"ΛΟΓΟΠΛΟΗΓΗΣΗ"
Ιούνιος 1998
Τεύχος 4

Επιστημονικός Υπεύθυνος:
Καθηγητής Γιώργος Καραγιάννης

Υπεύθυνη Έκδοσης:
Δρ. Ιωάννα Μαλαγαρδή

Συνεργάτες:
Δρ. Στέλιος Μπακαμίδης
Σπύρος Ράπτης
Αναστάσιος Πατρικάκος
Δέσποινα Σκούταρη
Αθανασία Φούρλα

Γραφίστας:
Άρτεμις Γλάρου

Διεύθυνση:
Ινστιτούτο Επεξεργασίας του Λόγου
Αρτέμιδος 6 & Επιδαύρου
151 25 Παράδεισος Αμαρουσίου
τηλ.: 6800959 • fax: 6854270
e-mail: ioanna@ilsp.gr
http:// www.ilsp.gr

Την ευθύνη των κειμένων έχουν οι συγγραφείς.

Η χρηματοδότηση της έκδοσης αυτής έγινε από τα προγράμματα της ΓΓΕΤ "Γραφεία Διαμεσολάβησης" και "Ανθρώπινα Δίκτυα".

Η "Λογοπλοήγηση" διανέμεται δωρεάν.

"LogoNavigation"
June 1998
Issue 4

Scientific Director:
Professor George Carayannis

Edition Executive:
Dr. Ioanna Malagardi

Collaborators:
Dr. Stelios Bakamidis
Spyros Raptis
Anastasios Patrikakos
Despina Scutari
Athanassia Fourla

Graphics Designer:
Artemis Glarou

Address:
Institute for Language and Speech Processing
Artemidos 6 & Epidavrou Str.
151 25 Marousi
Athens, Greece
tel: 301- 6800959 • fax: 301-6854270
E-mail: ioanna@ilsp.gr
http://www.ilsp.gr

The authors are responsible for text content.

Funding for this issue was carried out by the "Liaison Offices" and the "Human Networks" programmes of the General Secretariat for Research and Technology

"LogoNavigation" is distributed free of charge.

Πίνακας Περιεχομένων / *Table of Contents*

Εισαγωγικό Σημείωμα

Το τέταρτο τεύχος της "Λογοπλοήγησης" περιέχει ποικιλία θεματικών ενοτήτων που αφορούν επιστημονικές δραστηριότητες στο πλαίσιο της Γλωσσικής Τεχνολογίας. Τα θέματα που δημοσιεύονται στο παρόν τεύχος προέρχονται κυρίως από εκδηλώσεις που πραγματοποιήθηκαν στο πλαίσιο των δραστηριοτήτων του έργου "Ανθρώπινο Δίκτυο Γλωσσικής Τεχνολογίας".

Η πρώτη θεματική ενότητα περιέχει περιλήψεις των εισηγήσεων του σεμιναρίου που διεξήχθη στο Ινστιτούτο Επεξεργασίας του Λόγου στις 17 Φεβρουαρίου 1998. Στο σεμινάριο συμμετείχαν ερευνητές από την Ολλανδία και τη Σουηδία καθώς και ο υπεύθυνος της Γλωσσικής Τεχνολογίας στην Ευρωπαϊκή Ένωση. Ορισμένα από τα κείμενα των περιλήψεων των εισηγήσεων συνοδεύονται με αντίγραφα διαφανειών που προβλήθηκαν κατά τη διάρκεια της ομιλίας. Η δεύτερη θεματική ενότητα αντλεί το υλικό της από την ημερίδα με θέμα "Ηλεκτρονική Λεξικογραφία" που πραγματοποιήθηκε επίσης στο Ινστιτούτο Επεξεργασίας του Λόγου στις 14 Μαρτίου 1998. Στην συγκεκριμένη ενότητα δημοσιεύονται τα κείμενα των περιλήψεων των εισηγήσεων της ημερίδας. Αμφότερα τα γεγονότα διοργανώθηκαν από το ΙΕΛ στο πλαίσιο των δραστηριοτήτων του Ανθρωπίνου Δικτύου Γλωσσικής Τεχνολογίας. Η τρίτη θεματική ενότητα περιέχει κείμενα μελών του Ανθρωπίνου Δικτύου Γλωσσικής Τεχνολογίας σχετικά με το θέμα της Ηλεκτρονικής Λεξικογραφίας. Η τέταρτη θεματική ενότητα περιέχει παρουσιάσεις νέων βιβλίων σχετικών με θέματα Γλώσσας και Τεχνολογίας. Η πέμπτη θεματική ενότητα περιλαμβάνει θέματα σχετικά με τη δημιουργία γλωσσαρίου όρων Γλωσσικής Τεχνολογίας και Πληροφορικής. Τέλος η έκτη θεματική ενότητα περιέχει ειδήσεις σχετικές με τη Γλωσσική Τεχνολογία και την Πληροφορική.

Introductory Note

The fourth issue of "LogoNavigation" contains a variety of thematic units concerning scientific activities in the context of Language Technology.

The first thematic unit contains summaries of the introductions carried out in the context of the seminar that took place at the Institute for Language and Speech Processing (ILSP) on February 17 1998. Researchers from Holland and Sweden as well as the Head of the Language Technology section at the European Commission participated in this seminar. Certain summary texts are accompanied by copies of slides presented by the lecturers during their introductions. The second thematic unit derives its material from the information day titled "Electronic Lexicography", held also on the premises of ILSP on March 14 1998. The seminar as well as the information day were organised by ILSP in the context of the Greek Human Network of Language Technology. The third thematic unit contains texts related to the issue of Electronic Lexicography from other members of the Greek Human Network of Language Technology. The fourth thematic unit contains the presentation of newly-published books related to issues of Language and Technology. The fifth thematic unit discusses terminology issues in the Language and Information Technology fields. Finally, the sixth unit contains news related to events in Language and Information Technology.

I. Περιλήψεις Εισηγήσεων του Σεμιναρίου / *Summaries of the Seminar's Introductions*

1. Seminar and Meeting of the Greek Human Network

Professor George Carayannis
Institute for Language and Speech Processing
Artemidos 6 & Epidaurou
Paradeisos Amarousiou
151 25 Athens, Hellas
gcara@ilsp.gr

I would like to say welcome everybody. We have met many times already since the human network started working, two years ago. As you remember, we had organisational meetings, concertation meetings and information meetings.

Our Journal/Newsletter is improving in quality. I hope that the network was helpful to your work during its initial steps and that it will be useful in the future also, especially if it succeeds in offering more sophisticated services.

Today's meeting is very important as we have with us three distinguished friends as our invited speakers. The three subjects covered today are extremely interesting in the field of language engineering. The first is from the speech domain, the second from the written text domain. The third speech has a strategic character.

Our speakers are well known for their pioneering work. We benefit form their presence here in the framework of the ELSNET board meeting, which took place yesterday at ILSP. For those of you who do not know, ELSNET is the European Language and Speech Network which was created some years ago. It is the very first ESPRIT network created in Europe and as it is very successful and useful, it has an increasing number of members and activities. Some of our national organisations are also members of ELSNET, which is involved in many policy issues in LE, providing information to its members, enhancing concertation at the European level, improving relationships between academia and industry etc.

Like ELSNET, the Greek Human Network has started through a proposal, which was submitted under a call by the General Secretariat for Research and Technology. We do not know how long the Greek network will stay alive as the related project expires on April. We plan to have two more information meetings like today's, the first in March and the second in April Dr. Malagardi will give you more details about these meetings at the end of this event.

We will begin with Prof. Björn Granström who needs no introduction, as he is well known. Björn has devoted many years of R&D to speech processing and speech synthesis. You will be given a specialist view on the new trends in speech synthesis.

Steven Krauwer is an expert in Machine Translation issues. He has been involved in many European projects and has important expertise in that field. He has been directing MT operations in Holland for many years. MT is a domain where views do not always converge. Steven will speak about the future of MT.

Roberto Cencioni is our last speaker. He will also speak about the future. He is the person who has the broader view on the LE field, through so many projects that he manages on behalf of the European Commission. He is also planning the future. As he is in the process of designing the work programme of the 5th FP, we will be given the opportunity to hear very fresh and valuable information.

2. Machine Translation: Towards New Metaphors

Professor Steven Krauwer
President of ELSNET
Utrecht Institute of Linguistics OTS
Trans 10, 3512 JK Utrecht, The Netherlands
steven.krauwer@let.ruu.nl

1. Is Machine Translation possible?

The question whether Machine Translation (MT) is possible, is a very delicate one for a researcher to answer. If the answer is "yes" we need evidence that this is indeed the case, i.e. that there are Machine Translation systems out in the world that can produce adequate translations, or at least some sort of evidence that this will be the case in the foreseeable future. If, on the other hand, the answer is "no", a large number of academic and industrial researchers will have to justify the millions they have spent over the last 40 years, and are still spending, on the MT enterprise. Let us first have a closer look at the answer, and then at the question.

1.1. The answer

Those who have ever been confronted with the unedited output of an MT system, will easily admit that the result contained errors, ranging from small, stylistic deviations, to an absolute distortion of the meaning of the text in the best case and absolute nonsense in the worst case. Some of the errors may look rather superficial, and easy to repair in a next version of the system (e.g. unknown words), but others might be much more difficult to tackle. Let us take the following two sentences (inspired by an example from Jerry Hobbs) as our starting point:

(a) The policemen fired at the demonstrating students because they FEARED violence.
(b) The policemen fired at the demonstrating students because they WANTED revolution.

From a purely syntactic and lexical point of view the two sentences are almost identical: the underlying syntactic structure is the same, and the only real change is that "feared" in sentence (a) has been replaced by "wanted" in sentence (b). But there is a difference in interpretation, which goes further than just changing the verb. In sentence (a) most native speakers would interpret "they" as to refer to the policemen, whereas in sentence (b) "they" refers to the demonstrating students.

The reason is obvious: we have a common understanding of the roles of policemen and students in society (at least in our part of the world), and there policemen are supposed to protect us against dramatic changes in our way of living, such as revolutions, whereas students have the reputation to be amongst the first ones to adopt a critical attitude, and to advocate changes.

No human translator would ever have any difficulty in assigning the correct interpretation and translation to sentences (a) and (b), whereas it is difficult to imagine how one could set up an MT system in such a way that it would make the right choice here. Linguistic knowledge, however deep, will not give any clue, and the sort of knowledge that is needed, is very hard to encode and apply in any systematic way. Some may want to argue that there is no need for the MT system to resolve the anaphoric ambiguity here, as e.g. translation into French would not depend on the choice made here, but it is easy to see that if we replace "policemen" by "police women" there is no way to get around the problem. The necessary conclusion is that however far we get in building MT systems, we will never be able to guarantee that we get it right.

1.2. The question

Answering the question we asked above turns out to bring us in a slightly difficult position, so it might be wiser to ask ourselves whether the

question itself is a reasonable one to ask. Traditionally, the objective of any MT effort has been to provide a possibly perfect, cheaper and faster imitation of the human translator. Success was measured on the basis of the extent to which this simulation was successful, and as we all know, this success has never been very impressive.

As an alternative way of looking at things, we can ask ourselves whether the simulation of the human translator is a meaningful purpose in itself. And the answer is a clear no. Although it may be an interesting intellectual challenge, the ultimate goal of any translation activity is to solve a communication problem, where two parties who want to exchange information happen to speak different languages. From this point of view we can describe the purpose of a MT system as a solution for a communication problem. This allows us to define the notion “success of a translation system” as the degree to which the intended communication has been made successful.

2. A new metaphor

We propose to adopt a new metaphor, instead of the human translator, as the starting point for the construction and the use of MT systems. Our metaphor is "the traveling tourist and his host".

We all know from our travels abroad that communication in a different language community poses problems, which can sometimes be solved by one of the parties (e.g. one person decides to speak the other person’s language), or sometimes by cooperative action.

This leads us to the introduction of two types of tourists (or guests) and hosts: adaptive and unadaptive. The adaptive party is the party that accepts that communication across languages is problematic, and makes a special effort to facilitate communication. The unadaptive party is not prepared to compromise, and insists on using his own language the way he would do at home.

We can now make the following grid, based on the two parties’ adaptiveness, and investigate what sort of solutions MT or in a broader sense language and speech technology could offer:

adaptive host adaptive guest	adaptive host unadaptive guest
unadaptive host adaptive guest	unadaptive host unadaptive guest

3. A war or a series of battles

The traveling tourist metaphor breaks down the MT problem into at least four subproblems, the solution of which may require different types of approaches and technologies. It may look less elegant to have four different solutions to four different instantiations of the MT problem, rather than one magic formula that does all the work, but on the other hand one should realize that just like in e.g. transportation (where airplanes are faster than bikes, but more difficult to park in cities), in communication every situation is different, and may have different constraints and success criteria. Therefore, even within the four boxes of the grid one may have to further subdivide the problem into subproblems of a size and complexity that we can manage.

The MT problem should not be seen as a single war, aimed at solving the MT problem in one go, but rather as a series of local battles, each of which should serve to reduce, in specific situations, the communication problems caused by the fact that people speak different languages.

3.1. The adaptive host and the adaptive guest

If both parties are equally adaptive, the

chances to get the communication problem solved, are by far the greatest. A typical solution is for both parties to resort to a third language, known by both. The use of English as a lingua franca in tourism, international transportation and in science is a very well known example. It is important to note that, although neither party may have a perfect command of the foreign language, the communication is normally successful, even if it may be less efficient than it would have been between two native speakers of the language (although even there no 100% guarantee for success can be given).

There is still a role for language and speech technology here, such as language learning facilities, or electronic dictionaries and foreign language authoring aids.

If none of the parties has a good command of the common language, other facilities may come into play. An excellent example is offered by the Verbmobil project, where a German and a Japanese speaker use their own language to make meeting arrangements via the phone, and where English is used as an interlingua to sort out problems.

3.2. The adaptive guest and the unadaptive host
This is a very normal situation for e.g. tourists who speak a minority language, and who are traveling in a country where either the native language belongs to the majority languages, or where most people have received no foreign language education.

If the tourist has some managed to acquire some moderate knowledge of the language, he might still be able to communicate successfully, even if some of the details of the communication may escape him. In such a situation a small electronic dictionary might be of help. If the language gap is wider, more sophisticated facilities are needed. An interesting example is the DIPLOMAT system, designed to help UN soldiers in Bosnia to communicate with the local population, which helps asking questions in a foreign language, and translating the answers.

For electronic travel obvious (and already existing) facilities are e.g. a translating browser (even if the translations are not perfect, cf the way AltaVista makes use of the SYSTRAN translation system), or an information extractor, which extracts the key information from a foreign language text and presents it in the user's own language.

3.3. The unadaptive guest and the adaptive host.
Typically this applies to situations where mass tourism moves people who may have received no foreign language training to places where they can enjoy sun, sea, food and drinks, and who don't want to be confronted with the language barriers.

Hosts who want to please their guests have to be adaptive, e.g. by learning their language, or by employing staff with the right native language.

From a language and speech technology point of view, obvious facilities to support this are the use of controlled languages in authoring messages from host to guests (as long as the author sticks to the rules, error-free translation can be guaranteed), or by means of multilingual generation from tabular information (e.g. generating weather or avalanche reports in various languages from tables).

3.4. The unadaptive guest and the unadaptive host.
This is a fairly extreme situation, and it is clear that ideally one would have to resort to either a human translator or interpreter, or to Fully Automated High Quality Machine Translation -- if only it existed.

As a long term research topic, this should certainly be kept high on the agenda, together with the interpreting phone which would ideally allow me to speak my own language over the phone, and to be heard by my Japanese conversation partner as speaking fluent Japanese, in my own voice.

For the shorter term the best option seems to be to develop tools that increase the productivity of human translators, or the quality of their output. Typical examples are electronic term banks, on-line dictionaries and thesauri, or translation memories.

4. Concluding remarks

First of all it should be clear that MT is not one single simulation problem, but a complex of communication problems, each of which may require a different approach and solution. Each of the subproblems can (and even: should) be further subdivided.

The adaptivity grid we have presented here, can also help us in raising new questions, such as "is there a tool for an adaptive guest who wants to express himself in a foreign language unknown to him?". Note that an affirmative answer to such a question will not automatically imply that there will also be a market for the tool.

In the spirit of the above, it may not come as a surprise that for the future we would like to advocate the development of domain dependent, goal dependent, and discourse dependent MT facilities. In order to exploit the fact that these systems are constrained, it is important that they should not only be based on what we know about language (the rules, which express what we understand), but also on what we know about how language is used in specific communication contexts (i.e. statistical data which we can measure and extrapolate from).

3. Multi-modal Speech Synthesis with Applications

Professor Björn Granström
CTT - Center for Speech Technology
Speech, Music and Hearing
KTH, Stockholm, Sweden
(http://www.speech.kth.se)

1. Introduction

While speech synthesis implies only sound, there is much more in speech communication than just voice. Multi-modal speech synthesis is a very natural way of communication.

We believe that multi-modal speech synthesis has important advantages like increased intelligibility of synthetic speech and is quite important for application areas like:

- friendlier human-computer interfaces like, for example, animated talking agents
- language learning
- tools for the hard of hearing
- perceptual experiments
- games, toys, cartoons, film, etc.

2. Multi-modal Speech Synthesis

In order to realize multi-modal speech synthesis one needs to have "real" understandable faces. Different approaches can be used for multi-modal speech synthesis. One approach, for use in educational systems, is based on pre-recorded phrases where a sampled waveform is accompanied by a video. A different approach is based on the concatenation of sub-word units using LPC or PSOLA techniques, where natural speech is sliced up in smaller parts and then put together again. Then, one could use a "terminal-analog" model where a formant speech synthesizer is used in combination with a parametric 2D or 3D face model. Finally, articulatory models can be used for both speech and image synthesis.

2.1. Work at KTH

2.1.1. General Layout of the Text-to-Audiovisual Speech System

The approach we are following uses a synthetic parametric 3D face model including teeth and tongue, which is actually a mesh of 3D points and a skin built on top of it. This face is controlled by means of high-level articulatory parameters like the jaw opening, the lip rounding, the lip closing, the labio-dental occlusion etc. The mouth movements are controlled by rules in a text-to-speech synthesis system. Prosodic information is used to control the eyebrows, the gaze and the head motion.

The layout of the system is shown on FIGURE 1.

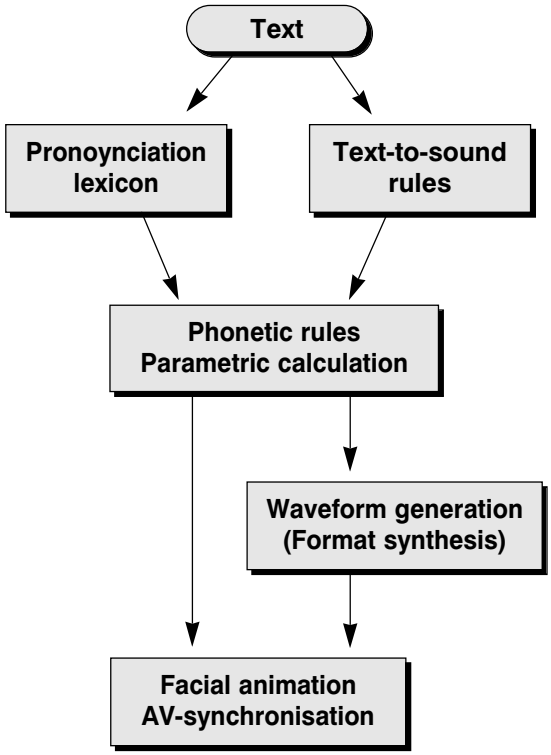


FIGURE 1. The layout of the text-to-audiovisual speech system.

2.1.2. Multi-modal Intonation: Eyebrow Movement vs. Intonation

The general idea is that eyebrow movement connects to intonation. There are papers in the literature arguing that there is a direct mapping from eyebrow movement to intonation. The following experiment were conducted at KTH:

- *No eyebrow movement.*
- *Eyebrow height is controlled by F0 curve.*
- *Eyebrow movement on highly focal accents.*
- *Eyebrow movement on first highly focal accent.*

In the first case the synthetic face looked rather unnatural. More naturalness and intelligibility were achieved in the other cases.

3. Animated Talking Agents: A New Paradigm for Human-Computer Interaction

One of the major application areas of multi-modal speech synthesis concerns the *animated talking agents*. Interfaces employing such agents could shift computing from the desktop-metaphor to the *person-metaphor*: you can “talk” to someone, or you can ask this “thing” or “person” or whatever to help you doing things like searching on the Internet. *Spoken dialog* is needed, of course, in these cases as well as *non-verbal communication* since there are many other things that can also be communicated. Additionally, this metaphor can increase the *believability*, but not necessarily the realism, of the system because the user surely want to know that this is *not* a person; it is just the computer.

3.1. Tasks of an Animated Agent

The main tasks of an animated talking agent can be summarized as follows:

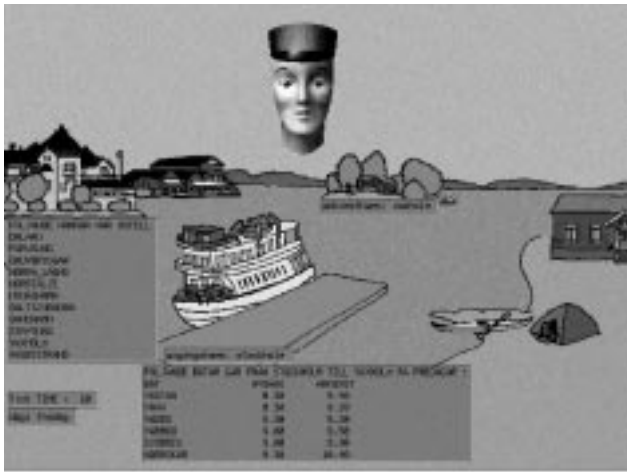
- to provide *intelligible synthetic speech*.
- to indicate *emphasis* and *focus* in utterances
- to support *turn-taking* since it is always important to know, for example, when user intervention is required.
- to provide *spatial references* by pointing, gazing, etc. so as to build a link between spoken and graphics modalities.
- to provide *non-verbal back-channeling*.
- to appropriately indicate the *system’s internal state*.

3.2. Related Projects

The animated agent technology has been

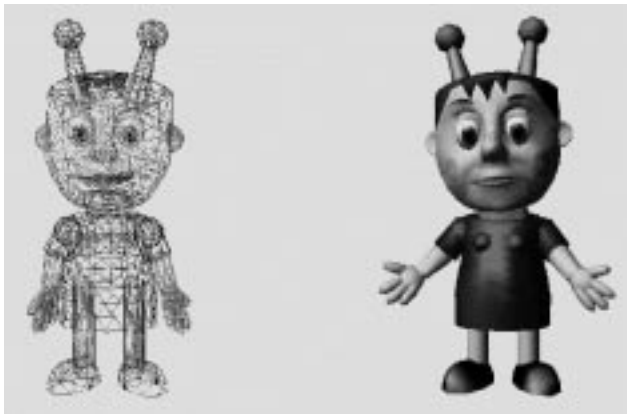
deployed in two projects, namely the Waxholm project and the Olga project.

3.2.1. The Waxholm Project



The aim of this project was to develop a tourist information application about Stockholm archipelago. The system used mixed initiative dialogues. It employed speech recognition for the input and multimodal speech synthesis and graphics for the output. The synthetic face was directly used for spatial references.

3.2.2. The Olga Project



Olga, is an antromorphic agent that can assist the user by providing consumer information. It combines speech, gesture, facial expressions and 2D graphics. It employs multimodal speech synthesis techniques as well as template-based gestures. The agent was chosen to be cartoon-like so as to keep user expectations low. Olga, is the result of the collaboration between four departments, namely CTT, DID, SU, and SICS.

4. The Teleface Project[1]

The Teleface project at the Department of Speech, Music and Hearing, KTH, aims at evaluating the possibilities of using synthetic visual speech in tools for hearing-impaired people. The project will include an effort to implement a demonstrator prototype of a telephone communication aid for the hard of hearing. This device will generate a synthetic face that articulates in synchrony with the telephone speech using only the information contained in the telephone speech signal, see FIGURE 2.

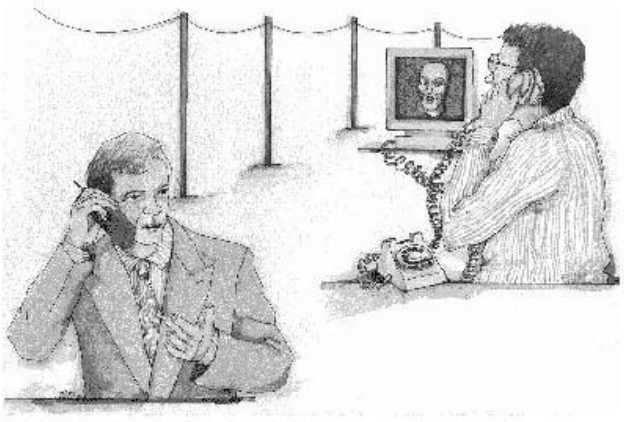


FIGURE 2. The intended teleface demonstrator

The project involves work in several areas: multimodal speech synthesis, audiovisual speech intelligibility studies, visual speech measurement and speech signal analysis. We are currently performing a suite of audio-visual intelligibility tests, where cross-combinations of synthetic and natural voices and faces are evaluated at varying signal-to-noise ratios. Subjects are hearing impaired as well as normal hearing people.

[1] *The text and figures in the Teleface section were taken by the web pages of KTH, and can be found in: http://www.speech.kth.se/teleface/ and in related pages linked to it.*

4.1. Method

4.1.1. Subjects

In the first test series the subjects were 18 fourth-year-students in engineering at KTH. The test was made as a part of a mandatory laboration in the Speech Communication course given by the department. A screening test was being used to check that all the subjects had a normal hearing level.

4.1.2. Stimuli

In the first test series we used lists consisting of VCV-words with 17 Swedish consonants /b, d, g, p, t, k, s, sj, tj, f, v, m, n, ng, j, l, r/, in symmetric context with the vowels /a/ and /o/. When performing tests on normal hearing persons using this material, audio has been degraded by adding white noise. The signal-to-noise-ratio in these tests was 3 dB. Each subject performed tests with 8 different combinations of voices and faces.

4.1.3.Procedure

The tests are being performed in a computer-based test environment that gives us the opportunity to play video sequences of the faces together with sound files of the voices. In the test series we can therefore evaluate the intelligibility of different audio-visual combinations. A monitor is used for presenting the visual stimuli and a loudspeaker for the audio. A forced choice response for the VCV-words is made using a mouse on a computer screen presenting all possible consonants. There is no time limit for the response.

4.1.4. Results

Data from the tests were analyzed using confusion matrices and feature analysis. Overall results are shown in figure 1. Adding a synthetic face to a natural male voice increases correct responses from 63% to 70%. Corresponding result for adding a natural face is 76%. Synthetic male voice gave 31% correct responses compared to 45% with a synthetic face added.

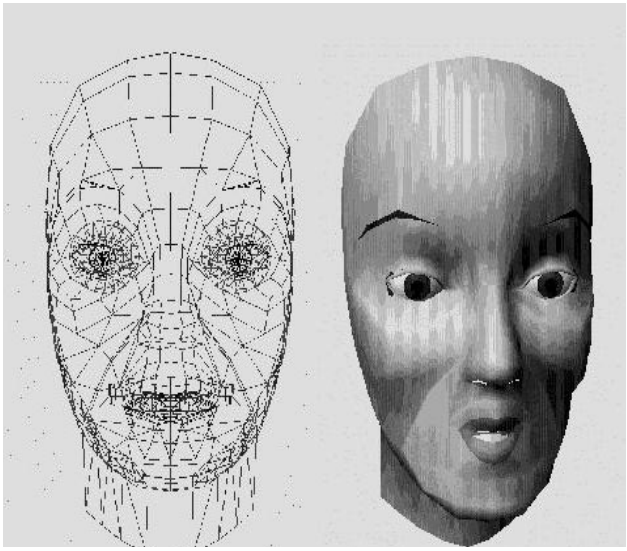


FIGURE 3. Animations: Holger

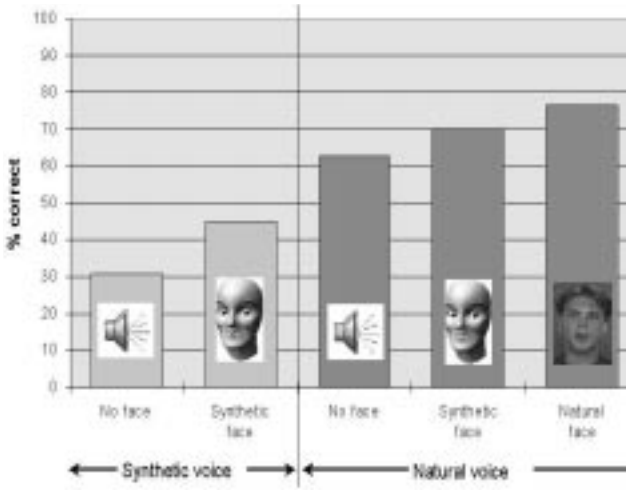


FIGURE 4. Results from intelligibility tests.
Number of correct responses (in %). Average for 18 subjects.

4. Language Technologies: Progress and Prospects

Mr. Roberto Cencioni
European Commission
DG XIII-E-5
roberto.cencioni@lux.dg13.cec.be

At the seminar held by the Greek Human Network for LE on the 17th of February 1997, Mr. Cencioni's speech focused on the progress and prospects for Language Engineering. The main points from this speech are the following (text followed by selected slides):

During the 4th Framework Programme and in the framework of the LE Programme for Language Engineering, three key technology lines were followed:

- Full multilinguality-Ability to work and communicate in one's language
- Natural interfaces based on natural speech rather than the keyboard
- Active content-Information retrieval and extraction, filtering, clustering and delivery

The Programme focused on the following application areas:

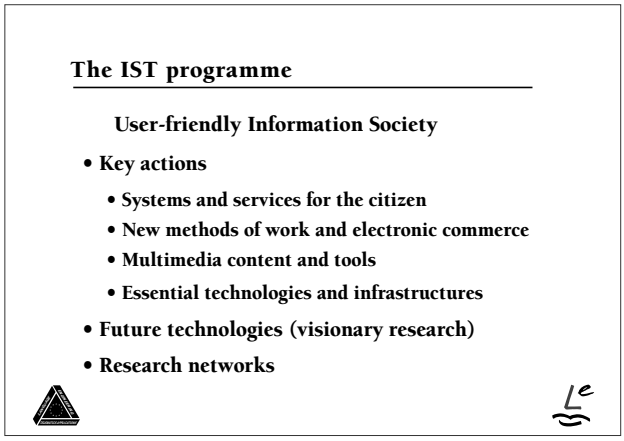
- Business information services
- Services of public interest
- International commerce
- Tele business
- Business (language) training

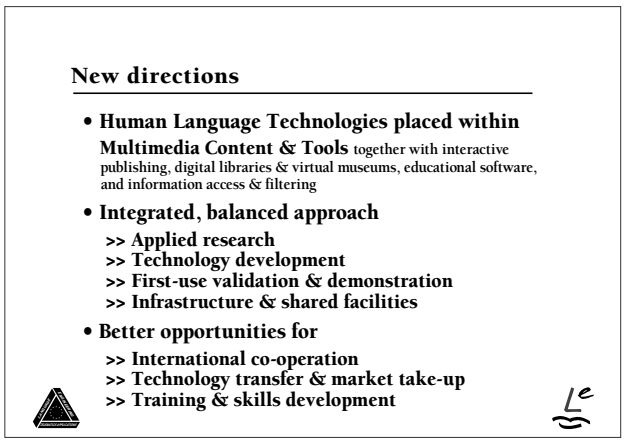
Apart from research institutes, other Programme participants included suppliers of Language Technology products and services as well as users from the private and public sector. LE was the largest dedicated R&D Programme (ECU 110 million in 1992-1998) and won recognition from both the industrial and political sectors. Significant commercial developments stemming from the LE Programme include: multilingual HTML editors & Web browsers, online search and translation services, dictation systems, translation memories etc.

The outlook for 1999-2002 includes:

- An integrated approach which will avoid the gaps between research and market take-up
- Building on strengths and concentrating on global challenges
- Increased reactiveness to industry-led demonstration
- Ease of transfer of key technologies to multiple languages through new forms of partnership
- Support of shared networks and facilities

More information regarding prospects for the 5th Framework Programme are outlined in the following slides.





Key dimentions

- Multi-purpose R&D (from FP3)
- User driven validation (from FP4)
- Transfer to other languages (new)
- R&D infrastructure (enhanced)

One size does NOT fit all!

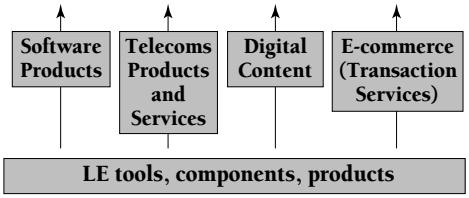
- "Replication" carried out in other parts of the programme

Intertwined drivers

- Multinational (global) business
- Multilingual communication
- Multimedia (digital) content
- Multimodel interfaces

Delivery channels

Language intensive processes and transactions within and between organisations



Conclusion

- Rapid acceleration in recent years, strong international competition
- Good prospects for future R&D despite severe budget constraints
- One focal point within IST, and many "satellites" (eg human-centred interfaces)
- Integrated chain and...
  - >> Demonstration & take-up
  - >> Links with national and EU programmes
  - >> Multidisciplinarity and skills development
  - >> New forms of partnership with users & integrators

human language technologies

human language technologies

dg 13 – e 5
http://www.echo.lu/langeng/

the big scene

global village: global competition	globalisation
computing, telecoms & media: new markets	convergence
ICT as commodity, mass markets, full participation	human centred

objective

empower people

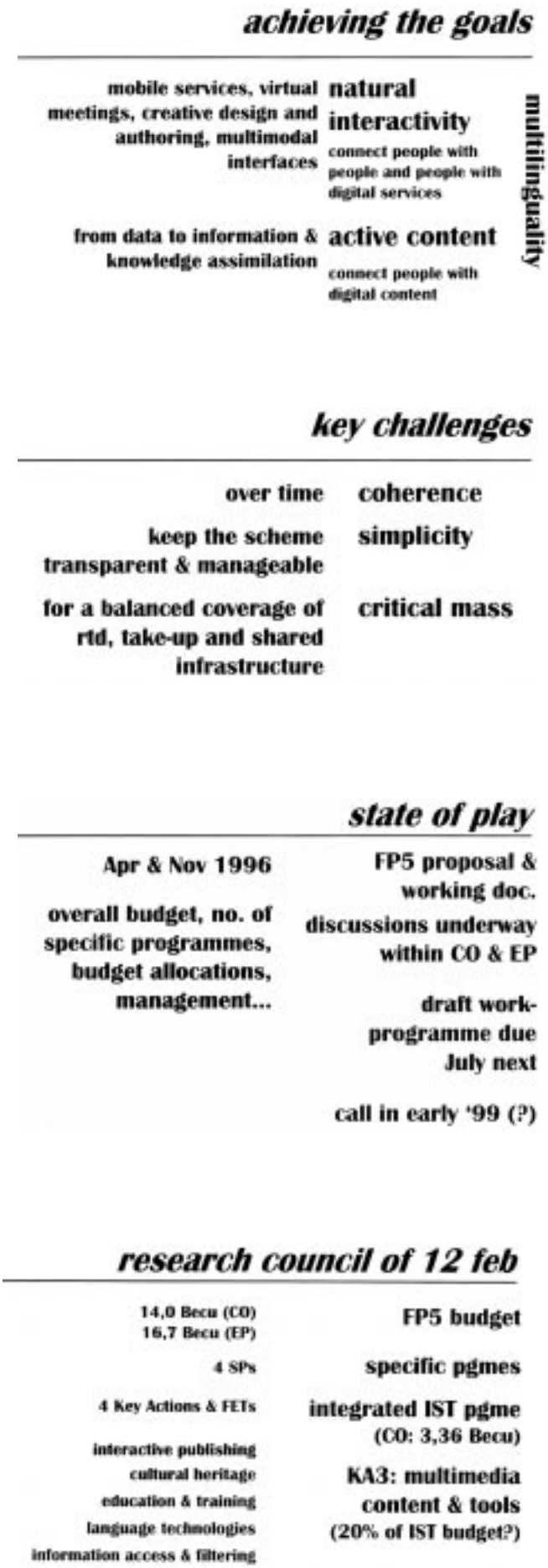
facilitate new business

building on european strengths

enable the information society

achieving the goals

always	full multilinguality
mobility & ubiquity, new models of socialising, life long learning	natural interactivity
efficiency & quality, equal opportunity	active content
new production & business models	natural interactivity & active content



II. Περιλήψεις εισηγήσεων ημερίδας με θέμα "Ηλεκτρονική Λεξικογραφία" / *Workshop on "Electronic Lexicography"*

1. Development of Bilingual Electronic Dictionaries at the Wire Communications Laboratory of the University of Patras.

H. Coutsogeorgopoulos, Researcher and Dr. E. Dermatas
Wire Communications Laboratory (WCL).
Department of Electrical Engineering,
University of Patras, Hellas
E-mail:dermatas@george.wcl2.ee.upatras.gr,
coutsoge@wlc.ee.upatras.gr

Abstract

This paper presents the dictionaries that were developed by the WCL in collaboration with Harper Collins Bilingual Dictionaries: an English-Greek Dictionary and a Greek-Framework. The role of the WCL was the translation of the English-Greek Dictionary into Greek, and the creation of the Greek Framework which serves as a base for translation into English and other languages. Both dictionaries were compiled by referring to electronic corpora. The English Framework was based on the corpus known as the "Bank of English", a database of the English language containing 300m. words from a very wide range of texts. The Greek Framework was based on a Corpus of Greek Texts that was created by the WCL according to the structure and specifications of the English corpus and contains over 50m. words. Its range of material includes newspapers and magazines, scientific and technical texts, advertising and tourist materials, books on literature, Minutes of the Greek Parliament, and generally texts that could reflect as much as possible the Greek language that is used today. Each dictionary contains over 40.000 lemmas and phrases printed in bold type, pronunciation in phonetic transcription, and

grammatical category for every lemma. Over 25.000 usage examples, thousands of collocates and synonyms, taken from the text corpora, show how to use the translation and find the meanings in context. Very frequent function words, as well as words with many uses and meanings that require detailed treatment, are marked out by the heading "Keyword", while a number of special indicators give more information of how a word is used, e.g. for colloquial or formal language (register) or words belonging to a special field. All lemmas, phrases and usage examples of the English-Greek Dictionary are translated into Greek. The Greek Framework contains in addition morphological information (endings, irregular verb formations, literary word-forms) with reference to Tables of Inflectional Paradigms for all declined lemmas, as well as alternative forms for spelling or morphological variants.

Ανάπτυξη Δίγλωσσων Ηλεκτρονικών Λεξικών στο Εργαστήριο Ενσύρματης Τηλεπικοινωνίας (ΕΕΤ) του Πανεπιστημίου Πατρών

Ελένη Κουτσογεωργοπούλου, Ερευνήτρια και
Δρ. Ευάγγελος Δερματάς
Εργαστήριο Ενσύρματης Τηλεπικοινωνίας (ΕΕΤ)
Τμήμα Ηλεκτρολόγων Μηχανικών & Τεχνολογίας Υπολογιστών
Πανεπιστήμιο Πατρών Πάτρα 261 10

Η ομάδα Γλωσσικής Επεξεργασίας του Εργαστηρίου Ενσύρματης Τηλεπικοινωνίας (ΕΕΤ) του τμήματος Ηλεκτρολόγων Μηχανικών του Πανεπιστημίου Πατρών έχει αναπτύξει μακροχρόνια δραστηριότητα στην ηλεκτρονική λεξικογραφία. Στην παρούσα εργασία περιγράφονται συνοπτικά το *Αγγλο-Ελληνικό Λεξικό* που δημιουργήθηκε σε συνεργασία με τον εκδοτικό οίκο Harper Collins Bilingual, και το Ελληνικό *"Λεξικό-Πλαίσιο"*. Το Ελληνικό *"Λεξικό-Πλαίσιο"* αποτελεί βάση για την δημιουργία δίγλωσσων λεξικών από την οποία ο εκδοτικός οίκος Harper Collins Bilingual θα δημιουργήσει το Ελληνο-Αγγλικό Λεξικό. Ο ρόλος του ΕΕΤ στο πρόγραμμα αυτό ήταν η μετάφραση του άγγλο-ελληνικού λεξικού στην Ελληνική γλώσσα και η σύνταξη του *Ελληνικού "Λεξικού-Πλαισίου"*.

Δεδομένου ότι η ανάπτυξη των λεξικών βασίζεται στην αντίληψη ότι η γλώσσα αποτυπώνεται όπως χρησιμοποιείται στην σημερινή φυσική ροή της, σε συγχρονικό επίπεδο, η δημιουργία τους επιτυγχάνεται με πρόσβαση και επεξεργασία ηλεκτρονικού σώματος κειμένων.

Σώμα Κειμένων (Corpus)
Εργαλεία πρόσβασης των λεξικογράφων σε παραδείγματα χρήσεως, σε συμφραστικό περιβάλλον (collocates) και στατιστικής επεξεργασίας αποτελούν την πιο σημαντική πρωτοτυπία στην ανάπτυξη των λεξικών αυτών.

Το Αγγλικό μέρος του λεξικού βασίζεται στο σώμα κειμένων της Αγγλικής το γνωστό ως "Bank of English" το οποίο έχει σήμερα περίπου 300 εκ. λέξεις και δίνει την δυνατότητα μεθοδικής επιλογής α) λέξεων, β) του περιβάλλοντος στο οποίο αυτές χρησιμοποιούνται και γ) σωρείας παραδειγμάτων χρήσεως και εκφράσεων. Με τον τρόπο αυτό καθορίζονται ακριβέστερα οι λεπτές αποχρώσεις των σημασιών.

Το Σώμα Ελληνικών Κειμένων που έχει δημιουργηθεί στο ΕΕΤ είναι προς το παρόν συλλογή γραπτών κειμένων της Ελληνικής γλώσσας και περιέχει 50εκ. λέξεις από κείμενα κυρίως της τελευταίας δεκαετίας, εκτός ορισμένων γνωστών λογοτεχνικών έργων που είναι παλαιότερα. Η οργάνωση του περιεχομένου έγινε με βάση την δομή και τις προδιαγραφές του Αγγλικού corpus και της σχετικής βιβλιογραφίας ([1]). Συγκεκριμένα περιλαμβάνει είδη κειμένων (εφημερίδες και περιοδικά, διαφημιστικά και τουριστικά έντυπα, αλληλογραφία, επιστημονικά και τεχνικά κείμενα, θεατρικά έργα, παιδική λογοτεχνία, τα

πρακτικά της Βουλής), ώστε να αντιπροσωπεύουν κατά το δυνατόν το Ελληνικό λεξιλόγιο. Αυτά συμμετέχουν στο σώμα κειμένων με συγκεκριμένες αναλογίες και ποσοστά ανά είδος. (Πίνακας Α)

	December 4, 1995 30351804.00 Words	Total	Target %	Target 100.000.000 Words	%
1. Press					**(30)**
Newspapers	4,565,448	15.04	30.44	15,000,000	15
Magazines	420,263	1.38	2.8	15,000,000	15
2. Literature					**(32)**
Fiction	5,859,351	19.3	29.3	20,000,000	20
Theatre	111,481	0.3	2.23	5,000,000	5
Poetry	293,676	0.97	14.68	2,000,000	2
Children's	1,727,805	5.56	34.56	5,000,000	5
3. Essays					**(15)**
Technical	460,313	1.52	11.51	4,000,000	4
Scientific	5,636,728	18.57	70.46	8,000,000	8
Reviews	817,560	2.69	81.76	1,000,000	1
Biographies	707,892	2.33	35.39	2,000,000	2
4. Law & administration					**(13)**
Law	4,268,426	14.06	42.68	10,000,000	10
Commercial	413,617	1.36	13.79	3,000,000	3
5. Current					**(2)**
Adverts	46,720	0.15	6.67	700,000	0.7
Tourism	0	0	0	700,000	0.7
Information	415,780	1.37	69.3	600,000	0.6
6. Others					**(4)**
Educational	133,877	0.44	4.46	3,000,000	3
Guide books	51,840	0.17	5.18	1,000,000	1
7. Oral					**(4)**
Parliament	4,421,027	14.57	147.37	3,000,000	3
Speeches	0	0	0	1,000,000	1
Total	**30,351,804**		**30.35**	**100,000,000**	**100**

Πίνακας Α. Ελληνικό Λεξικό - Σώμα κειμένων - Ποσοστά ανά είδος κειμένου

Το Αγγλο-Ελληνικό Λεξικό

Το λεξικό περιέχει περίπου 40.000 λήμματα στα οποία περιλαμβάνονται κύρια ονόματα, συντομογραφίες και αρκτικόλεξα. Ιδιαίτερη έμφαση δίδεται στο λεξιλόγιο τεχνικών και επιστημονικών όρων, καθώς και σε όρους από τον χώρο του εμπορίου και των επιχειρήσεων, ώστε το λεξικό να είναι χρήσιμο τόσον σε εξειδικευμένες ανάγκες εργασίας όσον και στην γενικότερη εκμάθηση της γλώσσας.

Για κάθε λήμμα παρέχεται:

- Φωνητική γραφή του λήμματος.
- Γραμματικές πληροφορίες.
- Μετάφραση στην Ελληνική.
- Παραδείγματα χρήσεως και μετάφραση στα Ελληνικά.
- Φράσεις: ενότητες λέξεων με συγκεκριμένη συντακτική δομή και συγκεκριμένη χρήση, π.χ. *as a matter of fact, Pap test,* με παραδείγματα χρήσεως και μετάφραση.
- Συμφραστικό περιβάλλον λημμάτων και των τύπων τους όπως βρίσκονται στα κείμενα.

Πρόκειται για δίγλωσσο λεξικό απευθυνόμενο σε ομιλητές δύο γλωσσών, στο οποίο η μετάφραση δεν δίνεται με ορισμούς αλλά με αντιστοιχίες λημμάτων από την μία γλώσσα στην άλλη, καθώς και με όλες τις δυνατές πληροφορίες (συμφραστικό περιβάλλον, παραδείγματα χρήσεως) ώστε η μετάφραση να είναι όσο το δυνατόν πιο ακριβής.

Κατά την μέθοδο που ακολουθούν οι Collins χρησιμοποιούνται μεταφραστές και λεξικογράφοι με μητρική γλώσσα την Αγγλική για το αγγλικό μέρος και με μητρική γλώσσα την Ελληνική για την ελληνικό μέρος. Το ΕΕΤ οργάνωσε ομάδα Ελλήνων μεταφραστών, τους εξεπαίδευσε στην ηλεκτρονική λεξικογραφία και επέβλεψε την μετάφραση στα Ελληνικά.

Στα ιδιαίτερα χαρακτηριστικά του λεξικού περι-

λαμβάνονται:

- Λέξεις κλειδιά (Keywords), π.χ. *of, get* για λέξεις με πολλές σημασίες και μεγάλη συχνότητα. Η λέξη KEYWORD εμφανίζεται σε ξεχωριστό πλαίσιο πάνω από το λήμμα ώστε να ξεχωρίζει από το υπόλοιπο κείμενο.

- Δείκτες: οι οποίοι δίνουν πληροφορίες ως προς την χρήση του λήμματος και του περιβάλλοντος στο οποίο χρησιμοποιείται, το υφολογικό επίπεδο, κ.α. Παραδείγματος χάριν, στο λήμμα *engine* που μεταφράζεται "μηχανή" δίδεται μέσα σε παρένθεση το (AUT, RAIL) ώστε ο χρήστης ξέρει ότι δεν χρησιμοποιείται αυτή η λέξη στο "φωτογραφική μηχανή". Στο λήμμα *teach* ο δείκτης (fig) δείχνει ότι η λέξη χρησιμοποιείται μεταφορικά

Η παρουσίαση του υλικού έχει ως κύριο στόχο να διευκολύνει τον χρήστη στον εντοπισμό της πληροφορίας που χρειάζεται. Ενδεικτικώς αναφέρονται τα εξής:

- λήμματα, παραδείγματα, φράσεις, και παράγωγα είναι τυπωμένα με έντονους χαρακτήρες,

- χρησιμοποιούνται ενδεικτικά σήματα για φράσεις, παραδείγματα και περιφραστικά ρήματα,

- οι φράσεις καταχωρίζονται σε ξεχωριστή γραμμή,

- τα παράγωγα καταχωρίζονται ως λήμματα ώστε να μη γίνεται το άρθρο μεγάλο και κουραστικό για το χρήστη.

Ελληνικό "Λεξικό-Πλαίσιο"
(Έλληνο-αγγλικό Λεξικό)

Για την σύνταξη του Ελληνικού Πλαισίου ακολουθήθηκε η ίδια μέθοδος με προδιαγραφές προσαρμοσμένες στην Ελληνική γλώσσα. Ο όρος "πλαίσιο" αντιστοιχεί στον αγγλικό όρο "framework" και είναι το Ελληνικό μέρος του δίγλωσσου λεξικού χωρίς την μετάφραση. Αποτελεί βάση για πολλαπλές χρήσεις, μεταξύ των οποίων να μεταφραστεί σε άλλες γλώσσες ή να εξαχθούν από αυτό μικρότερα λεξικά.

Επιλογή λημμάτων

Το λεξικό περιέχει άνω των 32.000 λημμάτων, 8.000 φράσεις και 25.000 παραδείγματα, στα οποία περιλαμβάνονται κύρια ονόματα, συντομογραφίες, αρκτικόλεξα και όλα τα χαρακτηριστικά που έχει και το Άγγλο-Ελληνικό λεξικό. Το λημματολόγιο δημιουργήθηκε από την στατιστική επεξεργασία των λέξεων του Σώματος Ελληνικών Κειμένων, και συμπληρώθηκε από υπάρχον λεξιλόγιο του ΕΕΤ, Ελληνικά μονόγλωσσα λεξικά και ειδικά λεξιλόγια.

Διαμορφώθηκαν λεπτομερή κριτήρια για την επιλογή των λημμάτων εκ των οποίων τα πιο χαρακτηριστικά αναφέρονται στο:

- πόσο απαραίτητη είναι η λέξη: π.χ. *έχω, νερό,*

- ποια συχνότητα εμφάνισης στο Σώμα Κειμένων,

- αν μπορεί να αντικατασταθεί από άλλη,

- αν είναι δόκιμη,

- πόσο χρήσιμη είναι στον υποτιθέμενο χρήστη.

Για να περιοριστεί κατά το δυνατόν η παράληψη λέξεων του βασικού λεξιλογίου καθορίστηκαν "κλειστές" κατηγορίες (π.χ. προθέσεις, ζώδια, εποχές) ώστε να μπορούν να ελεγχθούν με συνέπεια, αναλύθηκαν χωριστά και μετά ενώθηκαν με το υπόλοιπο λεξικό.

Ακολουθεί δείγμα λεξιλογίου με τις συχνότητες που βρέθηκαν στα κείμενα. (βλ. πίνακα Β). Μολονότι η συχνότητα εμφανίσεως των λέξεων στα κείμενα έχει συμβουλευτικό χαρακτήρα, είναι ενδεικτική για την επιλογή των συχνότερα εμφανιζομένων λημμάτων, για την ύπαρξη ή μη ορθο-

γραφικών ή άλλων εναλλακτικών τύπων καθώς και για την προτεραιότητα του ενός έναντι του άλλου. Επίσης, οι πίνακες συχνοτήτων παρέχουν σαφείς ενδείξεις για την ύπαρξη και χρήση λογίων καθώς και εκλαϊκευμένων ή ποιητικών λέξεων, ή τύπων τέτοιων λέξεων οποίοι επιλέγονται ως λήμματα στο λεξικό.

Αββάς	14	Αυγές	9
Αβγά	117	Αυγή	471
Αβγού	8	Αυγής	114
Αβγό	51	Αυγού	24
Αβγών	8	Αυγό	132
Αβεβαιότης	5	Αυγών	55
Αβεβαιότητα	269	Αυστηρό	177
Αβεβαιότητας	235	αυστηρό,	9
αβεβαιότητες	36	Αυστηρός	122
αβεβαιότητος	11	Αυτό	71037
Αβεβαιοτήτων	10	αυτό,	495
αβλαβές	6	αυτό:	5
αβλαβή	16	αυτό·	18
αβλαβής	19	Αυτός	14468
αβλαβούς	11	αυτός:	10
αβλεψία	11	Αυτόχθον	6
αβρόχοις	11	Αυτόχθονα	6
αδεία	10	Αυτόχθονες	19
αδείας	319	αυτόχθονων	8
αδειών	191	αυτόχθων	5
αδιακρίτως	45	αυτωνών	13
αδιαλείπτως	19	αυτώ	100
αυγά	262	αυτών	8853

Πίνακας Β. Συχνότητες εμφάνισης λέξεων στο Σώμα Ελληνικών Κειμένων

Ανάλυση λημμάτων

Η ανάλυση των λημμάτων είναι συστηματικά κωδικοποιημένη. Κάθε μορφολογική, σημασιολογική, και υφολογική ή άλλη πληροφορία φέρει έναν συγκεκριμένο κωδικό ώστε να γίνεται εύκολα ο χειρισμός και η ανάκτηση των δεδομένων, π.χ. *<POSP> ουσ* (ιαστικό)

Η ανάλυση περιέχει:

- Φωνητική γραφή του λήμματος.
- Γραμματικές/μορφολογικές πληροφορίες.
- Παραπομπή σε κλιτικό παράδειγμα.
- Σημασίες, συνώνυμα, συμφραστικό περιβάλλον.
- Φράσεις (εκφράσεις).
- Παραδείγματα χρήσεως από το κείμενο.
- Εναλλακτικούς τύπους.
- Υφολογικές διακρίσεις, π.χ. *μτφ., επισ., προφ.*
- Σήμανση ειδικού λεξιλογίου, π.χ. *ΨΥΧ., ΝΟΜ.*
- Σήμανση λημμάτων για εξαγωγή μικρότερων λεξικών, π.χ. *COMMON* όπου περιέχονται οι πιο κοινές και πιο συχνές λέξεις.

Λεξιθήρας: Υπολογιστικό Σύστημα Λεξικογραφίας βασισμένο σε σώμα κειμένων

Η ύπαρξη συστήματος διαχείρισης και επεξεργασίας μεγάλου πλήθους κειμένων φυσικής γλώσσας αποτελεί την βάση κατασκευής αποτελεσματικών και αποδεκτού χρόνου απόκρισης λεξικογραφικών εργαλείων. Ο Λεξιθήρας αποτελεί μία βάση δεδομένων ειδικού σκοπού προσανατολισμένη στην ελαχιστοποίηση του χρόνου απόκρισης των εργαλείων επεξεργασίας των κειμένων.

Οι βασικές λειτουργίες του Λεξιθήρα είναι οι ακόλουθες:

- *Βαθμίδα δημιουργίας βάσης δεδομένων κειμένων.* Η δομή της βάσης δεδομένων κατασκευάζεται με ρυθμό ενσωμάτωσης 8.3 εκατ. λέξεων ανά ώρα επεξεργασίας.

- *Λημματοποίηση του λεξικού των κειμένων.* Με την βοήθεια στοχαστικού μοντέλου και έναν αριθμό δόκιμων καταλήξεων, οι λέξεις του κειμένου ομαδοποιούνται σε κατηγορίες λέξεων που ανήκουν στο ίδιο λήμμα. Οι λεξικογράφοι πραγματοποιούν διορθώσεις στις εκτιμήσεις του στοχαστικού μοντέλου.

- *Συχνότητα εμφάνισης λημμάτων στα κείμενα.* Η πληροφορία αυτή δημιουργείται αυτόματα από τον Λεξιθήρα και χρησιμοποιείται ως βοηθητικό κριτήριο επιλογής των λημμάτων που θα επιλεγούν.

- *Διάκριση και καθορισμός των εννοιών - Επιλογή των παραδειγμάτων.*
  Το εργαλείο αυτό επιτρέπει στον λεξικογράφο να εργάζεται στην ανάπτυξη του λεξικού με την βοήθεια ενσωματωμένου επεξεργαστή κειμένου ο οποίος διαθέτει την δυνατότητα να αναλύει όλους τους τύπους του λήμματος που απαντούν στα κείμενα παρουσιάζοντας την συχνότητα εμφανίσεώς τους, και απεικονίζοντας όλα τα παραδείγματα χρήσεως που υπάρχουν στα κείμενα. Η πλέον χρήσιμη λειτουργία του εργαλείου εντοπίζεται στο γεγονός της αυτόματης μεταφοράς παραδειγμάτων χρήσης που επιλέγονται από τους λεξικογράφους στον επεξεργαστή κειμένου, δυνατότητα που αυξάνει σημαντικά την παραγωγικότητα τους και μειώνει την πιθανότητα λάθους.

- *Στατιστική επεξεργασία λέξεων.*
  Με επιλογή λέξεων δίνεται η δυνατότητα να υπολογίζονται στατιστικά μεγέθη όπως η από κοινού πληροφορία (mutual information) και η συχνότητα εμφάνισης ζευγών λέξεων σε γειτονικό περιβάλλον (concordance).
  Τα στατιστικά μεγέθη που υπολογίζονται είναι πολύ χρήσιμα στον εντοπισμό συχνά χρησιμοποιούμενων συμφραστικών λέξεων.

Ενδεικτική Βιβλιογραφία

1. Sinclair, J. (1991).
   Corpus, Concordance, Collocation.
   Oxford: Oxford University Press

2. Baker, M., Frances, G. and
   Tognini-Bonelli, E. (Eds) (1993).
   Text and Technology.
   In Honour of John Sinclair.
   Amsterdam: John Benjamins Publishing Co.

3. Collins (1992).
   COBUILD English Language Dictionary.
   London: Harper Collins Publishers

4. Webster's (1985)
   Ninth New Collegiate Dictionary.
   Springfield. MA: Merriam-Webster Inc., Publishers

5. Collins Bilingual Reference. (1993).
   Bilingual Dictionary Compilation Guidelines.
   Revised and adapted for GREEK

6. Τριανταφυλλίδη,
   Μ. Νεοελληνική Γραμματική Αθήνα: ΟΕΔΒ

7. Τζαρτζάνου, Α. Α.
   Γραμματική της Νέα Ελληνικής Γλώσσης.
   (Της Απλής Καθαρευούσης). Έκδοσις Β'.
   Αθήναι: Κ. Κακουλίδη

8. Βοσταντζόγλου, Θεολ. (1990):
   ΑΝΤΙΛΕΞΙΚΟΝ ή Ονομαστικόν της Νεοελληνικής Γλώσσης. Αθήναι

9. Τεγόπουλος - Φυτράκης. (1993).
   Ελληνικό Λεξικό.
   Αθήνα: Αρμονία Α.Ε.

10. Δημητράκου, Δ.
    Νέον Ορθογραφικόν Ερμηνευτικόν Λεξικόν.
    Αθήνα: Χρ. Γιοβάνη

11. Δορμπαράκης, Παν. Χ. (1993).
    Ετυμολογικό Ερμηνευτικό Λεξικό της Νεοελληνικής.
    Αθήνα: Σπουδή.

12. Μαρκαντωνάτος, Γ. (1992).
    ΛΕΞΙΚΟ Αρχαίων, Βυζαντινών και Λογίων Φράσεων της Νεας Ελληνικής.
    Αθήνα:Gutenberg

13. Πρωίας.
    Σύγχρονον Ορθογραφικόν Ερμηνευτικόν Λεξικόν της Ελληνικής Γλώσσης.
    Αθήναι: Σταμ. Π. Δημητρακου.

2. Multilingual Terminology Management for Distributed Digital Collections

Dr. Martin Doerr
Institute of Computer Science,
Foundation for Research and Technology Hellas
Heraklion-Crete, Greece
Vassilika Vouton, P.O.Box1385,
GR 71110 Heraklion,Crete,Greece
Email: martin@ics.forth.gr

Introduction

One can roughly separate the problem of heterogeneity in distributed digital collections into a structural one - the differences in the schemata or document structures, and a terminological one - the differences in data values, which may refer to the same real items. There are two kinds of references, those to actual things, **"instances"**, as "me, my house, my computer, my publications", and those to groups of things, either **concepts**, as "researchers, buildings, PCs, essays, roads", or **non-discrete sets**, as areas on the surface of earth. If one accesses a series of electronic collection, and wishes to retrieve data about certain things, there is the permanent problem of the identity of things referred by terms. Each social group, be it a scientific discipline or a nation, uses other terms, and even individuals may differ in their use of terms. This problem is tackled by the use of so-called "authorities", which define and standardize terminology of a certain group and domain for consistent use in documentation and retrieval. This works quite well on isolated databases, but is still insufficient for larger federations of databases. The hope to create a "world-wide" authority can be fairly regarded as an illusion.

Terminology Support

Authorities try to solve two problems: The identification of a notion, and the definition of a concept. For identification, linguistic expressions, so-called "terms", i.e. possible or preferred **noun phrases** or **names** are associated with a notion according to the practice of a social group. The notion in turn is described by attributes, as life-data of a person, free texts, geo-coordinates etc. The user may select due to these descriptions "good" terms, which can be used by the retrieval agent to match the notion behind with database records and occurrences in texts. For "instances", the translation problem is solved, when we gather the terms of all groups for each notion. See e.g. the *United List of Artist Names* from the Getty Information Institute.

For concepts, however, each group tends to have its own definitions, and items may be classified or referred by coarser or narrower concepts. Concepts are therefore correlated by equivalence and subsumption expressions in so-called "multilingual thesauri", which can be exploited by retrieval agents for matching data records about related items, but classified in foreign terms. As above, free texts (scope notes), images etc. support further the identification of a concept by a user. Hence the translation of concepts consists of an identification and a correlation problem of all concepts of all groups, which is obviously an open ended task, as continuously new concepts appear. Authorities, in particular thesauri, can be regarded and dealt with as knowledge bases, which comprise domain knowledge in form of terminological logic.

Current Situation

From the point of view of implementation and system integration, the current situation can be described as follows:

- Either a separate, not integrated thesaurus tool is used, or there is an idiosyncratic implementation of a thesaurus management within the local collection management system.
- Some libraries agree to use a foreign (typically English) thesaurus, as e.g. LCSH, ACM subject headings etc., thus giving poor support for the local language and any further specialization to local needs.

- Very few systems support automatic query term expansion, in the same or to other languages.
- Evolution of the thesaurus on an external tool and consistent migration of new or changed terms **into** a set of local collection management system and into other external thesaurus tools is typically not foreseen.

Hence valuable information remains inaccessible, and retrieved information is incomplete and inconsistent with the request, at least by far more than necessary.

The Architecture

This article describes a proposal for an architecture, which can render integrated terminology services on large federations of digital collections in a scalable and manageable way with similar quality as currently on some local systems. It builds on the experiences and system developments from several cooperations of the author [1], among which the AQUARELLE project (see e.g. http://aqua.inria.fr, [2]) is the furthest going in this direction. The equally important question of knowledge acquisition and the effective creation of thesaurus contents is deliberately not addressed here (see e.g. [3]).

For optimal results, the terms used for asset classification, in the search aid thesaurus and in the experts' terminology should be consistent. This led us to a three level architecture of components cooperating within an information access network:
(1) vocabularies in local databases,
(2) local thesaurus management systems of wider use and
(3) central term servers for retrieval.

For reasons of standardization of format and centralization of handling, we foresee a separate thesaurus manager to which the vocabularies of several local databases can be loaded, and in the sequence are organized as thesauri ("authorities") by some experts, following variations of the ISO2788 semantic structure. In addition, standard external vocabularies can be loaded to it. The authorities may be specific to one database, a user organization, or a whole language group.

The local vocabularies and terms that are already used for classification may need updating with the changes done at the thesaurus manager. This must be a semiautomatic process, which will be supported by a tool that compares the changes in the thesaurus manager and the use of terms in the local database, and makes proposals for the least changes to be made in the database. Therefore the thesaurus manager must maintain a history of semantic changes from release to release. The same data can be used to translate or transform terms in a query formulated according to the new thesaurus release against a database consistent with an older release.

Term servers are loaded with multiple thesauri from the local thesaurus management systems. Term servers are used as search aids. Equivalence expressions will be introduced between the terms in the different thesauri, which on one side help users to select correct terms for databases using authorities he/she is not familiar with. On the other side retrieval agents should be able to make such "translation" automatically, in case many different databases are addressed simultaneously. Such translation is approximate, and optimal translation of a complex query is an interesting research issue.

It should be mentioned, that there are two different notions of translation. In one case, the optimal terms one expects to find in one language/context/database for some concept are correlated to the **closest concepts** and their terms that one expects to find in another language/context/database for the same items. This is typically what we need, if we translate

queries automatically for different target databases. In the other case, we select **optimal circumscriptions** for a foreign concept, in local words not necessarily optimal in an equivalent local context. This is typically, what we need as user aid in order to make one understand a foreign vocabulary.

Equivalence expressions are not easily found, and their number increases with the possible combinations of thesauri. Therefore term servers may be cascaded to support multiple translation steps for scalability reasons. Finally, term servers must be updated with new releases of local thesauri, maintaining referential integrity of the equivalence expressions. Again, the history of changes is the key to that.

Within AQUARELLE we have enhanced our thesaurus management software SIS-TMS [4] to support cooperative development of multilingual thesauri. The system features release procedures with history of changes as described above and can also be configured as search aid thesaurus. By graphical visualization it allows for excellent understanding and control of complexly interlinked terminology structures. In parallel, ILSP has developed the AQUARELLE Term Server on top of the SIS-TMS server and an advanced module of their own for approximate, language neutral word matching, which tolerates a large variety of misspellings. The whole solution consists of a set of independent components with open interfaces. The system has found very good user response so far.

What to do now

We advocate for international cooperation to implement and experiment with a full architecture as described above, which means to provide solutions for term translation within complex query expressions, e.g. in Z39.50 protocol requests, for the interfacing between: term servers and retrieval agents (1), between thesaurus management systems and local databases (2) and between term servers (3). These three interfaces become really valuable, when an open standard communication protocol can be established, which allows to combine freely thesaurus management systems and their term servers with retrieval agents and collection classification systems.

Further the means for multilingual thesaurus contents production must be improved. On the one side, we have labor intensive human processes, by which international expert groups produce isolated terminology systems of high specialization and precision. These processes could be considerably enhanced by modern communication software. On the other side, we have a wealth of methods from computational linguistics, which are cheap, and create vocabularies of excellent balance and coverage, but not necessarily precise in special cases. The international framework of research politics and expert specialization does not really foster collaborations in the above methods. Therefore we are happy to participate together with ILSP and other partners in the European TELEMATICS project "Term-IT" [5], which has brought the spirits together and actually aims at a graceful combination of all known methods for better multilingual thesaurus production.

We believe, that the separation of the terminology service from the retrieval agents and collection management systems into an overall federated architecture in conjunction with improved methods for knowledge acquisition and term correlation, has the potential to make effective retrieval from a large number of multilingual data servers a reality. As well, the provision of a correlated terminology rather than reclassification of all data can make even highly specialized data widely accessible.

References

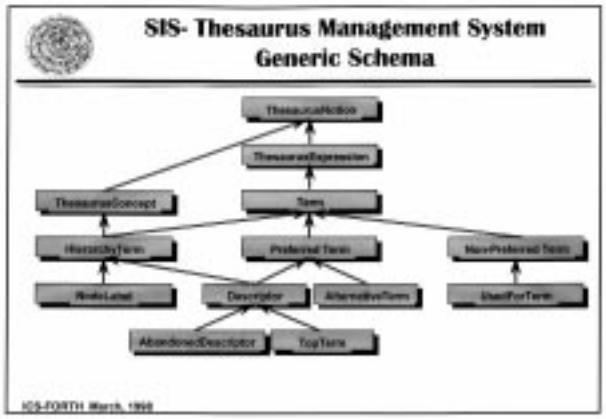
[1] M. Doerr,
"Authority services in global information spaces."
Heraklion - Crete,
Greece: FORTH, Institute of Computer Science-
Technical Report FORTH-ICS/TR-163, 1996
(http://www.ics.forth.gr/proj/isst/Publications/TechnicalReports.html)

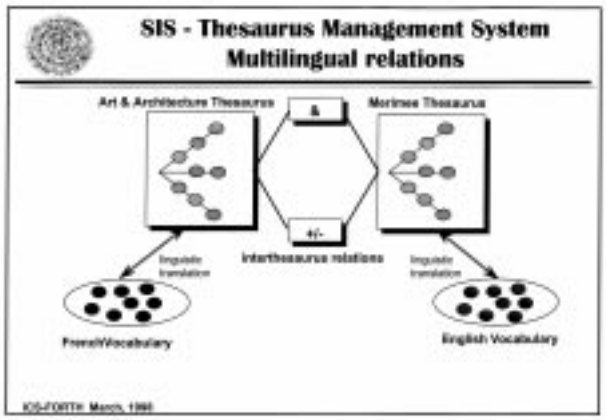
[2] M. Doerr, I. Fundulaki, V. Christofidis,
"The specialist seeks expert views: managing digital folders in the AQUARELLE project",
in: Museums and the Web 97: Selected Papers",
Archives & Museum Informatics, Pittsburg, 1997.
ISBN 1-885626-13-4.

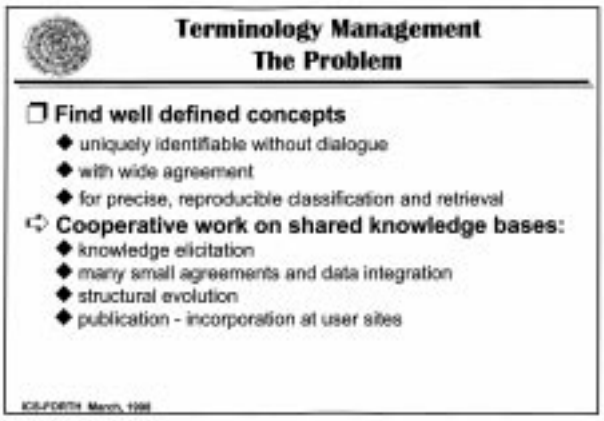
[3] M. Doerr,
"Reference Information Acquisition and Coordination",
in: "ASIS'97 - Digital Collections: Implications for Users, Funders, Developers and Maintainers",
Proceedings of the 60th Annual Meeting of the American Society for Information Sciences,
November 1-6 '97, Washington, Vol.34.
Information Today Inc.: Medford, New Jersey, 1997.
ISBN 1-57387-048-X.

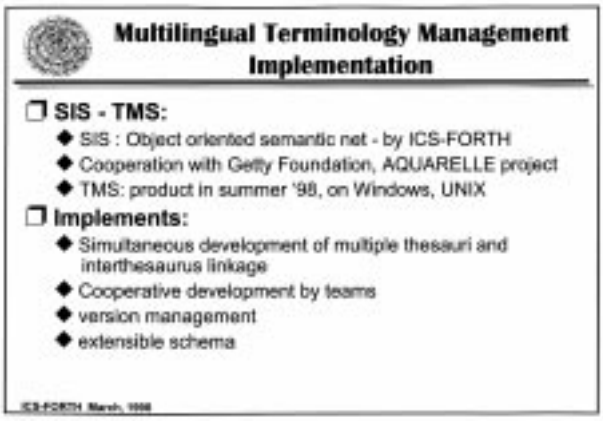
[4] "The SIS Thesaurus Management System SIS-TMS"
(http://www.ics.forth.gr/proj/isst/Systems/TMS/index.html)

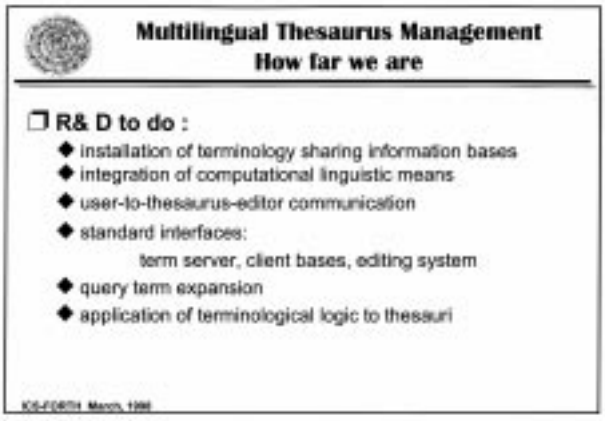
[5] http://www.mda.org.uk/term-it











3. The Design of the Macrostructure of a Computational Lexicon

M. Gavriilidou, P. Lambropoulou, E. Mantzari and S. Rousou, Linguists
Institute for Language and Speech Processing
Artemidos 6 & Epidaurou, Paradeisos Amarousiou
151 25 Athens, Hellas
E- mail: maria@ilsp.gr, penny@ilsp.gr, elena@ilsp.gr, sofia_r@ilsp.gr

Abstract

This paper presents specifications for the design of the macrostructure of a computational lexicon of Modern Greek for Natural Language Processing applications. The specifications for the lexicon's macrostructure concern criteria for the selection and the codification of the lemmata. The methodology adopted for the selection of the lemmata to be included in the lexicon is hybrid: it combines statistical processing of text corpora aiming at the production of a word frequency list and, subsequently, linguistic processing of this list, which, adhering to specific criteria aims at the compilation of the final macrostructure. Finally, criteria for the establishment of links between lemmata are presented.

Κατάρτιση Μακροδομής ενός Υπολογιστικού Λεξικού

Μ. Γαβριηλίδου, Π. Λαμπροπούλου, Έ. Μάντζαρη, και Σ. Ρούσσου, Γλωσσολόγοι
Ινστιτούτο Επεξεργασίας του Λόγου
Αρτέμιδος 6 & Επιδαύρου, Παράδεισος Αμαρουσίου
151 25 Αθήνα

0. ΕΙΣΑΓΩΓΗ

Αντικείμενο της παρούσας ανακοίνωσης είναι οι προδιαγραφές κατάρτισης της μακροδομής ενός **υπολογιστικού** λεξικού της **Νέας Ελληνικής (ΝΕ)**, η οποία βασίζεται σε κατάλογο λημμάτων που προέκυψε από στατιστική επεξεργασία ενός **σώματος γραπτών κειμένων**. Η κατασκευή ενός **υπολογιστικού** λεξικού απαιτεί τη συνεπή και συστηματική εφαρμογή ενός μοντέλου τυπικής αναπαράστασης του λεξιλογίου. Στο λημματολόγιο καταχωρίζονται αποκλειστικά μονολεκτικοί τύποι. Πολυλεκτικοί σχηματισμοί (λεξικές φράσεις, στερεότυπες φράσεις, κτλ.) αποτελούν αντικείμενο επόμενων επιπέδων περιγραφής, και επομένως σε αυτό το επίπεδο τα συστατικά τους καταχωρίζονται ανεξάρτητα. Συντακτική και σημασιολογική πληροφορία προβλέπεται να κωδικοποιηθεί σε επόμενα στάδια. Το λεξικό προορίζεται για εφαρμογές επεξεργασίας γραπτών κειμένων γενικής γλώσσας, υποστηρίζοντας τόσο τη διαδικασία της αναγνώρισης όσο και τη διαδικασία της παραγωγής (generation) λεξικών τύπων. Τέτοιες εφαρμογές είναι η λημματοποίηση, ο μορφοσυντακτικός χαρακτηρισμός γραπτών κειμένων, η ορθογραφική διόρθωση κτλ.

1. Αρχές Συγκρότησης του Σώματος Κειμένων

Το σώμα κειμένων αποτέλεσε την αρχική πηγή δεδομένων για την κατάρτιση του λημματολογίου. Οι προβλεπόμενες εφαρμογές του λεξικού (όπως ορίστηκαν παραπάνω) καθόρισαν τις παραμέτρους επιλογής του σώματος κειμένων, το οποίο απαρτίζεται από κείμενα **γενικής γραπτής σύγχρονης** (από το 1976 και μετά) γλώσσας, και αποτελεί υποσύνολο του Σώματος Κειμένων του Ινστιτούτου Επεξεργασίας του Λόγου.

Κείμενα		Αριθμός λέξεων
Βιβλία		**1.500.000**
	Λογοτεχνικά	644.000
	επιστημονικά / τεχνικά	856.000
Εφημερίδες		**7.500.000**
ΣΥΝΟΛΟ		**9.000.000**

2. Διαδικασία κατάρτισης λημματολογίου

Η μέθοδος που υιοθετήθηκε για την κατάρτιση του λημματολογίου είναι **υβριδική**, δηλαδή χρησιμοποιεί συμπληρωματικά δύο διαδικασίες: τη στατιστική επεξεργασία των δεδομένων ενός σώματος γραπτών κειμένων και τη γλωσσική επεξεργασία των δεδομένων που προκύπτουν από την πρώτη.

2.1. Στατιστική Επεξεργασία

Ο βασικός πυρήνας του λημματολογίου καθορίστηκε με βάση το μέτρο της συχνότητας εμφάνισης. Από τα συνολικά 51.764 λήμματα που προέκυψαν από τη λημματοποίηση και τη στατιστική επεξεργασία του σώματος κειμένων των 9.000.000 λέξεων, επιλέχθηκαν ως βασικός πυρήνας τα 20.000 συχνότερα λήμματα, με συχνότητα εμφάνισης ≥ 8.

2.2. Γλωσσική επεξεργασία

Η γλωσσική επεξεργασία συστηματοποιεί τα αποτελέσματα της στατιστικής επεξεργασίας. Περιλαμβάνει δύο διαδικασίες, κατά τις οποίες αφαιρούνται και προστίθενται λήμματα, με βάση συγκεκριμένες αρχές.

2.2.1. Διαδικασία αποκλεισμού λημμάτων

Δεδομένου ότι το γλωσσικό σύστημα που περιγράφεται στο λεξικό είναι το σύστημα της γενικής γλώσσας της ΝΕ, αποκλείστηκαν από τον αρχικό κατάλογο των λημμάτων:

- **λέξεις της αρχαίας ελληνικής** (ΑΕ) που εντοπίστηκαν σε αποσπάσματα ΑΕ κειμένων που χρησιμοποιούνται ως παραθέματα *(ειμί, όστις, κύων, μνα, ωσεί)*. Διατηρήθηκαν, ωστόσο, στον κατάλογο λέξεις που χρησιμοποιούνται και στη ΝΕ εκτός ΑΕ αποσπασμάτων (***σιδηρά*** κυρία) και λέξεις των οποίων ορισμένοι τύποι επιβιώνουν σε στερεότυπες εκφράσεις (***σώας*** τας *φρένας*, *υπό* ***μάλης***, *διά* ***χειρός***).
- **λέξεις έντονα διαλεκτικές** που εντοπίστηκαν κυρίως σε λογοτεχνικά κείμενα *(βαβούλι, τσαπέλα, φορτσέρι, πουλακίδα)*.
- **λέξεις που ανήκουν σε ειδικές υπογλώσσες** *(λιτή, μανέλα*). Παρέμειναν, ωστόσο, στον κατάλογο λέξεις ειδικών υπογλωσσών οι οποίες χρησιμοποιούνται συχνά στη γενική γλώσσα διατηρώντας την ίδια σημασία *(σάκχαρο, αφθώδης, μεθαδόνη)*.

2.2.2. Διαδικασία προσθήκης λημμάτων

Σύμφωνα με τις αρχές της μεθόδου που ακολουθήθηκε, το είδος των λημμάτων που προστίθενται καθορίστηκε με βάση διαδικασίες που προσιδιάζουν στο μορφολογικό επίπεδο και όχι με βάση διαδικασίες που προσιδιάζουν στα επίπεδα της σύνταξης και της σημασιολογίας (π.χ. συντακτικές δομές, σημασιολογικά πεδία ή λεξικές σχέσεις σημασίας).

Τα μορφολογικά κριτήρια στα οποία στηρίχθηκε ο εμπλουτισμός του λημματολογίου είναι τα εξής:

- ύπαρξη κοινών τύπων μεταξύ διαφορετικών λημμάτων,
- ύπαρξη εναλλακτικών μορφών για την ίδια λέξη.

Με βάση τα παραπάνω κριτήρια προστίθενται:

- Λήμματα που έχουν κοινούς λεξικούς τύπους με κάποιο από τα υπάρχοντα και
- ανήκουν στο ίδιο μέρος του λόγου: *(ο, η) γιατρός, (ο,η) εισαγγελέας, ο κρίνος/το κρίνο [των κρίνων].*
- ανήκουν σε διαφορετικά μέρη του λόγου: *μαθηματικός-ή-ό/(ο, η) μαθηματικός/τα μαθηματικά, ηχώ (ρήμα)/η ηχώ, λύνω/η λύση [λύσεις].*
- Λήμματα που εμφανίζουν
- εναλλακτικές ορθογραφήσεις: *αβγό/αυγό, τρένο/τραίνο.*
- σχέση τονικής διπλοτυπίας με κάποιο από τα υπάρχοντα: *αμερικανικός/αμερικάνικος, προπέρσινος/προπερσινός.*
- σχέση φωνητικής διπλοτυπίας με κάποιο από τα υπάρχοντα. Τα φωνητικά διπλότυπα προκύπτουν με βάση τα πάθη φωνηέντων και συμφώνων, π.χ. συναίρεση *(δεκαέξι/δεκάξι)*, αφαίρεση αρχικού φωνήεντος *(εβδομάδα/βδομάδα)*, αλλαγή στο αρχικό φωνήεν *(έξαφνα/άξαφνα)*, αφομοίωση *(σιρόκος/σορόκος)*, μετάθεση συμφώνων *(χούφτα/φούχτα)*, τροπή συμφωνικών συμπλεγμάτων *(ανοικτός/ανοιχτός, πτωχός/φτωχός)*.
- σχέση πολυτυπίας με κάποιο από τα υπάρχοντα σε ό,τι αφορά την κατάληξη: *η αρχιτέκτονας/η αρχιτεκτόνισσα, μουρμουρίζω/μουρμουράω.*

3. Κωδικοποίηση λημματολογίου

3.1. Λήμμα

Η λεξική μονάδα περιγραφής είναι το **λήμμα**, δηλαδή, ο τύπος που επιλέγεται να αντιπροσωπεύσει όλες τις κλιτές μορφές μιας λέξης, και που για τα ουσιαστικά είναι η ονομαστική ενικού, για τα ρήματα το πρώτο πρόσωπο ενικού αριθμού του ενεστώτα της οριστικής έγκλισης της ενεργητικής φωνής, κτλ. Οι περιπτώσεις που αποκλίνουν από την ανωτέρω κωδικοποίηση είναι οι εξής:

- απουσία ενικού αριθμού σε ουσιαστικά, οπότε κωδικοποιείται ως λήμμα η ονομαστική πληθυντικού: *πρόποδες, εγκαίνια.*
- απουσία αρσενικού γένους σε επίθετα, οπότε κωδικοποιείται ως λήμμα η ονομαστική του γένους που χρησιμοποιείται: *αποφράδα, εξώτερο.*
- απουσία θετικού βαθμού σε επίθετο ή επίρρημα, οπότε κωδικοποιείται ως λήμμα ο συγκριτικός ή ο υπερθετικός: *ανώτερος, υπέρτατος.*
- απουσία α' προσώπου ενικού αριθμού σε ρήματα. Πρόκειται για την περίπτωση των απρόσωπων και των τριτοπρόσωπων ρημάτων, οπότε κωδικοποιείται ως λήμμα το γ' ενικό πρόσωπο: *πρέπει, πρόκειται.*
- συστατικά στερεότυπων εκφράσεων τα οποία απαντούν σε συγκεκριμένους τύπους, οπότε κωδικοποιείται μόνο ο τύπος που χρησιμοποιείται: *[σώας τας] φρένας, [υπό] μάλης, [εν] αντιθέσει, [κατά] μόνας.*

3.2. Συνδέσεις μεταξύ λημμάτων

Στενές μορφολογικές και σημασιολογικές σχέσεις που εντοπίζονται μεταξύ των λημμάτων αποκρυσταλλώνονται μέσω της κωδικοποίησης δύο ειδών συνδέσεων των λημμάτων: σχέση οριζόντιας σύνδεσης και σχέση υπαγωγής.

3.2.1. Σχέση οριζόντιας σύνδεσης

Σχέση οριζόντιας σύνδεσης συνδέει λήμματα που αναγνωρίζονται ως διπλοτυπίες (όπως αυτές αναφέρθηκαν παραπάνω) με εξαίρεση τις εναλλακτικές ορθογραφήσεις. Παρόλη τη στενή μορφολογική και σημασιολογική σχέση των διπλοτυπιών, το υφολογικό τους φορτίο και η χρήση τους σαφώς διαφοροποιούνται: λέμε *έσκισα τη γάτα* αλλά όχι *έσχισα τη γάτα*, *τριχόπτωση και αρχές φαλάκρας* και όχι *καράφλας*. Τα λήμματα αυτά κωδικοποιούνται ως αυτόνομες λεξικές μονάδες, δεδομένου ότι δεν εμφανίζονται με αμοιβαία εναλλαγή σε όλα τα περιβάλλοντα χρήσης - η σύνδεσή τους, ωστόσο, καθιστά εμφανή τη σχέση τους.

3.2.2. Σχέση υπαγωγής

Δεδομένου ότι το λεξικό προορίζεται και για εφαρμογές αναγνώρισης, πρέπει να περιλαμβάνει τις διαφορετικές γραφηματικές εκδοχές με τις οποίες μια λέξη απαντά στο γραπτό λόγο, π.χ. *αυγό/αβγό, κασσέττα/κασέττα/κασέτα,* προκειμένου να τις αναγνωρίσει ως τύπους της ΝΕ (είναι προφανές ότι δεν κατατάσσονται σε αυτήν την κατηγορία ορθογραφικά λάθη του τύπου *πηγένο* αντί του *πηγαίνω* ή *τιρί* αντί του *τυρί*). Η σχέση οριζόντιας σύνδεσης δεν καλύπτει αυτήν την περίπτωση πολυτυπιών, δεδομένου ότι δεν πρόκειται για στενά συνδεόμενα λήμματα αλλά για εναλλακτικές ορθογραφήσεις του ίδιου λήμματος. Για αυτήν την περίπτωση χρησιμοποιείται η σχέση υπαγωγής, σύμφωνα με την οποία μία από αυτές τις ορθογραφήσεις επιλέγεται συμβατικά ως υπερ-λήμμα, στο οποίο υπάγονται οι υπόλοιπες. Ως υπερ-λήμμα επιλέχθηκε η ορθογραφική εκδοχή της Γραμματικής Τριανταφυλλίδη.

4. Συμπεράσματα

Σε αυτή την ανακοίνωση παρουσιάστηκαν οι προδιαγραφές κατάρτισης του λημματολογίου ενός υπολογιστικού λεξικού της ΝΕ γλώσσας.

Δεδομένου ότι το λεξικό προορίζεται για εφαρμογές Επεξεργασίας Φυσικής Γλώσσας, οι προδιαγραφές στοχεύουν στην **πλήρη** και **συστηματική** κωδικοποίηση του λεξιλογίου της γραπτής ΝΕ. Η συστηματική περιγραφή επιτυγχάνεται με

την ταξινόμηση και συστηματοποίηση των γλωσσικών δεδομένων μέσω των σχέσεων οριζόντιας σύνδεσης και υπαγωγής μεταξύ των λημμάτων. Η αξιοποίηση των μαρτυριών του σώματος γραπτών κειμένων συντελεί στην πληρότητα της περιγραφής.

5. Βιβλιογραφία

Goutsos D, O. Hatzidaki & P. King 1993: "A corpus-based approach to Modern Greek language research and teaching". Στο *I. Philippaki-Warburton, K.Nikolaidis, M. Sifianou* (εκδ.), *Themes in Greek Linguistics, Papers from the First International Conference on Greek Linguistics* (London: John Benjamins Publishing Company).

Mackridge P. 1990: *Η νεοελληνική γλώσσα* (Αθήνα: Εκδόσεις Πατάκη).

Αναστασιάδη - Συμεωνίδη Α. 1981: "Ένα είδος συνταγματικής μονάδας στα Νέα Ελληνικά". *Στο Μελέτες για την Ελληνική Γλώσσα*, Πρακτικά της 1ης Ετήσιας Συνάντησης του Τομέα Γλωσσολογίας της Φιλοσοφικής Σχολής του ΑΠΘ (Θεσσαλονίκη).

Μπασλής Γ. 1995: "Θηλυκά επαγγελματικά". *Στο Μελέτες για την Ελληνική Γλώσσα*, Πρακτικά της 15ης Ετήσιας Συνάντησης του Τομέα Γλωσσολογίας της Φιλοσοφικής Σχολής του ΑΠΘ (Θεσσαλονίκη).

Παυλίδου Θ. 1984: "Παρατηρήσεις στα θηλυκά επαγγελματικά". *Στο Μελέτες για την Ελληνική Γλώσσα,* Πρακτικά της 4ης Ετήσιας Συνάντησης του Τομέα Γλωσσολογίας της Φιλοσοφικής Σχολής του ΑΠΘ (Θεσσαλονίκη).

Ράλλη Α. 1990: "Λεξική Φράση: Αντικείμενο γλωσσολογικού ενδιαφέροντος". *Στο Μελέτες για την Ελληνική Γλώσσα,* Πρακτικά της 10ης Ετήσιας Συνάντησης του Τομέα Γλωσσολογίας της Φιλοσοφικής Σχολής του ΑΠΘ (Θεσσαλονίκη).

Τριανταφυλλίδης Μ. 1963: "Η "βουλευτίνα" και ο σχηματισμός των θηλυκών επαγγελματικών ουσιαστικών". *Στο Άπαντα*, 2 (Θεσσαλονίκη: ΑΠΘ-ΙΝΣ).

Τριανταφυλλίδης Μ. 1963: "Η δυναμικότητα των ασυμμόρφωτων λόγιων τύπων". *Στο Άπαντα*, 2 (Θεσσαλονίκη: ΑΠΘ-ΙΝΣ).

Τριανταφυλλίδης Μ. 1993: *Νεοελληνική γραμματική (της δημοτικής)*, ανατύπωση της έκδ. του ΟΕΣΒ (1941) με διορθώσεις (Θεσσαλονίκη: ΑΠΘ-ΙΝΣ).

Τσοπανάκης Αγ. 1994: *Νεοελληνική γραμματική.* (Εκδοτικός οίκος αδελφών Κυριακίδη και Βιβλιοπωλείον της Εστίας).

4. Lexipedia: A Multimedia Greek and Foreign Language Dictionary

Professor George Carayannis and Dr. Marianne Katsoyannou
Institute for Language and Speech Processing
Artemidos 6 & Epidaurou, Paradeisos Amarousiou
151 25 Athens, Greece
E-mail:gcara@ilsp.gr, marianna@ilsp.gr

Abstract

A multimedia Greek and foreign language dictionary directed to Greek children is presented. The dictionary contains 8000 entries, most of which have been taken from elementary school books. "Lexipedia" is a definition dictionary containing morphological information, usage examples and synonyms/antonyms. Morphological information is provided for each entry. A definition is provided for each different meaning of a word and a usage example or synonym/antonym (wherever possible) is given. Apart from pronunciation examples, images and sounds related to the entries are provided. The user is also able to sample the different tenses for an entry. The translation and the pronunciation of each entry is available in six languages (English, French, Spanish, German, Russian, Bulgarian). Interesting linguistic games are included in the dictionary, to make the learning process more amusing for the children.

Ένα Πολύγλωσσο Λεξικό Πολυμέσων

Καθηγητής Γιώργος Καραγιάννης και
Δρ. Μαριάννα Κατσογιάννου
Ινστιτούτο Επεξεργασίας του Λόγου
Αρτέμιδος 6 & Επιδαύρου, Παράδεισος Αμαρουσίου
151 25 Αθήνα
E-mail: gcara@ilsp.gr, marianna@ilsp.gr

1. Υπολογιστικός σχεδιασμός και τεχνολογικές καινοτομίες του προϊόντος

Στο πλαίσιο ανάπτυξης εκπαιδευτικών προϊόντων πολυμέσων, το Ινστιτούτο Επεξεργασίας του Λόγου

έχει προχωρήσει στο σχεδιασμό μιας σειράς λεξικών με το γενικό τίτλο Λεξιπαιδεία. Το εργαστηριακό πρωτότυπο το οποίο παρουσιάζεται έχει αναπτυχθεί με σκοπό να ανταποκριθεί στις ανάγκες της σειράς αυτής, η οποία απευθύνεται σε μαθητές διαφόρων ηλικιών. Ο σχεδιασμός προβλέπει τρία διαφορετικά προϊόντα: "Λεξιπαιδεία για το Δημοτικό", "Λεξιπαιδεία για το Γυμνάσιο" και "Λεξιπαιδεία για το Λύκειο". Δεδομένου ότι πρόκειται για λεξικά γενικής γλώσσας, το κυριότερο χαρακτηριστικό τους είναι ότι η πληροφορία εστιάζεται στη λεξική μονάδα (μονολεκτική ή πολυλεκτική), η οποία παρουσιάζεται ως στοιχείο ενός συστήματος του οποίου τα μέλη βρίσκονται σε συνεχή αλληλεξάρτηση. Ένα δεύτερο χαρακτηριστικό είναι η πρωτότυπη επεξεργασία των ερμηνευμάτων, τα οποία περιέχουν πολλών ειδών πληροφορίες: γραμματικές, μορφολογικές, συντακτικές, φρασεολογικές, σημασιολογικές, ... Η επεξεργασία αυτή γίνεται με γνώμονα τις πραγματικές ανάγκες του χρήστη, θεωρώντας ότι ο σκοπός του λεξικού γλώσσας δεν είναι μόνο είναι να καταλάβει ή να αναγνωρίσει ο χρήστης τη σημασία της λέξης, αλλά και να μπορεί να τη χρησιμοποιήσει σωστά.

Από την υπολογιστική άποψη, το κυριότερο χαρακτηριστικό του προγράμματος είναι οι τεχνολογικές καινοτομίες τις οποίες υιοθετεί ο σχεδιασμός του. Ο τρόπος με τον οποίο δίνονται οι πληροφορίες για τα λήμματα του λεξικού βασίζεται στην εκμετάλλευση της τεχνολογίας των πολυμέσων: η Λεξιπαιδεία περιλαμβάνει εικόνες, κίνηση, ήχους και βιντεοσκοπήσεις, που χρησιμοποιούνται τόσο στην παρουσίαση των ερμηνευμάτων, όσο και στα διάφορα παιχνίδια που συνοδεύουν το γλωσσικό υλικό. Η χρήση μορφολογικού λεξικού επιτρέπει να δίνεται όχι μόνο η πλήρης κλίση όλων των λημμάτων, αλλά και οι στοιχειώδεις μορφολογικές πληροφορίες για λέξεις οι οποίες δεν περιέχονται στο λεξικό. Ένα ειδικό λογισμικό επιτρέπει την αυτόματη φωνητική μεταγραφή των λέξεων και ένα δεύτερο χρησιμοποιείται για τον συλλαβισμό. Η χρήση συνθετικής φωνής είναι επίσης σημαντική, γιατί επιτρέπει την εκφώνηση όλων των λημμάτων, ενώ οι λειτουργίες αναγνώρισης φωνής μπορούν να χρησιμοποιηθούν σε ορισμένα παιχνίδια. Σκοπός είναι να αποτελέσει η χρήση του υπολογιστή όχι μόνο διευκόλυνση, αλλά και κίνητρο για τους νεαρούς χρήστες, οι οποίοι συνήθως δεν είναι εξοικειωμένοι με τη χρήση λεξικού. Σημειώνουμε εξάλλου ότι η Λεξιπαιδεία είναι το πρώτο λεξικό στην Ελλάδα το οποίο υλοποιείται απευθείας σε βάση δεδομένων - και όχι μεταφέροντας σε ηλεκτρονική μορφή τα δεδομένα κάποιου έντυπου λεξικού. Αυτό σημαίνει ότι το λεξικογραφικό εργαλείο αποκτά μεγαλύτερη εμβέλεια, ενώ παράλληλα αναπτύσσεται η δυνατότητα ελέγχου των δεδομένων με ιδιαίτερη ακρίβεια και αξιοπιστία.

Αξίζει ιδιαίτερα να αναφερθούν οι διάφοροι τρόποι αναζήτησης ενός λήμματος στο περιβάλλον της Λεξιπαιδείας. Υπάρχει καταρχήν ο κλασικός τρόπος, όπου ο χρήστης πληκτρολογεί την προς αναζήτηση λέξη, υπάρχουν όμως και άλλοι τρόποι αναζήτησης: μπορεί κανείς να πληκτρολογήσει μία κατάληξη ή ένα άλλο τμήμα λέξης και να δει όλες τις λέξεις που περιέχουν, π.χ. το τεμάχιο -επι-. Ένας ακόμη πιο ενδιαφέρων τρόπος αναζήτησης είναι αυτός που επιτρέπει τον εντοπισμό ορισμένων διγραφιών: ο χρήστης πληκτρολογεί τη λέξη, βάζοντας ένα ερωτηματικό στη θέση του γράμματος για το οποίο αμφιβάλλει. Σε απάντηση, το σύστημα εμφανίζει στην οθόνη τον τύπο ή τους τύπους με τους οποίους έχει λημματοποιηθεί η λέξη: για παράδειγμα, ο χρήστης που διστάζει μεταξύ των *αδελφός* και *αδερφός* ή *στεναχώρια* και *στενοχώρια* θα πάρει δύο απαντήσεις. Το ίδιο ισχύει και για λέξεις όπως *βρόμικος* και *βρώμικος* ή *χλομός* και *χλωμός*, οι οποίες ορθογραφούνται με δύο τρόπους. Οι δύο απαντήσεις οδηγούν κάθε φορά στην ίδια οθόνη-καρτέλα, η οποία περιέχει τις δύο ορθογραφίες της λέξης συνοδευόμενες από κοινό ερμήνευμα. Αυτός ο τρόπος αναζήτησης έχει περισσότερο ενδιαφέρον όταν πρόκειται για ομόηχες λέξεις με διαφορετικές γραφές: για παράδειγμα, αν ο χρήστης δεν θυμάται πώς γράφεται η λέξη /'klisi/ πληκτρολογεί: κ λ ? σ η και πληροφορείται ότι υπάρχει ένα λήμμα *κλήση* και δύο λήμματα *κλίση*, με διαφορετικές ση-

μασίες. Αφού ο χρήστης επιλέξει με ένα από τους παραπάνω τρόπους τη λέξη που θέλει να μελετήσει, περνάει στην οθόνη-καρτέλα που περιέχει τις πληροφορίες του ερμηνεύματος, οι οποίες παρουσιάζονται στη συνέχεια.

2. Λεξικογραφικός σχεδιασμός

2.1. Επιλογή και κατηγοριοποίηση λημμάτων

Ο αρχικός κατάλογος λημμάτων που περιλαμβάνονται στο λεξικό συγκροτήθηκε χρησιμοποιώντας τα γλωσσάρια των σχολικών εγχειριδίων. Το σύνολο των λέξεων που συγκεντρώθηκαν με τον τρόπο αυτό:

α) Συμπληρώθηκε ύστερα από αντιπαραβολή με το βασικό λεξιλόγιο της νέας ελληνικής και τις πηγές του ΙΕΛ σχετικά με τη συχνότητα εμφάνισης των λέξεων.

β) Τροποποιήθηκε, σύμφωνα με ορισμένα κριτήρια επιλογής, από τα οποία άλλα είναι καθαρά λεξικογραφικά και άλλα έχουν σχέση με την ηλεκτρονική μορφή του λεξικού. Για παράδειγμα, δεν λημματοποιούνται οι τύποι που μπορούν να προκύψουν από παραγωγή ή οι μετοχές σε -μένος (όπως τα υποκοριστικά ή τα παραθετικά των επιθέτων), εφόσον αυτές οι πληροφορίες δίνεται αυτόματα μέσω ειδικών λειτουργιών του λεξικού (κλίση, παραγωγή, ...).

Από τις παραπάνω διαδικασίες προέκυψε μία βασική μακροδομή, της οποίας ο τελικός έλεγχος και η συμπλήρωση γίνονται τόσο με λημματοποίηση των ορισμών, ώστε να εξασφαλίζεται ότι όλες οι λέξεις που χρησιμοποιούνται θα αποτελούν και λήμματα του λεξικού, όσο και με ταξινόμηση του λεξιλογίου σε σημασιολογικά πεδία. Η χρήση των σημασιολογικών πεδίων έχει καταρχήν σκοπό να μην επεξεργάζονται οι συντάκτες του λεξικού τα λήμματα με αλφαβητική σειρά για να αποφεύγονται οι επαναλήψεις και η κυκλικότητα στους ορισμούς, χρησιμοποιείται όμως και σαν διαδικασία ελέγχου, ώστε να μην υπάρχουν εννοιολογικά κενά.

2.2. Δομή και οργάνωση του ερμηνεύματος

Οι πληροφορίες οι οποίες δίνονται για κάθε λήμμα αποτελούν ισάριθμα πεδία μιας βάσης δεδομένων και οργανώνονται με τον παρακάτω τρόπο:

Γραφή, προφορά και μορφολογική πληροφορία: για κάθε λέξη-λήμμα δίνεται η γραφή, η μεταγραφή σε φωνητικό αλφάβητο και η δυνατότητα να ακούσει ο χρήστης την προφορά της λέξης. Ακολουθεί η πληροφορία για τη γραμματική κατηγορία, στην οποία αναφέρεται και το γένος για τα ουσιαστικά ή η συζυγία για τα ρήματα. Δεν υπάρχει παραπομπή σε πίνακες κλιτικών παραδειγμάτων, εφόσον ο χρήστης μπορεί να δει την πλήρη κλίση της λέξης, πατώντας το κατάλληλο κουμπί.

Ορισμός και κατάταξη σημασιών: Το μεγαλύτερο βάρος δίνεται στη διατύπωση του ορισμού, που γίνεται προσπάθεια να διατυπώνεται με πλήρη πρόταση (και όχι με απλή παράθεση συνωνύμων), και να αποτελείται από το υπερώνυμο του λήμματος συνοδευόμενο από τα διακριτικά του χαρακτηριστικά. Τα χαρακτηριστικά τα οποία, αν και σημαντικά για την κατανόηση μιας λέξης, δεν είναι υποχρεωτικά, χαρακτηρίζονται ως "συνήθη", ενώ αυτά τα οποία αποκλίνουν από τη λογική του ορισμού μπορούν να εμφανίζονται στο παράδειγμα.

Για να ανταποκρίνεται ο ορισμός στις ανάγκες του χρήστη (κατανόηση ή αναγνώριση της σημασίας μιας λέξης), πρέπει καταρχήν η λέξη που θέλουμε να ορίσουμε και το προτεινόμενο σημασιολογικό ισοδύναμό της να μην έχουν κοινή ετυμολογία, ώστε να μην έχουμε ταυτολογικούς ορισμούς. Όταν δεν υπάρχει άλλη λύση, κάτι που συμβαίνει κυρίως στους ορισμούς των επιθέτων και των επιρρημάτων, φροντίζουμε να υπάρχει δυνατότητα αυτόματης πρόσβασης στο σχετικό λήμμα. Την ίδια λύση ακολουθούμε και για ορισμένα ουσιαστικά που δηλώνουν δράση ή ενέργεια και των οποίων ο ορισμός περιλαμβάνει το αντίστοιχο ρήμα, όπως για παράδειγμα στο *ψάρεμα*, που ορίζεται ως το *να ψαρεύεις*, με αυτόματη πρόσβαση στο *ψαρεύω*.

Γίνεται επίσης προσπάθεια ώστε η λέξη η οποία δίνεται σαν σημασιολογικό ισοδύναμο του λήμματος

να είναι της ίδιας γραμματικής κατηγορίας με αυτό (ουσιαστικό = ουσιαστικό, ρήμα = ρήμα). Την κυριότερη εξαίρεση στον παραπάνω κανόνα αποτελούν τα επίθετα, για τα οποία ο ορισμός αποτελείται συνήθως από αναφορική πρόταση εισαγόμενη με το "που". Άλλο χαρακτηριστικό του ορισμού, το οποίο αναφέρεται εδώ γιατί δεν ανευρίσκεται σε όλα τα παιδικά λεξικά, είναι ότι αποτελείται από μία μόνο πρόταση. Παρά την λεξικογραφική αυστηρότητα των αρχών που ακολουθούνται, δίνεται προσοχή στη διατύπωση των ορισμών, ώστε να είναι κατανοητοί από χρήστες νεαρής ηλικίας.

Ο ορισμός μπορεί να περιλαμβάνει και ενδείξεις για τον γνωστικό τομέα στον οποίο ανήκει η λέξη ή μία από τις σημασίες της, όπως: (μαθημ.), (φυσ.), κ.λπ. Μέχρι τώρα έχουμε συνηθίσει να βλέπουμε τους χαρακτηρισμούς αυτούς μέσα σε παρενθέσεις, στη Λεξιπαιδεία όμως θα αντιπροσωπεύονται από εικονίδια. Η κατάταξη των σημασιών δεν γίνεται με βάση την ιστορική τους εξέλιξη. Συνήθως προτάσσεται η συχνότερη, η πιο συγκεκριμένη, ή η κυριολεκτική, σημασία. Ένα πρόβλημα που παραμένει είναι η σειρά των ορισμών στα ρήματα: δεδομένου ότι οι διαφορετικές συντακτικές χρήσεις ενός ρήματος δίνουν συνήθως διαφορετικούς ορισμούς, ο συντάκτης του λεξικού υποχρεώνεται να επιλέξει μεταξύ δύο διαφορετικών τρόπων κατάταξης: ο πρώτος είναι αυτός που αναφέραμε παραπάνω, ο δεύτερος λαμβάνει υπόψη τον αριθμό των ορισμάτων (οπότε αναφέρεται πρώτα η αμετάβατη χρήση του ρήματος).

Συνώνυμα, αντίθετα και σχόλια: για κάθε σημασία του λήμματος δίνονται τα συνώνυμα και αντίθετα, στα οποία ο χρήστης μπορεί να παραπεμφθεί αυτόματα με την τεχνολογία του υπερκειμένου. Η καρτέλα μπορεί επίσης να περιέχει σχόλια, τα οποία περιλαμβάνουν κυρίως υφολογικές και φρασεολογικές ενδείξεις ή γραμματικές παρατηρήσεις, για παράδειγμα "δεν σχηματίζει πληθυντικό" ή "χρησιμοποιείται κυρίως στη λογοτεχνία". Εδώ εντάσσονται και ορισμένες απλές παρατηρήσεις ετυμολογικού χαρακτήρα, όπως: "*ο δείκτης* λέγεται έτσι γιατί τον χρησιμοποιούμε για να δείχνουμε" ή "η λέξη *πλήκτρο* προέρχεται από το ρήμα *πλήττω* που σημαίνει *χτυπώ*".

Χρηστικά παραδείγματα, εκφράσεις και παροιμίες: Τα παραδείγματα, για τα οποία γίνεται προσπάθεια να προέρχονται από τα σχολικά βιβλία, μπορούν να περιέχουν επιπλέον πληροφορία, η οποία δεν έχει θέση στον ορισμό και/ή προσθέτει χαρακτηριστικά χρήσιμα για την κατανόηση μίας έννοιας, π.χ. το παράδειγμα που αντιστοιχεί στο λήμμα *ερπετό* αναφέρει "πολλαπλασιάζονται με αυγά". Ιδιαίτερη σημασία δίνεται στο φρασεολογικό πλαίσιο στο οποίο συνήθως συναντάται το λήμμα και το οποίο δεν περιέχεται πάντοτε στον ορισμό, π.χ. για μία λέξη όπως *ακτινοβολία*, φροντίζουμε να εμφανίζεται στο παράδειγμα το ρήμα *εκπέμπω*. Τα παραδείγματα αποτελούνται πάντοτε από πλήρεις προτάσεις, το δε περιεχόμενό τους προέρχεται, όταν αυτό είναι δυνατόν, από θέματα της σύγχρονης ζωής τα οποία μπορούν να κινήσουν το ενδιαφέρον ενός παιδικού ή νεανικού κοινού, ή από την Ελληνική παράδοση. Ένα τελευταίο χαρακτηριστικό των παραδειγμάτων είναι η (σχετική) διαχρονικότητά τους, ώστε να μπορούν να θεωρηθούν επίκαιρα σε μεταγενέστερες εκδόσεις του λεξικού.

Σε χωριστό πεδίο της βάσης δεδομένων καταχωρούνται παροιμίες ή στερεότυπες εκφράσεις, των οποίων η πρωτοτυπία είναι ότι συνδέονται με όλες τις λέξεις στις οποίες αναφέρονται. Έτσι, το *λύνει* και *δένει* εμφανίζεται τόσο στο ρήμα *λύνω* όσο και στο *δένω*, το *η ισχύς εν τη ενώσει* δίνεται στα λήμματα *ισχύς* και *ένωση*, κλπ. Οι εκφράσεις αυτές συνοδεύονται πάντοτε από την ερμηνεία τους και, όταν αυτό κρίνεται απαραίτητο, από σχολιασμό για τη χρήση τους.

Μετάφραση: κάθε σημασία του λήμματος μεταφράζεται σε έξι ξένες γλώσσες (αγγλική, γαλλική, ισπανική, γερμανική, ρωσική, βουλγαρική). Το τελικό προϊόν θα μπορεί να εμφανίζεται ως δίγλωσσο ή πολύγλωσσο και να ενεργοποιεί, σύμφωνα με την επιλογή του χρήστη μία ή περισσότερες από τις ξένες γλώσσες στις οποίες μεταφράζονται τα λήμματα.

Παιχνίδια: δεδομένου ότι το προϊόν απευθύνεται σε παιδιά, συνοδεύεται από μία σειρά παιχνιδιών, τα

οποία έχουν εκπαιδευτικό στόχο. Τα παιχνίδια της Λεξιπαιδείας σχεδιάζονται με τρόπο ώστε η χρήση τους να ενεργοποιεί να ελέγχει ή και να αναπτύσσει γλωσσικές γνώσεις και ικανότητες, ενώ καταβάλλεται ιδιαίτερη προσπάθεια ώστε το περιεχόμενο να συνδέεται, σε μικρότερο ή μεγαλύτερο βαθμό με αυτό του λεξικού. Δεδομένου ότι προβλέπεται η ένταξη της Λεξιπαιδείας σε ένα παιδικό περιβάλλον χρήσης, το οποίο αυτή τη στιγμή είναι υπό σχεδιασμό, το περιβάλλον αυτό θα συνδέσει στενά τα παιχνίδια με το κυρίως λεξικό.

Βιβλιογραφία

BEJOINT, Henri - THOIRON, Philippe, 1996, *Les dictionnaires bilingues*, Editions Duculot, 1996, 256p.

ΕΚΠΑΙΔΕΥΤΗΡΙΑ ΔΟΥΚΑ-ΙΝΣΤΙΤΟΥΤΟ ΕΠΕΞΕΡΓΑΣΙΑΣ ΤΟΥ ΛΟΓΟΥ, Εκπαιδευτικό Πολυλεξικό: *Προδιαγραφές απαιτήσεων - ανάλυση αναγκών,* (Παραδοτέο ΕΠΕΤ ΙΙ - ΔΙΑΛΟΓΟΣ), Αθήνα, 1995, 91 σ.

ΙΝΣΤΙΤΟΥΤΟ ΝΕΟΕΛΛΗΝΙΚΩΝ ΣΠΟΥΔΩΝ - Ίδρυμα Μανόλη Τριανταφυλλίδη, *Λεξικό της νέας ελληνικής γλώσσας (δείγμα: ψηφία ζ, η, θ, ι),* Θεσσαλονίκη, 1987, 124 σ.

PICOCHE, Jacqueline 1996, *Precis de lexicographie francaise; l'etude et l'enseignement du vocabulaire,* Nathan Universite, 191p.

SINCLAIR, John (ed), *Collins Cobuild, English Language Dictionary,* Collins - The University of Birmingham, 1988 (1987), 1703 p.

Οπτικοί δίσκοι

Dorling Kindersley Multimedia, *The Dorling Kindersley Childres's Dictionary.*

Ινστιτούτο Επεξεργασίας του Λόγου, *Λογομάθεια - Διδασκαλία της σύγχρονης ελληνικής γλώσσας.*

Microsoft Explorapedia, *Children's interactive Encyclopedia.*

III. Συνεισφορές Μελών του Ανθρωπίνου Δικτύου Γλωσσικής Τεχνολογίας / *Greek Human Network of Language Technology Members' Texts*

1. Electronic Lexicography

Professor Christoforos Charalambakis
Professor in Linguistics
Department of Education
University of Athens
44 Hippokratus str., 106 80 Athens - Greece
Tel.+ Fax: 00-30-1-3640719

Abstract

In this paper, under the title Electronic Lexicography, the main points presented are:

- Clarifications of terms and concepts
- Electronic Lexicography and Texts Corpora
- Electronic Lexicography: Research Objects-Problems and Prospects
- The Use of Dictionaries as Texts
- Lexical Representations on the Basis of Psychological Theories
- Future Prospects

Ηλεκτρονική Λεξικογραφία

Καθηγητής Χριστόφορος Χαραλαμπάκης
Καθηγητής Γλωσσολογίας
Παιδαγωγικό Τμήμα
Πανεπιστήμιο Αθηνών
Ιπποκράτους 44, 106 80 Αθήνα

1. Αποσαφήνιση όρων και εννοιών

Η *ηλεκτρονική λεξικογραφία* είναι ένας σχετικά νέος επιστημονικός κλάδος[1] ο οποίος βρίσκεται

[1] Στο Ινστιτούτο Επεξεργασίας του Λόγου λειτουργεί ειδικό "Τμήμα Ηλεκτρονικής Λεξικογραφίας" το οποίο έχει αναπτύξει εντυπωσιακή δραστηριότητα στην αυτόματη κατασκευή λεξικών και τον σχεδιασμό εξειδικευμένου λεξικογραφικού περιβάλλοντος. Παράλληλα έχει καταρτίσει Σώμα κειμένων αναφοράς της νεοελληνικής γλώσσας με περισσότερες από 30 εκατομμύρια τρέχουσες λέξεις και το οποίο εμπλουτίζεται συνεχώς.

σε άμεση σχέση με μια σειρά άλλων γνωστικών αντικειμένων.[2] Ο όρος αυτός, ο οποίος τείνει να καθιερωθεί στη σχετική πενιχρή ελληνική βιβλιογραφία,[3] ταυτίζεται συχνά με την *υπολογιστική λεξικογραφία* (computational lexicography), ενώ στην ουσία πρόκειται για δυο διαφορετικά στάδια κατά την επιστημονική μελέτη του λεξιλογικού θησαυρού μιας γλώσσας. Στη δεύτερη περίπτωση έχουμε το πρώτο αφηρημένο επίπεδο των υπολογιστικών σχεδιασμών, με τις οποίες ασχολείται η πληροφορική, ενώ στην πρώτη το αποτέλεσμα αυτών των διεργασιών με το "γέμισμα" του λεξικού, διαδικασία η οποία δεν μπορεί να γίνει χωρίς την ουσιαστική παρέμβαση του λεξικογράφου. Σε αντιστοιχία με την υπολογιστική γλωσσολογία θα ήταν ίσως προτιμότερο να καθιερωθεί ο όρος *υπολογιστική λεξικογραφία* και για τους δύο αυτούς επιμέρους κλάδους. Το *ηλεκτρονικό λεξικό* (electronic dictionary), αντίθετα, πρέπει να διαφοροποιείται από το *υπολογιστικό*. Για διαφημιστικούς λόγους προβάλλονται ορισμένα λεξικά ως ηλεκτρονικά (λ.χ. το *Μείζον Ελληνικό Λεξικό* Τεγόπουλου - Φυτράκη, 1997 και το *Υπερλεξικό*, 1998), στην ουσία όμως δεν πληρούν τις προϋποθέσεις των υπολογιστικών λεξικών. Κυκλοφορούν απλώς σε μορφή CD και γι' αυτό θα μπορούσαν να χαρακτηριστούν "ψευδο-ηλεκρονικά" ή "κλειστά" λεξικά. Στην ηλεκτρονική λεξικογραφία γίνεται ουσιώδης διάκριση ανάμεσα στο "lexicon", για να δηλωθεί μια σειρά από φορμαλιστικά λήμματα, και το "dictionary" για το συνηθισμένο τυπωμένο λεξικό. Πιο εξειδικευμένες είναι οι προσεγγίσεις demo και book. Η πρώτη αναφέρεται στα "χειροποίητα", μικρής έκτασης υπολογιστικά λεξικά και η δεύτερη στα μεγάλης έκτασης λεξικά, όπως το σύστημα Wordnet του Miller, 1985, (βλ. πιο κάτω, 3.2) το οποίο παρουσιάζει αυστηρά φορμαλιστική δομή για υπολογιστική και ψυχολογική χρήση.

Η υπερώνυμη έννοια στην οποία εντάσσεται η *υπολογιστική λεξικογραφία* είναι ασφαλώς η *τεχνητή νοημοσύνη* (artificial intelligence). Ο επιμέρους αυτός επιστημονικός κλάδος της Πληροφορικής ασχολείται με την ανάθεση σε μηχανές εντολών που απαιτούν πολύπλοκες νοητικές διεργασίες.[4] Πρόκειται για εντολές που αποσκοπούν στην προσομοίωση της ανθρώπινης ευφυίας, με στόχο, από το ένα μέρος τη δημιουργία ακόμα πιο αποτελεσματικών μηχανών (κομπιούτερ, ρομπότ), και από τα άλλο την κατανόηση της νοημοσύνης και των διανοητικών γενικότερα ικανοτήτων του ανθρώπου. Στην πρώτη περίπτωση διακρίνουμε την τάση της *τεχνολογικής* (engineering approach) και στη δεύτερη της *γνωστικής προσέγγισης* (cognitive-science approach). Η σημαντικότερη εμφανής διαφορά ανάμεσα στις δύο προσεγγίσεις είναι τα κριτήρια επίτευξης του επιδιωκόμενου στόχου: Αυτό που είναι επιτυχία για τον μηχανικό (η δημιουργία δηλ. μηχανής που θα ξεπερνούσε τις διανοητικές ικανότητες του ανθρώπου) κρίνεται ως αποτυχία από τον γνωστικό επιστήμονα.

Για την τεχνολογική προσέγγιση είναι ευρύτατα διαδεδομένος ο όρος *Επεξεργασία Φυσικής Γλώσσας*[5] (natural language processing, NLP), ενώ για τη γνωστική ο όρος *Υπολογιστική γλωσσολογία* (computational linguistics, CL). Τελικά, ο επιμέρους κλάδος της *Υπολογιστικής γλωσσολογίας* έχει αποκτήσει κατά τα τελευταία χρόνια σχετική αυτονομία. Αξιοποιεί τα πορίσματα τόσο της Γλωσσολογίας όσο και της (εφαρμοσμένης) Πληροφορικής, η οποία ασχολείται με την μηχανική επεξεργασία φυσικών γλωσσών σε όλα τα επίπεδα περιγραφής. Τα σημαντικότερα ερευνη-

[2] Ένα μέρος του προβληματισμού για την ηλεκτρονική λεξικογραφία, σε σχέση με την αξιοποίηση των σωμάτων κειμένων, με απασχολεί στη μελέτη: "Λεξικογραφική πρακτική με βάση σώματα κειμένων", στον τόμο: Χαραλαμπάκης, 1997, 345-360.

[3] Σύγκρ. Γ. Μπαμπινιώτης, "Ηλεκτρονική επεξεργασία της ελληνικής γλώσσας", στον τόμο: *Η γλώσσα ως αξία. Το παράδειγμα της Ελληνικής,* Αθήνα 1994, 213-219.

[4] Εντυπωσιακές είναι η σχετικές δραστηριότητες του Εργαστηρίου Τεχνητής Νοημοσύνης του Τμήματος Πληροφορικής του Οικονομικού Πανεπιστημίου Αθηνών το οποίο διευθύνει ο καθηγητής Ι. Κόντος. Βλ. *Λογοπλοήγηση*, 3, 1997, 24-27.

[5] Μια σειρά προβλημάτων που σχετίζονται με τον επιστημονικό αυτό κλάδο, όπως είναι ο μερισμός και η γραμματική, η αναπαράσταση του λόγου, η παραγωγή γλώσσας (language generation), η επίλυση φαινομένων αμφισημίας κ.ά, εξετάζονται στον τόμο: Ralli κ.ά., 1997.

τικά αντικείμενα της επιστήμης αυτής είναι:

- Ανάπτυξη φορμαλισμών για ακριβή και μηχανικά μεταφράσιμη παράσταση γλωσσικών δεδομένων ή μοντέλων.
- Ανάπτυξη διαδικασιών ανάλυσης και παραγωγής κειμένων φυσικής γλώσσας (μερισμός (parsing), μηχανική μετάφραση).
- Μοντέλα για μίμηση γλωσσικής συμπεριφοράς (λ.χ. στρατηγικές του διαλόγου, συστήματα ερώτησης - απάντησης).
- Τράπεζες εργασίας (workbanks) για γραμματικά και άλλα μοντέλα τα οποία καθιστούν δυνατό τον έλεγχο των κανόνων, και
- Προγράμματα για συλλογή και στατιστική επεξεργασία μεγάλου πλήθους γλωσσικών δεδομένων, λ.χ. για αυτόματη λημματοποίηση (ένταξη δηλ. τύπων μιας λέξης στα ανάλογα λεξήματα), για την παραγωγή πινάκων συχνότητας, για την αυτόματη δημιουργία πινάκων σύμφωνα με ορισμένα λήμματα, για τη δημιουργία πινάκων λέξεων με το συγκείμενό τους (concordances). Βλ. και *αυτόματη* ή *μηχανική μετάφραση* (machine translation), *συστήματα ειδημόνων* (expert systems).

Απόλυτη σχεδόν αυτονομία έχει αποκτήσει η *Γλωσσολογία σώματος κειμένων* (corpus linguistics) η οποία με την εξέλιξη της υψηλής τεχνολογίας και της τηλεματικής έχει σημειώσει εντυπωσιακή πρόοδο. Ευρύτερη έννοια έχει ο όρος *γλωσσικοί πόροι* (language resources) στον οποίο εντάσσονται γραπτά και προφορικά σώματα κειμένων, λεξικές βάσεις δεδομένων, γραμματικές και ορολογίες.[6]

2. Ηλεκτρονική λεξικογραφία και σώματα κειμένων

Ο όρος *σώμα υλικού* ή *σώμα κειμένων* (corpus) δηλώνει γενικά ένα πεπερασμένο πλήθος συγκεκριμένων γλωσσικών εξωτερικεύσεων που αποτελεί την εμπειρική βάση για γλωσσολογικές έρευνες. Εδώ θα μας απασχολήσουν τα μηχανικά αναγνώσιμα ή ηλεκτρονικά σώματα κειμένων (machine-readable or computer corpora). Η αξία και ο τρόπος δημιουργίας του σώματος κειμένων εξαρτάται από τα εξειδικευμένα ερωτήματα και τις μεθοδολογικές προϋποθέσεις του θεωρητικού πλαισίου της έρευνας, όπως είναι π.χ. η διαφορετική εκτίμηση των εμπειρικών δεδομένων στον δομισμό και τη γενετική- μετασχηματιστική γραμματική. Ο N. Chomsky, σε αντίθεση με τον L. Bloomfield, θεώρησε ως μοναδική αξιόπιστη πηγή για τη μελέτη της γλώσσας τη διαίσθηση του μητρικού ομιλητή και αμφισβήτησε την αξία των σωμάτων κειμένων. Οι μετέπειτα εξελίξεις στο χώρο της ηλεκτρονικής λεξικογραφίας έδειξαν ότι ο Chomsky είχε άδικο για τους παρακάτω λόγους:[7]

1. Η διαχωριστική γραμμή ανάμεσα στο *σώμα κειμένων* και τη *διαίσθηση* του μητρικού ομιλητή δεν είναι ορθή, αφού στις γλωσσικές αναλύσεις είναι απαραίτητα και τα δύο.

2. Η γενετική - μετασχηματιστική γραμματική βασίζεται στη διαίσθηση του μητρικού ομιλητή, χωρίς να λαμβάνει υπόψη της τον μη φυσικό ομιλητή που γνωρίζει επαρκώς την ξένη γλώσσα και ο οποίος χρειάζεται περισσότερο από τον πρώτο τις μαρτυρίες του σώματος κειμένων.

3. Η θεμελιώδης διάκριση ανάμεσα στη *γλωσσική ικανότητα* (competence) και τη *γλωσσική πραγμάτωση* (performance) έχει τεθεί υπό μερική αμφισβήτηση, όπως έδειξαν οι νεότεροι επιστημονικοί κλάδοι της κοινωνιογλωσσολογίας, της ψυχογλωσσολογίας, της πραγματογλωσσολογίας και της ανάλυσης της ομιλίας. Η εφαρμοσμένη γλωσσολογία κατέστησε εξ άλλου σαφές ότι για την κατάκτηση και την εκμάθηση της γλώσσας παίζει καθοριστικό ρόλο ο τρόπος με τον οποίο χρησιμοποιείται.

4. Ο ιδανικός μητρικός ομιλητής/ακροατής είναι

[6] Βλ. το κεφάλαιο Language resources (σσ. 381-407) στον τόμο Varile - Zampolli, 1997, με πλούσια σχετική βιβλιογραφία.

[7] Βλ. Leech, 1996, 73-80, και ιδιαίτερα τις σσ. 74-75.

μια άκρως προβληματική έννοια, αφού στην πράξη ο κάθε ομιλητής είναι κάτοχος διαφόρων κοινωνιόλεκτων και γεωγραφικών ποικιλιών, χρειάζεται επομένως το σώμα κειμένων για καθολική εποπτεία της γλωσσικής χρήσης.

5. Συστηματικές έρευνες σωμάτων κειμένων έδειξαν ότι υπάρχουν εκατοντάδες περιπτώσεων που δεν μπορούν να κατηγοριοποιηθούν ή να οδηγήσουν σε γενικεύσεις με βάση τη διαίσθηση.

6. Κατά τη επεξεργασία φυσικής γλώσσας με τη βοήθεια ηλεκτρονικού υπολογιστή είναι απαραίτητη η μελέτη της γλωσσικής πραγμάτωσης όπως εμφανίζεται σε αυθεντικά κειμενικά δεδομένα.

Από το άλλο μέρος, επιβάλλεται να τονιστεί ότι και τα σώματα κειμένων δεν αποτελούν πανάκεια. Όσο μεγάλη έκταση και αν έχουν δεν μπορούν να πλησιάσουν το "ιδανικό σώμα κειμένων" που θα περιείχε όλα τα γραπτά και προφορικά κείμενα μιας γλώσσας. Είναι επίσης σχεδόν αδύνατο να συμπεριληφθούν όλες οι γλωσσικές ποικιλίες σε μια δεδομένη χρονική στιγμή, ενώ και η διάσταση του χρόνου επαυξάνει τα θεωρητικά και πρακτικά προβλήματα.

Τα παραδοσιακά Αρχεία γλωσσικού υλικού σε χειρόγραφα ή πληκτρολογημένα δελτία έχουν σήμερα εντελώς ξεπεραστεί. Τα μηχανικά αναγνώσιμα σώματα κειμένων παρουσιάζουν δύο πολύ σημαντικά πλεονεκτήματα: α). Μπορεί να γίνει αυτόματη επεξεργασία τους (αλφαβητική ταξινόμηση, συντακτική ανάλυση (parsing) κ.ά.) με απίστευτη ταχύτητα, συνέπεια και ακρίβεια, που δεν θα μπορούσε να επιτύχει ακόμα και έμπειρη πολυμελής ερευνητική ομάδα. β). Η μεταφορά του υλικού γίνεται αυτόματα. Η "δημοσίευσή" του (υπό μορφή μαγνητικής ταινίας, δισκέτας ή ψηφιακού δίσκου, CD) είναι προσιτή στον χρήστη σε οποιοδήποτε σημείο της γης.

Ιδιαίτερη αναφορά πρέπει να γίνει στα *προ-επεξεργασμένα σώματα κειμένων* (pre-processed corpora), όπως είναι αλφαβητικά λεξιλόγια, συνήθως ενός συγγραφέα, που παραθέτουν τη λέξη στο κέντρο του στίχου με το συγκείμενό της και σχετική παραπομπή στην αρχή της κάθε γραμμής (KWIC concordance). Πολύ χρήσιμο είναι το γραμματικό μαρκάρισμα (grammatical tagging), δηλ. ο προσδιορισμός του μέρους του λόγου στο οποίο ανήκει κάθε λέξη. Αυτό είναι το πρώτο στάδιο μερισμού (parsing), με άλλα λόγια της συντακτικής ανάλυσης του σώματος κειμένων.

Οι έρευνες που βασίζονται σε εκτενή σώματα κειμένων έχουν ποικίλες εφαρμογές, όπως στη λεξικογραφία, τη διδασκαλία της γλώσσας, ιδιαίτερα για την εκμάθησή της με τη βοήθεια ηλεκτρονικού υπολογιστή, την αυτόματη ή μηχανική μετάφραση, την επεξεργασία της ομιλίας (speech processing) ή γενικότερα την τεχνολογία φωνής[8] (σύνθεση και αναγνώριση φωνής: speech synthesis and recognition) κ.ά.

3. Ηλεκτρονική λεξικογραφία: Ερευνητικά αντικείμενα - Προβλήματα και προοπτικές

Ένα από τα σοβαρότερα προβλήματα που αντιμετωπίζει σήμερα η ηλεκτρονική λεξικογραφία είναι ο τρόπος επιλογής των σωμάτων κειμένων, έτσι ώστε να αποτελούν αντιπροσωπευτικό δείγμα του συνόλου της γλώσσας. Σε πολλές χώρες δημιουργούνται εθνικά σώματα κειμένων (national text corpora), για την επιλογή όμως του υλικού δεν μπορεί να υπάρξει συμφωνία ανάμεσα στους ειδικούς. Όπως τονίζει χαρακτηριστικά ο Wilks, κ.ά. (1996, 8-9), ένα σώμα κειμένων που θα κρατούσε την ισορροπία ανάμεσα στις διάφορες δυναμικές τάσεις της γλώσσας "αποτελεί στην πράξη χίμαιρα. Υπάρχουν μόνο σώματα κειμένων για ειδικούς σκοπούς και επιμέρους θέματα". Ως λύση προτείνουν οι παραπάνω ερευνητές την καταγραφή των πυρηνικών σημασιών των λέξεων από τα υπάρχοντα λεξικά και την προσαρμογή τους σε ένα επιμέρους σώμα κειμένων,

[8] Γ. Καραγιάννης, "Φωνή (ή επεξεργασία της φωνής)", *Πάπυρος - Λαρούς - Μπριτάννικα*, 60, 1994, 251-257.

ανεξάρτητα από το θεματικό υπόβαθρο και το σκοπό συγκρότησής του. Για τον προσδιορισμό των σημασιών των λέξεων ο κάθε λεξικογράφος έχει τρεις βασικές πηγές στη διάθεσή του: α) ένα επιλεγμένο σώμα κειμένων, β) τη διαίσθησή του ως μητρικού ομιλητή και γ) την αξιοποίηση πληροφοριών από άλλα λεξικά.

Το δεύτερο δισεπίλυτο πρόβλημα της υπολογιστικής γλωσσολογίας, ιδιαίτερα στα πλαίσια της τεχνητής νοημοσύνης, παραμένει ο υπολογιστικός προσδιορισμός της σημασίας των λέξεων.[9] Το ερευνητικό ενδιαφέρον για τον συμβολικό ή μη χαρακτήρα της σημασίας, χωρίς να αποκλείονται άλλες αξιοπρόσεκτες ενδιάμεσες περιπτώσεις, παραμένει ζωηρό. Έχουν γίνει επίσης προσπάθειες ερμηνείας της σημασίας ως αναφοράς, ως φυσικού χειρισμού (physical manipulation), ως δράσης ή συμπεριφοράς κ.ά. Η *διαδικαστική σημασιολογία* (procedural semantics), που αναπτύχθηκε στα πλαίσια της τεχνητής νοημοσύνης, θεωρεί ότι οι σημασίες ως σύμβολα είναι περισσότερο διαδικασίες παρά οντότητες (entities) οποιουδήποτε είδους. Για την *υπολογιστική σημασιολογία* (computational semantics) παρουσιάζουν ιδιαίτερο ενδιαφέρον οι ακόλουθες θεωρίες: Ο *λειτουργισμός* του G. Frege (Fregean functionalism), ο *συνδετικισμός* (connectionism), *η αλυσίδα του Μάρκοφ, η στερεοτυπική θεωρία της σημασίας* κ.ά. Τα βασικά θεωρητικά ζητήματα της σημασίας δεν φαίνεται να απασχολούν τους παραδοσιακούς λεξικογράφους.

Για τη δημιουργία υπολογιστικών λεξικών είναι καθοριστικός ο ρόλος των *σημασιολογικών πρωτογόνων* (semantic primitives) τα οποία βρίσκονται σε όλα τα συστήματα κατανόησης φυσικών γλωσσών και ονοματίζουν βασικές έννοιες που υπόκεινται στην ανθρώπινη σκέψη.

Η πρακτική που ακολουθείται στα παραδοσιακά λεξικά για τον προσδιορισμό των σημασιών ενός λήμματος παρουσιάζει αρκετές δυσκολίες και το πρόβλημα αυτό μεταφέρεται και στα ηλεκτρονικά. Το *Λεξικό της νέας Ελληνικής γλώσσας* του Γ. Μπαμπινιώτη, Αθήνα 1998, εντοπίζει 52 σημασίες του ρήματος *κόβω*, ενώ το *Νέο ελληνικό λεξικό της σύγχρονης δημοτικής γλώσσας* του Ε. Κριαρά, Αθήνα 1995, καταγράφει για το ίδιο ρήμα 20 μεταβατικές και 7 αμετάβατες, με ορισμένες επιπλέον υποκατηγοριοποιήσεις. Η κατάτμηση των χρήσεων σε διακριτές σημασίες είναι ένα καθαρά υποκειμενικό θέμα. Η αξία, πάντως, ενός λεξικού κρίνεται από τη μεθοδολογία που ακολουθεί ως προς την επεκτασιμότητά του, τη δυνατότητα δηλ. να εντάξει νέες σημασίες που δεν υπάρχουν ήδη στο λεξικό με βάση κειμενικά σώματα. Τα *συμβατικά λεξικά* (conventional dictionaries) δεν μπορούν να αποτελέσουν τη βάση στην οποία θα μπορούσε να στηριχτεί η λεξική γνώση, αφού δεν διαθέτουν τη γνώση του πραγματικού κόσμου, όπως το *διανοητικό* ή *ανθρώπινο λεξικό* (mental or human lexicon) στο οποίο υπάρχουν οι απαραίτητοι μηχανισμοί αναγνώρισης των νέων σημασιών των λέξεων.

Η *κυκλικότητα* στα λεξικά (dictionary circles) είναι ένα άλλο σημαντικό ερευνητικό θέμα της υπολογιστικής λεξικογραφίας. Ο όρος κυκλικότητα δηλώνει την εμφάνιση ενός λήμματος ως κρίκου κάποιας αλυσίδας τουλάχιστον δύο φορές.[10] Η ερμηνεία των λημμάτων γίνεται συχνά με συνώνυμες λέξεις ή με λέξεις, η ερμηνεία των οποίων πρέπει να αναζητηθεί στα σχετικά λήμματα. Οι αλυσίδες ερμηνείας είναι απαραίτητο να καταλήγουν σε γνωστές λέξεις. Η σημασιολογία του ορισμού (the semantics of definition) σε ένα λεξικό συνδέεται με την αναπαραστατική του εργονομία (representational ergonomics). Η υπογλώσσα (sublanguage) του ορισμού του λήμματος παρουσιάζει ασυνέπειες, ενώ το λεξιλόγιό της δεν ερ-

[9] Τα θέματα αυτά συζητούνται εκτενώς στον Wilks κ.ά., 1996, A short history of meaning, σσ. 11-28. Ο τίτλος του βιβλίου *Electric words* παραπέμπει σαφώς στην ηλεκτρονική λεξικογραφία. Δεν χρησιμοποιείται όμως ο όρος *electronic* για να δηλωθεί ενδεχομένως η συνυποδήλωση των "ηλεκτρικών λέξεων" ως "ηλεκτρισμένων", που προκαλούν φόρτιση, διέγερση, ρίγη συγκίνησης. Βλ. και Tomaszczyk - Lewardowska (εκδ.), 1991.

[10] Βλ. Ι. Κόντος, 1997.

μηνεύεται ικανοποιητικά. Όταν ο χρήστης του λεξικού κατανοεί αυτή την υπογλώσσα τότε δεν τίθεται θέμα. Όταν όμως το λεξικό αποτελεί τη βάση για υπολογιστικούς πόρους, τότε είναι καθοριστικής σημασίας ένα σαφώς διαμορφωμένο σύστημα για τον ορισμό των λημμάτων. Αυτό μπορεί να γίνει με ένα ελεγχόμενο λεξιλόγιο, συνήθως 2.000 περίπου λέξεων, οι οποίες όμως είναι κατά κανόνα πολύσημες. Έτσι το μη ελεγχόμενο λεξιλόγιο εμφανίζει κατά μέσο όρο 2 σημασίες έναντι 12 του ελεγχόμενου (Wilks κ.ά., 1996, 99).

Τα λεξικά, ως ανθρώπινα δημιουργήματα, είτε πρόκειται για το έργο ενός μόνο λεξικογράφου (πράγμα σπάνιο σήμερα) είτε για το αποτέλεσμα ερευνητικής ομάδας με πλήρη ηλεκτρονικό εξοπλισμό, παρουσιάζουν κατά κανόνα ατέλειες και αδυναμίες. Ο Wilks κ.ά. (1996, 63 κ.ε.) διακρίνουν 5 βασικά είδη λεξικών: το λεξικό της καθιερωμένης γλώσσας, το συνωνυμικό, το δίγλωσσο[11], το υφολογικό και την κονκορντάντσια.

3.1. Τα λεξικά ως κείμενα

Η βασική μεθοδολογική αρχή που ακολουθεί ο Wilks κ.ά. (1996, 102) είναι ότι τα λεξικά πρέπει να αντιμετωπίζονται ως πραγματικά κείμενα, όπως αποδεικνύει το γεγονός ότι οι λέξεις που χρησιμοποιούνται στα ερμηνεύματα είναι εξίσου αμφίσημες με τα λήμματα. Αυτό σημαίνει ότι η υπολογιστική λεξικογραφία έχει στη διάθεσή της ένα ακόμα αξιοπρόσεκτο σώμα κειμένων.

Τα σπουδαιότερα απ' αυτά τα λεξικά, τα οποία εμφανίζουν πρωτότυπη εσωτερική σχηματοποίηση (internal formalization) με χρήση ηλεκτρονικού υπολογιστή, είναι το *Longman Dictionary of Contemporary English* (LDOCE), βλ. Procter, 1978, και το *Collins Cobuild* (Sinclair, 1987). Το LDOCE στη μηχανικά αναγνώσιμη μορφή του περιέχει 41.100 λήμματα. Τα ερμηνεύματα δίνονται με ένα ελεγχόμενο λεξιλόγιο 2.000 λέξεων. Οι γραμματικοί κώδικες συμπεριλαμβάνουν 110 περίπου συντακτικές κατηγορίες. Περιέχει ακόμα box codes με μια σειρά πρωτογόνων σε ιεραρχική δομή, όπως "αφηρημένο", "συγκεκριμένο", "έμψυχο" κ.ά. Το θεματικό κωδικό σύστημα περιλαμβάνει 124 βασικές κατηγορίες με 369 υποδιαιρέσεις.

Το *Cobuild* έσπασε τα δεσμά της συμβατικής λεξικογραφίας καθώς η σύνταξή του βασίστηκε σε ένα εκτενές αντιπροσωπευτικό σώμα κειμένων με 20 εκατομμύρια τρέχουσες λέξεις. Τα επεξηγηματικά παραδείγματα ανταποκρίνονται σε πραγματικές χρήσεις και δεν είναι κατασκευασμένα όπως στα παραδοσιακά λεξικά. Μια από τις σημαντικότερες καινοτομίες του λεξικού αυτού είναι ότι οι ορισμοί αποτελούν πλήρεις προτάσεις στις οποίες χρησιμοποιούνται τα εκάστοτε λήμματα και έτσι καταγράφεται στην ουσία ένα ακόμα παράδειγμα για τη χρήση της λέξης. Όσο και αν η κατηγοριοποίηση των σημασιών των λημμάτων παραμένει υποκειμενική, το *Cobuild* είναι πολύ πιο αξιόπιστο από τα άλλα λεξικά, αφού η συχνότητα των χρήσεων έχει ελεγχθεί από το σώμα κειμένων.

3.2. Λεξικές αναπαραστάσεις με βάση ψυχολογικές θεωρίες

Η σύνταξη λεξικών στηρίζεται κατά κανόνα σε γλωσσολογικές θεωρίες. Αξίζει όμως να γίνει αναφορά σε λιγότερο γνωστά συστήματα ψυχολογικών θεωριών οι οποίες εδράζονται σε πειραματικά μοντέλα που αναφέρονται στον τρόπο με τον οποίο σκέπτονται και χρησιμοποιούν οι άνθρωποι τις λέξεις. Το γνωστότερο απ' αυτά είναι το σύστημα λεξικής οργάνωσης WordNet[12] το οποίο εντάσσει τα λεξικά στοιχεία σε συνωνυμικές ομάδες και κατόπιν τις συνδέει μεταξύ τους με βάση μια σειρά από σχέσεις. Οι σημασίες των λέξεων διακρίνονται ρητά, δεν αριθμούνται όμως, όπως στα παραδοσιακά λεξικά. Οι σημα-

[11] Για δίγλωσσα ηλεκτρονικά λεξικά βλ. τις παρατηρήσεις του P. Scharpe, "Electronic dictionaries with particular reference to the design of an electronic bilingual dictionary for English - speaking learners of Japanese", *International Journal of Lexicography*, 8, 1995, 45 κ.ε.

[12] Beckwith R. κ.ά., "Word Net: a lexical database organized on psycholinguistic principles", *Proceedings of the first international lexical acquisition workshop* (IJCAI-89), Detroit.

σίες που εμφανίζει ένα συνωνυμικό σύνολο συνδέονται με άλλες συνωνυμικές ομάδες με βάση τις λεξικές σχέσεις της *υπερωνυμίας/υπωνυμίας, μερωνυμίας/ολωνυμίας* (meronymy/holonymy), δηλ. τη σχέση του μέρους με το όλο, *τροπωνυμίας* (troponymy), των τροπικών σχέσεων των ρημάτων, *αντωνυμίας* (antonymy), δηλ. των αντιθέτων εννοιών κ.ά.

Μια από τις σημαντικότερες βάσεις γνώσεων (knowledge bases) για γενικούς σκοπούς είναι το πρόγραμμα CyC (βλ. Wilks κ.ά., 1996, 127-128). Πρόκειται για την κωδικοποίηση ενός εκατομμυρίου λημμάτων από μια εγκυκλοπαίδεια, μια επένδυση χρόνου που υπολογίζεται σε απασχόληση δύο ατόμων επί ένα αιώνα. Η διαφαινόμενη αποτυχία της έρευνας θα επηρεάσει τη στάση ορισμένων ερευνητών απέναντι στη μεθοδολογία της τεχνητής νοημοσύνης. Ο τρόπος αναπαράστασης της γνώσης σ' αυτό το πρόγραμμα θέτει σε νέα βάση το πρόβλημα της σχέσης του λεξικού με την εγκυκλοπαίδεια. Είναι, πάντως, απαραίτητο να τονιστεί ότι η γνώση των λέξεων είναι στενά συνδεδεμένη με τη γνώση του κόσμου και γι' αυτό δεν υπάρχει λόγος να διαχωρίζεται το λεξικό από την εγκυκλοπαίδεια.

3.3. Μελλοντικές προοπτικές

Οι προοπτικές για το μέλλον είναι λαμπρές, καθώς υπάρχει διεθνώς μεγάλο ενδιαφέρον για την προώθηση της ηλεκτρονικής λεξικογραφίας σε θεωρητικό και πρακτικό επίπεδο.[13] *Η Γλωσσική τεχνολογία*[14] (Language engineering) άρχισε να αναπτύσσεται με ταχύτατους ρυθμούς και στην Ελλάδα με εντυπωσιακά ερευνητικά αποτελέσματα και πληθώρα ερευνητικών προγραμμάτων που βρίσκονται σε εξέλιξη, όπως μπορεί να διαπιστώσει κανείς, ξεφυλλίζοντας απλώς τα τρία προηγούμενα τεύχη της *Λογοπλοήγησης* (1996-1997), του Ενημερωτικού Δελτίου Ανθρώπινου Δικτύου Γλωσσικής Τεχνολογίας που εκδίδει το Ινστιτούτο Επεξεργασίας του Λόγου (ΙΕΛ) με επιστημονική ευθύνη του Διευθυντή του Ινστιτούτου, καθηγητή Γιώργου Καραγιάννη. Ιδιαίτερα αισιόδοξες είναι οι προοπτικές για την ανάπτυξη ειδικά των ηλεκτρονικών εκπαιδευτικών λεξικών, όπως είναι το πρωτοποριακό πρόγραμμα *Λεξιπαιδεία* του ΙΕΛ, το *Ηλεκτρονικό λεξικό προφοράς και χρήσης της σύγχρονης ελληνικής γλώσσας για ξένους* του Πανεπιστημίου Πατρών κ.ά.

Ένα νέο ερευνητικό πεδίο έχει εμφανιστεί τα τελευταία μόλις χρόνια και ήδη έχει λάβει εκρηκτικές διαστάσεις: Πρόκειται για την *κυβερνολεξικογραφία* (Cyberlexicography), η οποία ασχολείται με τη συμπίληση ("συρραφή") ή δημιουργία λεξικών με βάση τα δεδομένα και τις πληροφορίες που παρέχει το Internet. Οι διαθέσιμοι κυβερνολεξικογραφικοί πόροι είναι εντυπωσιακοί. (Βλ. τον σχετικό κατάλογο με on line λεξικά και εγκυκλοπαίδειες στον Carr, 1997). Τα ειδικά υπολογιστικά προγράμματα για την αναζήτηση λέξεων - κλειδιών στο Internet, οι γνωστές "μηχανές" (engines), παρέχουν στους λεξικογράφους, τους γλωσσολόγους και στους λεξικόφιλους (lexicophiles) πρόσβαση σε εκατομμύρια πολύγλωσσα κείμενα με δισεκατομμύρια λέξεις. Μια νέα εξέλιξη στη λεξικογραφική πρακτική είναι η "από κάτω προς τα πάνω" λεξικογραφία (Bottom-up lexicography). Πριν από την εμφάνιση του Internet η σύνταξη των λεξικών γινόταν "από πάνω προς τα κάτω", από τους αρχισυντάκτες και τους εκδότες, για να φτάσει το προϊόν στους χρήστες. Τώρα εμφανίζεται δειλά η *συνεργατική λεξικογραφία* (collaborative lexicography) με το ηλεκτρονικό ταχυδρομείο. Η συμβολή χρηστών του Internet στην ενημέρωση και βελτίωση λημμάτων του ηλεκτρονικού λεξικού μπορεί να αποβεί καθοριστική ως προς την άμεση συμμετοχή των πολιτών σε θέματα γλωσσικής χρήσης.

[13] Για τις σχετικές εξελίξεις σχετικά με τη σύνταξη μηχανικά αναγνώσιμων λεξικών από σώματα κειμένων και άλλες πηγές, όπως και για τις μελλοντικές προοπτικές, βλ. Wilks κ.ά., 1996, 239-255.

[14] Η Ευρωπαϊκή Ένωση αποδίδει μεγάλη σημασία στην ανάπτυξη της γλωσσικής τεχνολογίας, όπως διαφαίνεται από τις δραστηριότητες της Επιτροπής (Διεύθυνση XIII, Λουξεμβούργο). Από το 1998 οι σχετικές δραστηριότητες εντάσσονται στην *Τεχνολογία της ανθρώπινης γλώσσας* (Human language technology). Βλ. Varile - Zampolli, 1997.

4. Βιβλιογραφία

Amsler, R. A., 1980: *Dictionary Databases,* Cambridge: Cambridge University, Computer Laboratory.

Atkins B. T. S. - Zamboli A. (εκδ.), 1994: *Computational approaches to the lexicon,* New York: Oxford University Press.

Carr M., 1997: "Internet dictionaries and lexicography", *International Journal of Lexicography*, 10, 209-221.

Κόντος Ι., 1997: "Γλωσσική τεχνολογία και ευφυείς πράκτορες", *Λογοπλοήγηση*, 3, 24-27.

Leech G. N., 1996: "Corpora", Malmkjaer K. (εκδ.), *The Linguistics encyclopedia,* Routledge: London and New York, 73-80.

Miller G., 1985: "Wordnet: a dictionary browser", *Proceedings of the First International Conference on Information in Data,* Waterloo, Ontario: University of Waterloo Centre for the New Oxford English Dictionary.

Procter P. (εκδ.), 1988: *Longman Dictionary of Contemporary English,* Harlow, Essex, England: Longman Group.

Ralli A. - Grigoriadou M. - Philokyprou G. - Christodoulakis D. - Galiotou E. (εκδ.), 1997: *Working papers in natural language processing,* Αθήνα: Εκδόσεις Δίαυλος.

Sinclair J. M. (εκδ.), 1987: *Cobuild Dictionary of the English language:* Glasgow: Collins.

Tomaszczyk J. - Lewardowska B. (εκδ.), 1991: *Meaning and lexicography,* Menlo Park, California: Benjamins.

Varile G. B. - Zampolli A. (εκδ.), 1997: Survey of the state of the art in human language technology, Pisa: Giardini editori e stampatori. [*Linguistica Computazionale*, τόμ.12-13].

Wilks Y. A. - Slator B. A. - Guthrie L. M., 1996: *Electric words. Dictionaries, Computers, and meanings*, Cambridge, Massachusetts: Massachusetts Institute of Technology.

Χαραλαμπάκης Χρ., 1997: *Νεοελληνικός λόγος. Μελέτες για τη γλώσσα, τη λογοτεχνία και το ύφος*, 2η έκδ. βελτιωμένη και επαυξημένη, Αθήνα.

2. Lexical Analysis and World Knowledge

Dr. Ioanna Malagardi
Institute for Language and Speech Processing
Artemidos 6 & Epidaurou, Paradeisos Amarousiou
151 25 Athens, Hellas
ioanna@ilsp.gr

Abstract

In the present paper we study methods that can be followed when a natural language processing system uses during the lexical processing phase a dictionary for the correlation of words. The result of this correlation is necessary for the following phases of syntactic, semantic and pragmatic analysis and the production of the final result. Two special cases are presented to illustrative cases which concern namely on the one hand the finding of implicit relations between nouns and on the other hand between a verb and its complement nouns.

Λεκτική Ανάλυση και Γνώση του Κόσμου

Δρ. Ιωάννα Μαλαγαρδή
Ινστιτούτο Επεξεργασίας του Λόγου
Αρτέμιδος 6 & Επιδαύρου, Παράδεισος Αμαρουσίου
151 25 Αθήνα

1. Εισαγωγή

Ένα σύστημα επεξεργασίας λόγου κατά τη φάση της λεκτικής επεξεργασίας κειμένου αξιοποιεί ένα λεξικό για τη συσχέτιση λέξεων που είναι απαραίτητη για τις περαιτέρω φάσεις όπως είναι η συντακτική, η σημασιολογική και η πραγματολογική ανάλυση για την παραγωγή του αναμενόμενου αποτελέσματος. Το αναμενόμενο αποτέλεσμα μπορεί να είναι απάντηση σε ερώτηση, εξαγωγή πληροφορίας, μετάφραση σε άλλη γλώσσα, περίληψη κειμένου κ.ά. [3]. Σε ορισμένες περιπτώσεις στις οποίες θα αναφερθούμε παρακάτω τα απαιτούμενα λεξικά πρέπει να περιέχουν και πληροφορίες ή γνώσεις σχετικές με

τον μικρόκοσμο στον οποίο λειτουργεί το σύστημα. Παραδοσιακή πηγή τέτοιων πληροφοριών είναι τα ερμηνεύματα των λεξικών [4]. Όμως η νεότερη εξέλιξη της γλωσσικής τεχνολογίας απαιτεί την αξιοποίηση και πρόσθετων γνώσεων που πρέπει να αξιοποιηθούν για τους σκοπούς της συγκεκριμένης εφαρμογής.

2. Παραδείγματα αξιοποίησης γνώσης του κόσμου

Δύο περιπτώσεις που έχουμε ασχοληθεί είναι ο προσδιορισμός της υπονοούμενης σχέσης μεταξύ ονομάτων και της υπονοούμενης σχέσης μεταξύ ρήματος και ονόματος λαμβάνοντας υπόψη τις κατάλληλες οντολογίες [6].

Για την πρώτη περίπτωση θα αναφερθούμε στο παράδειγμα του ζεύγους λέξεων *"αποθήκη ξυλείας"*. Το ζεύγος αυτό θα μπορούσε να σχετίζεται με ποικιλία σχέσεων όπως, *προέλευση, ιδιότητα, αιτία, περιεχόμενο* κ.ά. [11]. Κάθε μία από αυτές τις σχέσεις μπορεί να εξηγηθεί εάν παραφράσουμε το ζεύγος.

Εάν ως σχέση ληφθεί η προέλευση τότε θα έπρεπε να παραφραστεί ως εξής:
Η αποθήκη δημιουργείται από την ξυλεία.

Εάν ως σχέση ληφθεί η ιδιότητα τότε θα έπρεπε να παραφραστεί ως εξής:
Η αποθήκη είναι κατασκευασμένη από ξυλεία.

Εάν ως σχέση ληφθεί η αιτία τότε θα έπρεπε να παραφραστεί ως εξής:
Η αποθήκη προκύπτει από την ξυλεία.

Εάν ως σχέση ληφθεί το περιεχόμενο τότε θα έπρεπε να παραφραστεί ως εξής:
Η αποθήκη περιέχει ξυλεία ή η ξυλεία φυλάσσεται στη αποθήκη.

Η τελευταία σχέση που είναι η και η πιο αποδεκτή βάσει της γνώσης του μικρόκοσμου μπορεί να επιλεγεί με τον εξής αλγοριθμικό συλλογισμό:

Η αποθήκη είναι δοχείο. Η ξυλεία είναι ύλη. Τα δοχεία μπορεί να περιέχουν ύλη. Επομένως η αποθήκη περιέχει την ξυλεία.

Για τη δεύτερη περίπτωση θα αναφερθούμε στο εξής παράδειγμα: *"Η εταιρεία διέθεσε στην ομάδα έναν υπολογιστή για να τον χρησιμοποιήσει για έλεγχο"*. Το πρόβλημα στο παράδειγμα αυτό είναι η συσχέτιση του ρήματος *"χρησιμοποιήσει"* με τα *ονόματα* [5]. Στην περίπτωση αυτή αναζητείται από το σύστημα ο προσδιορισμός της σχέσης του ρήματος και κάποιων ονομάτων, δηλαδή ο καθορισμός των ορισμάτων του όπως του υποκειμένου και του αντικειμένου. Η σχέση μπορεί να επιλεγεί με τον εξής αλγοριθμικό συλλογισμό:

Τα ονόματα που εμφανίζονται στην πρόταση είναι τα εξής: εταιρεία, ομάδα, υπολογιστής και έλεγχος. Από τη χρησιμοποιούμενη οντολογία προκύπτει ότι η ομάδα και η εταιρεία είναι φορείς, ο υπολογιστής είναι μηχανή και ο έλεγχος μία δράση. Από τη γνώση του μικρόκοσμου προκύπτει ότι υποκείμενο του ρήματος "χρησιμοποιήσει" πρέπει να είναι φορέας και αντικείμενο πρέπει να είναι κάτι άψυχο εξαιρουμένων των περιπτώσεων μεταφοράς. Επομένως η σχέση μεταξύ εταιρείας ή η ομάδας με το ρήμα θα είναι σχέση υποκειμένου ενώ η σχέση του ρήματος με τον υπολογιστή θα είναι σχέση αντικειμένου. Η τελική επιλογή του υποκειμένου απαιτεί τη χρήση πολυπλοκότερης γνώσης, τύπου σεναρίου, βάσει της οποίας συνάγεται ότι κάτι που δίνει κάποιος δεν μπορεί στη συνέχεια να το χρησιμοποιήσει. Δηλαδή ο χρήστης πρέπει να είναι ο αποδέκτης του αντικειμένου που θα χρησιμοποιήσει και όχι ο χορηγός. Επομένως η ομάδα είναι το όνομα που σχετίζεται με το ρήμα "χρησιμοποιήσει" με σχέση υποκειμένου και ο υπολογιστής είναι το όνομα που σχετίζεται με το ρήμα "χρησιμοποιήσει" με σχέση αντικειμένου.

3. Άντληση των απαιτούμενων γνώσεων από λεξικά

Οι γνώσεις που αναφέρθηκαν παραπάνω μπορούν να αντληθούν από διάφορες πηγές και με

διάφορους τρόπους. Μία σημαντική πηγή για την άντληση γνώσεων, όπως αυτές που αναφέρθηκαν παραπάνω είναι τα παραδοσιακά ερμηνευτικά και εννοιολογικά λεξικά [1], [2], [7], [8], [9], [10]. Είναι δυνατόν σε ορισμένες περιπτώσεις να απαιτείται η συμπλήρωση της προσφερόμενης από τα λεξικά γνώσης είτε με γνώσεις της καθημερινής ζωής τις οποίες προϋποτίθεται ότι έχει ο αναγνώστης του λεξικού, όχι όμως ο υπολογιστής, είτε με επιστημονικοτεχνικές γνώσεις που πρέπει να αντληθούν από τους ειδικούς. Οι γνώσεις αυτές πρέπει να παρασταθούν με τον κατάλληλο τρόπο, ώστε να είναι εφικτό να χρησιμοποιηθούν από ένα αυτόματο σύστημα επεξεργασίας λόγου. Για την αποσαφήνιση των παραπάνω θα παρουσιαστούν τα αποτελέσματα της μελέτης παραδειγμάτων που προέρχονται από διάφορα λεξικά, καθώς επίσης και οι τυχόν απαιτούμενες πρόσθετες γνώσεις, ώστε να είναι δυνατή η αυτόματη επίλυση των προβλημάτων του προσδιορισμού της σχέσης μεταξύ των λέξεων. Για τη μελέτη αυτή χρησιμοποιήθηκαν τα εμφανιζόμενα ως πρώτα κατά σειρά ερμηνεύματα των λημμάτων. Πέραν των ερμηνευμάτων αυτών προστέθηκαν και άλλα κατά περίπτωση ώστε να γίνει εμφανές το πρόβλημα της ενδεχόμενης αμφισημίας που προστίθεται στα παραπάνω προβλήματα. Τα ερμηνεύματα ελήφθησαν από τα εξής Λεξικά: Δ. Δημητράκου (Δ.Δ.): 1956, Σταματάκου (Ι.Σ.): 1971, Ι. Τεγόπουλου - Φυτράκη (Τ. - Φ.): 1997 και Γ. Μπαμπινιώτη (Γ.Μ.):1998. Η σειρά που αναφέρονται τα λεξικά καθώς και τα ερμηνεύματα των λημμάτων τους παρακάτω είναι χρονολογική.

- **αποθήκη**
- μέρος ένθα αποτίθενται πράγματα προς φύλαξιν (Δ.Δ.)
- χώρος ασφαλής εν ω αποτίθενται προς φύλαξιν διάφορα είδη (Ι.Σ.)
- κλειστός χώρος, κτίσμα, όπου τοποθετούνται διάφορα είδη για φύλαξη (Τ.-Φ.)
- ο χώρος ή το κτίσμα μέσα στο οποίο φυλάσσονται διάφορα είδη (Γ.Μ.)

- **έλεγχος**
- επιχείρημα προς ανακατασκευήν (Δ.Δ.)
- η έρευνα περί της αληθείας ή ψεύδους, περί της αξίας ή απαξίας κ.ο.κ. προσώπου τινός ή πράγματος (Ι.Σ.)
- έρευνα για την αλήθεια, την αιτία, την ορθότητα, την ικανότητα, τη λειτουργία αντικειμένων, προσώπων, καταστάσεων, ενεργειών (Τ.-Φ.)
- η έρευνα για τη εξακρίβωση ή πιστοποίηση της εγκυρότητας, της ορθότητας, της αλήθειας ή της πραγματικής αξίας (στοιχείων, δεδομένων, λόγων, ενεργειών κ.λπ.) (Γ.Μ.)

- **εταιρεία**
- όμιλος ανθρώπων αποτελούντων σωματείον επί τινι σκοπώ (Δ.Δ.)
- συνεταιρισμός προσώπων προς διεξαγωγήν εμπορικών πράξεων (Δ.Δ.)
- ομάς συνεταίρων, όμιλος ανθρώπων αποτελούντων σύνδεσμον επί τινι σκοπώ (Ι.Σ.)
- ομάδα ανθρώπων που συνεργάζονται για την επίτευξη κοινού σκοπού (Τ.-Φ.)
- η ένωση προσώπων για την επιδίωξη κοινού κέρδους με κοινές, ίσες ή άνισες εισφορές (Τ-Φ.)
- ομάδα προσώπων που έχουν κοινούς στόχους και συνεργάζονται για την πραγματοποίησή τους (Γ.Μ.)
- η σύμβαση, με τη οποία δύο ή περισσότερα πρόσωπα ενώνονται για την επιδίωξη κοινού σκοπού, ιδίως οικονομικού (Γ.Μ.)

- **ξυλεία**
- το σύνολον των ξύλων των προερχομένων εξ υλοτομίας δασών (Δ.Δ.)
- το σύνολον των ξύλων των προερχομένων εξ υλοτομίας δασών (Ι.Σ.)
- το σύνολο των ξύλων που προέρχονται από την υλοτομία (Τ.-Φ.)
- το σύνολο των ξύλων που προέρχονται από την υλοτομία των δασών (Γ.Μ.)

- **ομάδα**
- σύνολον, άθροισμα ομοειδών (Δ.Δ.)
- ένωσις ανθρώπων συνδεομένων δια κοινού

έργου ή σκοπού (Δ.Δ.)
- άθροισμα ατόμων ή ομοειδών πραγμάτων λαμβανόμενον ως εν όλον (Ι.Σ.)
- άθροισμα ατόμων ή ομοειδών πραγμάτων που λαμβάνονται ως ενιαίο σύνολο (Τ.-Φ.)
- άθροισμα προσώπων ή (σπαν.) πραγμάτων, τα οποία συνδέει κάτι κοινό και εκλαμβάνονται ως ενιαίο σύνολο (Γ.Μ.)

- **<u>υπολογιστής</u>**
- βοηθητικός λογιστής (Δ.Δ.)
- λογιστικός υπάλληλος (Ι.Σ.)
- ηλεκτρονικός υπολογιστής: η μηχανή που λειτουργεί με συγκεκριμένο πρόγραμμα, εκτελεί με ταχύτητα και ακρίβεια υπολογισμούς, και αποθηκεύει στοιχεία στη μνήμη τα οποία μπορεί να ανακληθούν (Τ.-Φ.)
- αυτός που ενεργεί με βάση τα ιδιοτελή του συμφέροντα (Γ.Μ.)
- ηλεκτρονικός υπολογιστής: η ηλεκτρονική συσκευή, με την οποία χειρίζεται και επεξεργάζεται κανείς πληροφορίες και δεδομένα με υψηλή ταχύτητα και ακρίβεια, με βάση συγκεκριμένα κάθε φορά προγράμματα (Γ.Μ.)

- **<u>χρησιμοποιώ</u>**
- ποιούμαι χρήσιν τινός, μεταχειρίζομαί τι (Δ.Δ.)
- κάμνω χρήσίν τινος (Ι.Σ.)
- κάνω χρήση ενός πράγματος, μεταχειρίζομαι κάτι για έναν σκοπό (Τ.-Φ.)
- εκμεταλλεύομαι (Τ.-Φ.)
- μεταχειρίζομαι (κάτι) για να επιτύχω ορισμένο σκοπό (Γ.Μ.)
- κάνω συστηματική χρήση (Γ.Μ.)
- εκμεταλλεύομαι (κάποιον, κάτι) για δική μου ωφέλεια (Γ.Μ.)

Εξετάζοντας τους παραπάνω συλλογισμούς για την ερμηνεία των δύο προτάσεων

Α. Έκτισε αποθήκη ξυλείας.
Β. Η εταιρεία διέθεσε στην ομάδα έναν υπολογιστή για να τον χρησιμοποιήσει για έλεγχο.
η γνώση η απαιτούμενη γι' αυτούς έχει ως εξής:

Α. Γνώση για την απόδειξη ότι η *αποθήκη περιέχει την ξυλεία*:
1. Η αποθήκη είναι δοχείο.
2. Η ξυλεία είναι ύλη.
3. Τα δοχεία μπορεί να περιέχουν ύλη.

Β. Γνώση για την απόδειξη ότι *η ομάδα είναι το όνομα που σχετίζεται με το ρήμα "χρησιμοποιήσει" με σχέση υποκειμένου*:
1. Η ομάδα και η εταιρεία είναι φορείς.
2. Ο υπολογιστής είναι μηχανή.
3. Ο έλεγχος είναι μία δράση.
4. Υποκείμενο του ρήματος "χρησιμοποιήσει" πρέπει να είναι φορέας.
5. Αντικείμενο του ρήματος "χρησιμοποιήσει" πρέπει να είναι κάτι άψυχο.
6. Κάτι που δίνει κάποιος δεν μπορεί στη συνέχεια να το χρησιμοποιήσει
7. Ο χρήστης πρέπει να είναι ο αποδέκτης του αντικειμένου που θα χρησιμοποιήσει.

Το επόμενο βήμα είναι η άντληση της γνώσης αυτής από τα ερμηνεύματα των λεξικών. Ως παράδειγμα θα εξεταστεί η γνώση που αναφέρεται στην περίπτωση Α. Παίρνοντας την περίπτωση της οντολογικής γνώσης Α.2. αναζητούμε τον τρόπο εξαγωγής από τα ερμηνεύματα των λεξικών της πληροφορίας ότι δεδομένης της *ξυλείας* αυτή ανήκει στην κατηγορία των υλικών αντικειμένων ή αλλιώς είναι *ύλη*, χωρίς αυτό να έχει δοθεί ρητά. Τα ερμηνεύματα της λέξης *ξυλεία*, όπως αναφέρονται παραπάνω περιέχουν ως κοινό πυρήνα της ονοματικής φράσης τη φράση *το σύνολο των ξύλων.*

Αναζητώντας στη συνέχεια το ερμήνευμα της λέξης ξύλο στα λεξικά βρίσκουμε τα εξής:

ξύλο
- η υπό τον φλοιόν των δένδρων ινώδης και σκληρά ουσία (Δ.Δ.)
- σκληρά και ινώδης ουσία η ευρισκομένη υπό τον φλοιόν των δένδρων (Ι.Σ.)
- ινώδης και σκληρή ουσία (Φ.-Τ.)

- σκληρή, ινώδης ύλη μέσα από τον εξωτερικό φλοιό δένδρων ή θάμνων (Γ.Μ.)

Από τα παραπάνω μπορεί να εξαχθεί ότι η *ξυλεία είναι ύλη* λαμβάνοντας υπόψη τα ερμηνεύματα που δίδονται στα Λεξικά για τις λέξεις *ξυλεία* και *ξύλο*. Η εξαγωγή αυτή πρέπει να στηριχθεί στο εξής σχήμα συμπερασμού: *Αν Χ είναι Ψ και Ψ είναι Ζ ⇒ Χ είναι Ζ.* Για τον άνθρωπο η χρήση του σχήματος είναι αυτονόητη. Για την υλοποίηση με υπολογιστή δεν είναι αυτονόητη και πρέπει να εμφανισθεί ως μέρος του προγράμματος. Παρόμοιες εργασίες εξαγωγής της γνώσης από λεξικά μπορούν να γίνουν και για την υπόλοιπη απαιτούμενη γνώση που αναφέραμε παραπάνω. Η σχετική έρευνα και υλοποίηση για τα θέματα αυτά βρίσκεται σε εξέλιξη. Παρόμοιες έρευνες έχουν γίνει μόνο για ξένες γλώσσες [12].

4. Επίλογος

Η λεκτική ανάλυση είναι απαραίτητη φάση λειτουργίας ενός συστήματος επεξεργασίας φυσικής γλώσσας. Όταν η λεκτική ανάλυση αφορά υπονοούμενες σχέσεις μεταξύ λέξεων τότε απαιτείται γνώση του κόσμου. Η εξαγωγή γνώσης του κόσμου που απαιτείται για τη λεκτική ανάλυση κειμένων είναι δυνατόν να γίνει από τα ερμηνεύματα λεξικών. Η εξαγωγή αυτή είναι χρήσιμη για τη λειτουργία ενός συστήματος επεξεργασίας λόγου με τελικό σκοπό την απάντηση ερωτήσεων, την εξαγωγή πληροφορίας, τη μετάφραση σε άλλη γλώσσα, την περίληψη κειμένου κ.ά. Η εξαγωγή αυτή μπορεί να γίνει είτε από τον δημιουργό του συστήματος επεξεργασίας λόγου μελετώντας τα σχετικά λεξικά, είτε από ένα σύστημα εξαγωγής πληροφορίας και γνώσης από λεξικά σε μορφή αναγνώσιμη από μηχανή.

5. Βιβλιογραφία

[1]. Βοσταντζόγλου Θ. 1962:
Αντιλεξικόν ή Ονομαστικόν της Νεοελληνικής Γλώσσης. (Αθήνα).

[2]. Δημητράκος Δ. 1956: *Επίτομον Λεξικόν Ορθογραφικόν- Ερμηνευτικόν όλης της Ελληνικής Γλώσσης.*
(Αθήνα: Εκδόσεις Δ. Δημητράκου).

[3]. Κόντος Ι. 1996: *Τεχνητή Νοημοσύνη και Λογομηχανική* (Επεξεργασία Λόγου).
(Αθήνα: Εκδόσεις Ε. Μπένου).

[4]. Κόντος Ι. & Μαλαγαρδή Ι. & Πέγκου Μ. 1997:
"Επεξεργασία Ερμηνευμάτων Ρημάτων με Υπολογιστή"
3ο Διεθνές συνέδριο για την Ελληνική Γλώσσα.
Πανεπιστήμιο Αθηνών. Φιλοσοφική Σχολή.
Τμήμα Φιλολογίας Τομέας Γλωσσολογίας.
(Αθήνα). (προς δημοσίευση).

[5]. Μαλαγαρδή Ι. 1995: *Συγκριτική Ανάλυση "να" και "για να" Δομών της Νέας Ελληνικής με Αντίστοιχες Δομές της Γερμανικής και Εφαρμογή στη Μηχανική Μετάφραση.*
Αδημοσίευτη Διδακτορική Διατριβή.
Πανεπιστήμιο Αθηνών.

[6]. Μαλαγαρδή Ι. 1996:
Προσδιορισμός με Υπολογιστή της Υπονοούμενης Σχέσης μεταξύ των Συστατικών Ονοματικών Φράσεων σε Υπογλώσσες.
17η Συνάντηση Εργασίας ΑΠΘ, Θεσσαλονίκη.

[7]. Μπαμπινιώτης Γ. 1998:
Λεξικό της Νέας Ελληνικής Γλώσσας. Ερμηνευτικό, Ορθογραφικό, Ετυμολογικό, Συνωνύμων- Αντιθέτων κ.ά
(Αθήνα: Κέντρο Λεξιλογίας)

[8]. *Roget's International Thesaurus.*1977:
Fourth Edition. Revised by R. L. Chapman.
(Harper Collins Publishers: London and Glasgow).

[9]. Σταματάκος Ι. 1971:
Λεξικόν της Νέας Ελληνικής Γλώσσης Καθαρευούσης και Δημοτικής
(Αθήνα: Εκδοτικός Οργανισμός "Ο Φοίνιξ").

[10]. Τεγόπουλος- Φυτράκης 1997:
Μείζον Ελληνικό Λεξικό
(Αθήνα: Εκδόσεις Αρμονία Α.Ε.).

[11]. Τζάρτζανος Α. 1989:
Νεοελληνική Σύνταξις. Τομος Β
(Θεσσαλονίκη: Εκδόσεις Αφοι Κυριακίδη).

[12]. Wilks Y. A. et al.: 1996, *Electric Words: Dictionaries, Computers, and Meanings.*
(ACL-MIT Press: Cambridge, Mass. & London, England).

3. Brief Description of EC-Systran

Olga Yannoutsou, Linguist and
Athanassia Fourla, Translator
Institute for Language and Speech Processing
Liaison Department- EUROMAT
Artemidos 6 & Epidavrou
Paradeissos Amaroussiou 15125, Hellas
email: soula@ilsp.gr email: olga@ilsp.gr

Abstract

The present paper briefly presents EC-Systran, the Machine Translation System of the European Commission. The system structure is described and special reference is made to its dictionaries as well as to the development of the English into Greek language pair.

Συνοπτική Περιγραφή του EC-Systran

Όλγα Γιαννούτσου, Γλωσσολόγος και
Αθανασία Φούρλα, Φιλόλογος-Μεταφράστρια
Ινστιτούτο Επεξεργασίας του Λόγου
Τμήμα Συνδέσμου- Γραφείο EUROMAT
Αρτέμιδος 6 & Επιδαύρου 15125, Παράδεισος Αμαρουσίου

Περίληψη

Το παρόν κείμενο παρουσιάζει συνοπτικά το σύστημα Μηχανικής Μετάφρασης της Ευρωπαϊκής Επιτροπής, EC-Systran. Περιγράφεται η δομή του συστήματος και γίνεται ιδιαίτερη μνεία στα λεξικά του καθώς και στην ανάπτυξη του ζεύγους Αγγλικά προς Ελληνικά.

EC-Systran: Γενικά

Το σύστημα Μηχανικής Μετάφρασης της Επιτροπής, EC-Systran, από τη δεκαετία του 70 και μετά αναπτύσσεται από την Ευρωπαϊκή Επιτροπή για τις εσωτερικές της ανάγκες από γλωσσολόγους και ειδικούς στον τομέα της τεχνολογίας των πληροφοριών. Τα γλωσσικά ζεύγη με τα οποία μεταφράζει σήμερα το σύστημα Μηχανικής Μετάφρασης της Επιτροπής είναι τα εξής:

- Αγγλικά προς Γαλλικά, Ιταλικά, Γερμανικά, Ολλανδικά, Ισπανικά, Πορτογαλικά, **Ελληνικά**
- Γαλλικά προς Ισπανικά, Αγγλικά, Γερμανικά, Ολλανδικά, Ιταλικά
- Γερμανικά προς Αγγλικά, Γαλλικά
- Ισπανικά προς Γαλλικά, Αγγλικά
- Ελληνικά προς Γαλλικά (το ζεύγος βρίσκεται στην αρχή της ανάπτυξής του και δεν είναι ακόμη λειτουργικό).

Το σύστημα Μηχανικής Μετάφρασης της Επιτροπής είναι το μόνο λειτουργικό σύστημα αυτόματης μετάφρασης που έχει στα γλωσσικά του ζεύγη την Ελληνική γλώσσα.

Η ποιότητα μετάφρασης του συστήματος ποικίλλει σημαντικά ανάλογα με το ζεύγος γλωσσών που χρησιμοποιείται, το θέμα και το είδος του εγγράφου. Το σύστημα Μηχανικής Μετάφρασης της Επιτροπής όπως και όλα τα συστήματα αυτόματης μετάφρασης μεταφράζει καλά κείμενα με σταθερή δομή και ορολογία, δεν μεταφράζει για παράδειγμα λογοτεχνικά κείμενα, ή προφορικό λόγο. Η ανάπτυξη του συστήματος έχει βασιστεί σε κείμενα της Επιτροπής, τα περισσότερα εκ των οποίων είναι διοικητικά και αναφέρονται σε κοινοτικά προγράμματα και ως εκ τούτου το σύστημα έχει υψηλότερο ποσοστό κατανόησης σε τέτοιου είδους κείμενα.

Η ανάπτυξη του ζεύγους Αγγλικά -> Ελληνικά, που είχε πρωτοξεκινήσει στην Αμερική, από το 1989 συνεχίστηκε στο Λουξεμβούργο με συγχρηματοδότηση από τη Γενική Γραμματεία Έρευνας και Τεχνολογίας και από την Επιτροπή Ευρωπαϊκών Κοινοτήτων στο πλαίσιο προώθησης των Ελληνικών στην Ευρώπη (τα υπόλοιπα ζεύγη χρηματοδοτούνται 100% από την Κοινότητα).

Τον Σεπτέμβριο του 1994 μετά από απόφαση της ΓΓΕΤ, τμήμα της Ελληνικής ομάδας ανάπτυξης συνεχίζει το έργο της στο Ινστιτούτο Επεξεργασίας του Λόγου, τόσο για να εξυπηρετήσει τις ανάγκες του Δημόσιου Τομέα της χώρας παρέχοντας δωρεάν μεταφραστικές υπηρεσίες μέσω

του συστήματος, όσο και για να συνεχίσει την ανάπτυξη του ζεύγους Αγγλικά->Ελληνικά σύμφωνα με τις ανάγκες του.

Από το 1994 μέχρι σήμερα η ομάδα ανάπτυξης της Αθήνας έχει συνεργαστεί με υπηρεσίες από όλα σχεδόν τα Ελληνικά Υπουργεία και έχει αναπτύξει τα λεξικά του συστήματος επικεντρώνοντας ιδιαίτερα στις ανάγκες ορολογίας στους τομείς της Γεωργίας, του Αθλητισμού, της Πυρόσβεσης, της Εκπαίδευσης και της Πληροφορικής.

Τεχνικά χαρακτηριστικά του EC-SYSTRAN

Το EC-SYSTRAN αποτελείται από τρία βασικά τμήματα:

1. το βασικό σύστημα που ελέγχει
2. τα γλωσσικά προγράμματα (ρουτίνες) και
3. τα λεξικά του

Το βασικό σύστημα είναι γραμμένο σε assembler, είναι βασισμένο σε αρχιτεκτονική IBM και στο λειτουργικό σύστημα MVS. Το EC-SYSTRAN έχει μικτά χαρακτηριστικά τόσο από συστήματα μηχανικής μετάφρασης σχεδιασμένα με transfer approach, όσο και από συστήματα σχεδιασμένα με direct approach.

Στο παρακάτω σχήμα φαίνεται η διαδικασία μετάφρασης ενός κειμένου από το σύστημα.

ΔΙΑΔΙΚΑΣΙΑ ΜΕΤΑΦΡΑΣΗΣ
ΕΙΣΑΓΩΓΗ ΚΕΙΜΕΝΟΥ ΠΡΟΣ ΜΕΤΑΦΡΑΣΗ
ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ
Καθορισμός των μεταφραστικών ενοτήτων (προτάσεων)
ΑΝΙΧΝΕΥΣΗ ΛΕΞΕΩΝ ΤΟΥ ΚΕΙΜΕΝΟΥ ΣΤΟ ΒΑΣΙΚΟ ΛΕΞΙΚΟ Ανεύρεση μεμονωμένων λέξεων
ΠΡΩΤΗ ΑΝΙΧΝΕΥΣΗ ΕΚΦΡΑΣΕΩΝ ΤΟΥ ΚΕΙΜΕΝΟΥ ΣΤΟ ΕΙΔΙΚΟ ΛΕΞΙΚΟ Ανεύρεση των συνδυασμών δύο ή περισσοτέρων λέξεων οι οποίες έχουν κωδικοποιηθεί στο ειδικό λεξικό ανεξάρτητα από το γλωσσικό τους περιβάλλον
ΑΝΑΛΥΣΗ
ΕΠΙΛΥΣΗ ΟΜΟΓΡΑΦΩΝ Διαλεύκανση της γραμματικής κατηγορίας ομόγραφων λέξεων
ΑΝΙΧΝΕΥΣΗ ΕΚΦΡΑΣΕΩΝ ΤΟΥ ΚΕΙΜΕΝΟΥ ΣΤΟ ΕΙΔΙΚΟ ΛΕΞΙΚΟ Ανεύρεση των εκφράσεων που έχουν κωδικοποιηθεί με βάση το γλωσσικό περιβάλλον της γλώσσας-πηγής
Α' ΓΛΩΣΣΟΛΟΓΙΚΗ ΑΝΑΛΥΣΗ Αναγνώριση και χαρακτηρισμός κυρίων και δευτερευουσών προτάσεων
ΑΝΑΓΝΩΡΙΣΗ ΤΩΝ ΕΠΙΦΑΝΕΙΑΚΩΝ ΣΥΝΤΑΚΤΙΚΩΝ ΔΟΜΩΝ
ΑΝΙΧΝΕΥΣΗ ΤΩΝ ΠΑΡΑΤΑΚΤΙΚΩΝ ΔΟΜΩΝ
Β' ΓΛΩΣΣΟΛΟΓΙΚΗ ΑΝΑΛΥΣΗ Αναγνώριση των βαθειών συντακτικών δομών μεταξύ μεμονωμένων λέξεων και συνδυασμών λέξεων
ΜΕΤΑΦΟΡΑ
ΑΝΙΧΝΕΥΣΗ ΕΚΦΡΑΣΕΩΝ ΤΟΥ ΚΕΙΜΕΝΟΥ ΣΤΟ ΕΙΔΙΚΟ ΛΕΞΙΚΟ Ανεύρεση των εκφράσεων που έχουν κωδικοποιηθεί με βάση το γλωσσικό περιβάλλον της γλώσσας-στόχου
ΜΕΤΑΦΡΑΣΗ ΤΩΝ ΠΡΟΘΕΣΕΩΝ
ΕΝΕΡΓΟΠΟΙΗΣΗ ΡΟΥΤΙΝΩΝ
ΣΥΝΘΕΣΗ
ΜΕΤΑΒΑΣΗ ΣΤΗ ΓΛΩΣΣΑ-ΣΤΟΧΟ Απόδοση μεμονωμένων λέξεων και συντακτικών σχημάτων στη γλώσσα-στόχο
ΜΟΡΦΟΛΟΓΙΚΗ ΠΑΡΑΓΩΓΗ Μορφολογική σύνθεση των λέξεων στη γλώσσα-στόχο
ΑΝΑΔΙΑΤΑΞΗ Τοποθέτηση των λέξεων με τους κανόνες που καθορίζουν τη σειρά όρων στη γλώσσα-στόχο
ΤΕΛΙΚΗ ΕΠΕΞΕΡΓΑΣΙΑ Αποκατάσταση της μορφής του κειμένου
ΕΞΑΓΩΓΗ ΜΕΤΑΦΡΑΣΘΕΝΤΟΣ ΚΕΙΜΕΝΟΥ

Τα Λεξικά του EC-SYSTRAN

Σημαντικό ρόλο για την κωδικοποίηση στα λεξικά του συστήματος παίζει η γραμμική δόμηση της πληροφορίας, όπου σε κάθε λέξη αντιστοιχεί ένα σύνολο από bytes καθένα από τα οποία μεταφέρει ένα μέρος πληροφορίας συντακτικής/γραμματικής/σημασιολογικής κ.λπ. Η αποκωδικοποίηση των πληροφοριών που φέρει κάθε byte βοηθά στην επιτυχή αντιμετώπιση μεταφραστικών προβλημάτων και στον ακριβή εντοπισμό της αιτίας τους.

Τα βασικά λεξικά του συστήματος είναι δύο: το STEM και το IDLS και πρόκειται για λεξικά μιας κατεύθυνσης.

Α. Το STEM

Το STEM είναι λεξικό που έχει μονολεκτικές καταχωρήσεις μόνο στη γλώσσα πηγή και εκεί συμπληρώνονται οι μεταφράσεις των λέξεων σε όλες τις γλώσσες στόχους. Υπάρχει δηλαδή ένα Αγγλικό STEM για όλες τις γλώσσες και σε κάθε Αγγλική λέξη υπάρχουν τα μεταφραστικά αντίστοιχα για τις άλλες γλώσσες.

Όλες οι γραμματικές και συντακτικές πληροφορίες βρίσκονται κωδικοποιημένες στα λεξικά του συστήματος Μηχανικής Μετάφρασης της Επιτροπής.

Μία λέξη κωδικοποιείται πάντα βάσει των κειμένων που υπάρχουν και όχι, για παράδειγμα, ανοίγοντας ένα λεξικό. Πριν αποδοθεί σημασία σε μία λέξη γίνεται ηλεκτρονική αναζήτηση των εμφανίσεων του εν λόγω λήμματος στα υπάρχοντα κείμενα. Δηλαδή επισημαίνονται οι εμφανίσεις του λήμματος στο εκάστοτε περιβάλλον του, έτσι ώστε η μετάφραση που θα αποδοθεί να καλύπτει τις περισσότερες των περιπτώσεων. Για τις υπόλοιπες περιπτώσεις, η λέξη κωδικοποιείται ως έκφραση. Έτσι λοιπόν ερμηνεύεται το πώς το σύστημα "μαθαίνει" να μεταφράζει βάσει των κειμένων που ήδη υπάρχουν. Για παράδειγμα η λέξη minutes στα λεξικά του EC-SYSTRAN έχει αποδοθεί με τη σημασία "πρακτικά συνεδριάσεων" και όχι "λεπτά" γιατί στις περισσότερες περιπτώσεις σημαίνει πρακτικά, εφόσον στο σύστημα μεταφράζονται κατά κύριο λόγο διοικητικά κείμενα. Βέβαια, υπάρχει και η σημασία "λεπτά", αλλά κωδικοποιημένη ως κανόνας με άλλες παραμέτρους. Επειδή, τα περισσότερα κείμενα είναι διοικητικά, το EC-SYSTRAN έχει προσανατολιστεί στη μετάφραση τέτοιων κειμένων.

Το STEM έχει τις εξής πληροφορίες:

- τη λέξη
- βασικές γραμματικές πληροφορίες, πχ. αριθμός, πτώση, γένος, χρόνος ρήματος κλπ
- κώδικα δήλωσης ