which a word form stems. The information stored in each layer is treated autonomously, whereas at the same time interacts with the other two layers of the system described in Sections 3 and 4. On figure 1 below is shown the structure of the morphological layer.

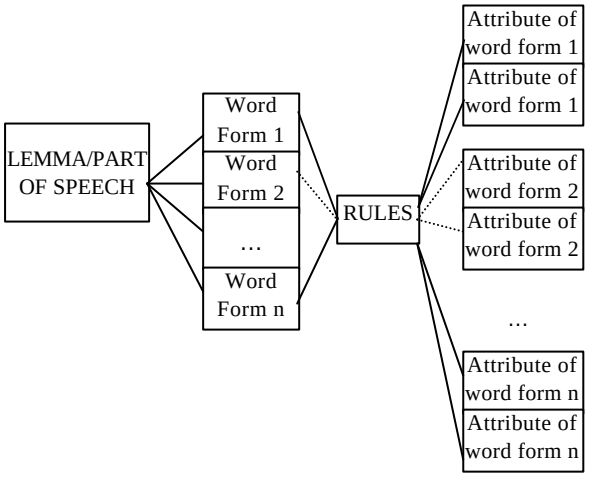

Figure 1: Basic structure of the morphological layer

Once the morphological information is stored into the database, the user can ask for all possible word forms of a lemma or all possible attributes of any word form of a lemma – part of speech, or make all kinds of queries related to the attributes of word forms.

3. LEXICOGRAPHIC INFORMATION LAYER

The lexicographic information layer comprises and handles lexicographic information, such as word meanings, idiomatic expressions, etc., providing in this way the infrastrusture for the storage of a lexicon. As shown in figure 2, for each lemma, one or more word meanings are provided, as well as the lemma' s etymology.

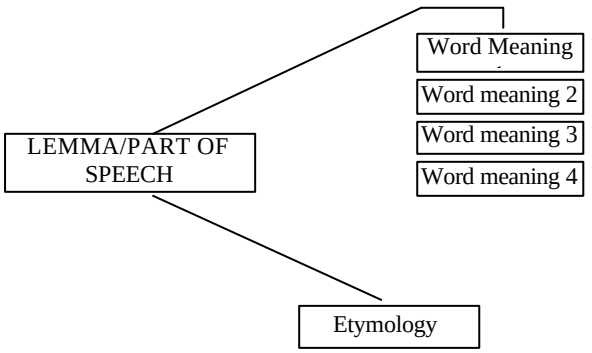

Figure 2: Basic structure of the lexicographic information layer

Furthermore, each word meaning is accompanied by a number of attributes / characteristics, as shown in figure 3. Specifically, each word meaning can be accompanied by comments (eg. the Greek word "Üíèñùðï ð' is interpreted as "êüóìï'ð only when used in plural), synonyms, antonyms, usage examples, idiomatic expressions containing the specific word meaning, syntactic information, domain specific terminology, as well as stylistic remarks (eg. formal / informal use, etc.).

The lexicographic information layer is structured in a way so as to provide to users a number of services, such as "return all usage examples for the lemma <x>, or "return all word meanings labelled as medical domain specific terminology".

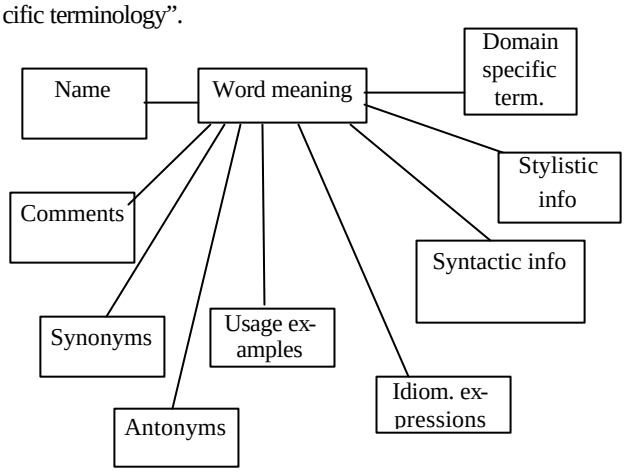

Figure 3: Attributes / characteristics of word meanings.

4. LANGUAGE INTERNAL RELATIONS LAYER

The Language Internal Relations Layer is based on the EuroWordNet architecture [9][18], in an effort to achieve maximal compatibility with it. This layer aims at providing the infrastructure for a semantic network of relations between word meanings of the Greek language.

The main entities of this layer are based on EuroWordNet and are the following: A) top concept, B) base concept, C) synset and D) lemma. A) The *top concept* is the main element of the EuroWordNet top ontology, a total of 63 fundamental semantic distinctions which are language-independent, used to relate and classify lemmas [9], [18]. This 3order entity ontology is used in the language internal relations layer of our database to assign semantic characteristics to the

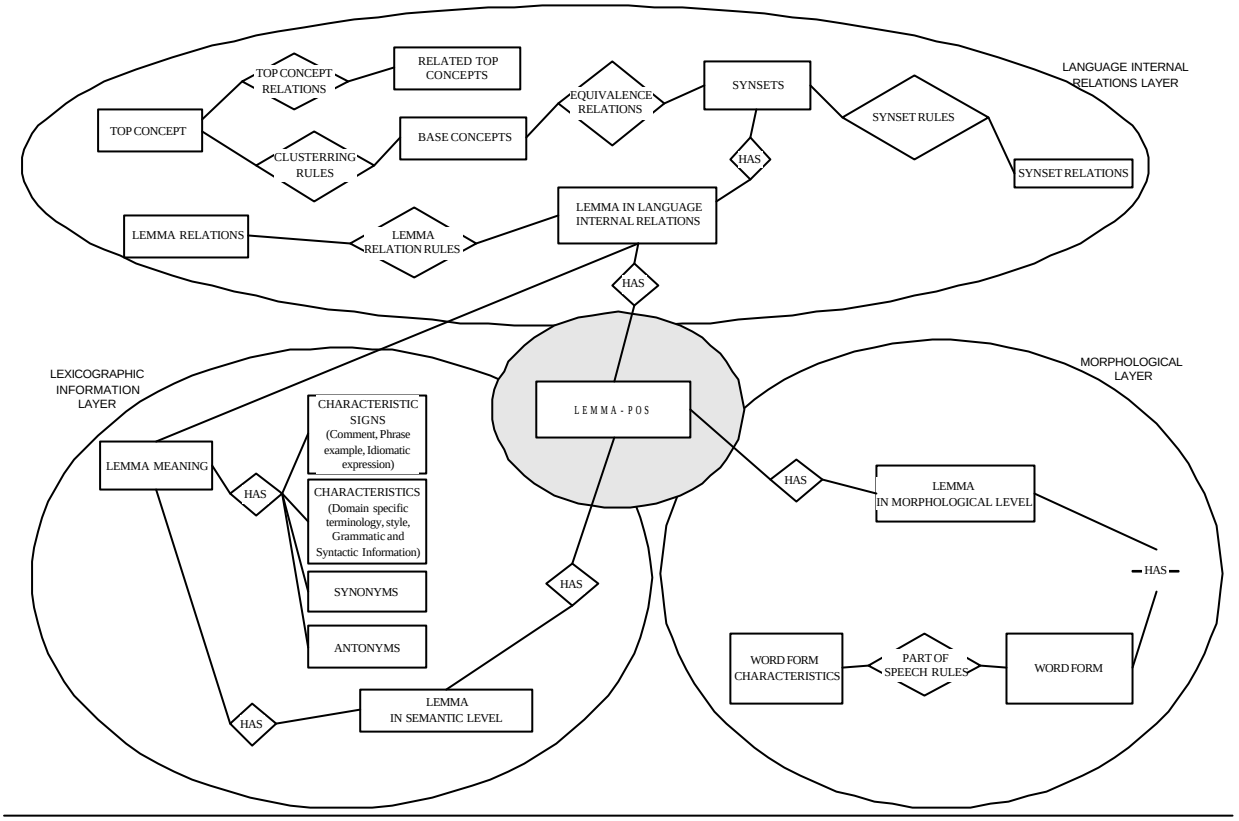

Figure 4: Entity – Relation Diagram

base concepts. For example, the «Ï íôüôçôá 1ò ÔÜÎçò» (1st order entity) has hyponym «ÊáôáãùãÞ» (Origin), the «êáôÜóôáóç» (Situation) has hyperonym the «Ï íôüôçôá 2ò ÔÜÎçò» (2nd Order Entity). B) The second entity of this layer is the *Base concept*. Base consepts are concepts with most semantic relations that function as anchors to attach other concepts. The base concepts are related to sysnets, through specified equivalence relations [18]; a base concept example is: «öõôü» (plant) which is related via hyponymy to concepts such as «ðï ñôï êáëéÜ» (orange tree) «ëåìï íéÜ» (lemon tree), etc. C) A *Synset* is a set of lemmas with the same part of speech related to each other through synonymy [5], [17]; a synset example for the Greek language is the following: {ßðõï ò, Üëï ãï , Üôé} (theylal mean "horse" and are used to refer to the same concept in different contexts). D) *Lemma* is the combination of the pair lemma – part of speech.

The language internal relations layer is divided in four parts: i) the top concepts part, ii) the base concepts part iii) the synset – relation part and iv) the lemma – relation part. These parts are hierarchically related to each other, through specific language internal relations based on the EuroWordNet bibliography, i.e. the top concepts part is related to the the base concepts part, etc. Specifically, top concepts are related to base concepts through language internal relations (eg. hyperonymy, hyponymy, etc.), according to specific part of speech constraints. Base concepts are related to synsets through equivalence relations (eg. synonymy, near synonymy, has holonym, etc.), according to specific part of speech constraints. Synsets are related to other synsets through language internal relations (synonymy, antonymy, causes, etc.), according to specific part of speech constraints. Finally, lemmas are related to each other (through antonymy, pertains_to relation, etc), according to specific part of speech constraints. [5], [18].

5. THE INTERCONNECTION OF THE THREE LAYERS

The system is designed in such a way that redundancy is avoided since the information needed by more than one layers is stored only in one layer and can be used by the others as well. Consequently, the effects of any extensions or alterations to any of the layers are localized without affecting the general architecture. Finally, it is structured in a way that provides to the user a variety of services regarding the data stored in the system, such as the ability of inserting, selecting, deleting, updating a lemma and its attributes (meanings, synonyms, antonyms, word forms, etc), or answer to queries such as "Return all the synonyms of lemma <X>, along with usage examples".

As shown in Figure 4, all three layers are interconnected via the common entity *lemma-pos*, where each lemma is codded according to its part of speech. Therefore, each

lemma is considered as a pair of lemma and part of speech. The reason for using this combination is to avoid semantic and morphological ambiguities, which are frequent for the Greek language as it is a highly inflectional one [7], [15]. Another reason is that in the language internal relations layer, lemmas, synsets, base concepts and top concepts are related according to their part of speech. In addition, it should be mentioned that each lemma – POS is examined according to its meaning extracted from the lexicographic information layer of the RDBMS. Each layer asks for permission to use a lemma – POS; if the specific lemma – POS exists, the user of the layer is able to insert, select, delete, or update the lemma's attributes, such as meanings, synonyms, antonyms, word forms, etc. If the lemma does not exist, then the lemma and its attributes can be inserted.

6. CONCLUSIONS AND PERSPECTIVES

A three-layer modular architecture of a flexible system, capable of storing and handling large amounts of heterogeneous lexical information was presented in this paper. Each of the three layers can handle the information autonomously and interacts with the others in order to provide the required services to the end user.

For the time being we have implemented the above architecture using a commercial RDBMS. Our aim is to design an Open Language Engineering System (OLES): a language engineering system that provides the framework in which more cognitive models can be developed and even imported within the architecture. We also intent to develop open tools for associating lemmas that have been imported into the system, in an effort to make possible for other developers to create new services or even enhance the old ones.

Once the information is imported into the database and a certain number of services is available, a web interface can be added to the whole infrastructure. In this way it will be possible either to browse the lexical information or use a network-based API to create new NLP applications.

7. REFERENCES

[1] Cole, R. et al., 1997, Survey of The State of The Art in Human Language Technology. Cambridge University Press.

[2] Domenig, M., Ten Hacken, P., 1992, Word Manager: A System for Morphological Dictionaries. Olms, Hildesheim.

[3] EAGLES project, 1996, Creating standards on Electronic Lexicons, Interim Reports.

[4] EAGLES, 1996, Computational Lexicons Working Group Reading Guide.

[5] Fellbaum, C., 1998, WordNet, An Electronic Lexical Database.

[6] Gakis P., Orphanos G. and Iordanidou A., 1999, Computational Processing of Modern Greek: Recording the Morphosyntactic Ambiguity (in Greek), Glossa, 49, Ekdosis Neas Pedias.

[7] Gibbon, D. 1997. "Computational Lexicography for Speech and Language", course materials, ELSNET '97 summer school

[8] Grishman, R., Macleod, C., Meyers, A., 1994, Comlex Syntax: Building a Computational Lexicon. Project Report.

[9] Ide, N., Greenstein, D., 1998, EuroWordNet. Computers and the Humanities. 32 Double Special Issue on EuroWordNet.

[10] International Journal of Lexicography, 1994, WordNet: Five Papers on WordNet.

[11] Mackridge, P., 1987. The Modern Grek Language. A Descriptive Analysis of Standard Modern Greek. Clarenton Press, Oxford.

[12] Ntoulas, A., Stamou S., Tsakou I., Tsalidis C., Tzagarakis M., and Vagelatos A., 2000, "Use Of A Morphosyntactic Lexicon As The Basis For The Implementation of the Greek WordNet." Proceedings of the 2nd International Conference on Natural Language Processing, Patras, Greece, pp. 49-56.

[13] Orphanos G. and Tsalidis C., 1999, "Combining handcrafted and corpus-acquired Lexical Knowledge into a Morphosyntactic Tagger", Proceedings of the 2nd Research Colloquium for Computational Linguistics in United Kingdom (CLUK), Essex, UK.

[14] Orphanos G., 2000, Computational Morphosyntactic Analysis of Modern Greek, Ph.D. Dissertation (in Greek), Computer Engineering and Informatics Department, University of Patras, Greece.

[15] Orphanos G., Gakis P. and Iordanidou A. 1999. "Morphosyntactic Ambiguity in Modern Greek: The case of Adjective-Noun-Adverb" (in Greek), Proceedings of the 20th Annual Meeting of the Department of Linguistics, Thessaloniki, Greece.

[16] Tsalidis, C., Orphanos, G., 1995 Word Description Languages, In Proceedings of the 1st Workshop in Natural Language Processing, Athens, Greece, pp. 239-253.

[17] Vagelatos A., Triantopoulou T., Tsalidis C., Christodoulakis D., 1995, "Utilization of a Lexicon for Spelling Correction in Modern Greek". 10th Annual Symposium on Applied Computing (SAC '95) - Special Track on Artificial Inteligence, ACM Computing Week, February 1995, Nashville, Tenesse, U.S.A.

[18] Vossen, P., 1998, EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers.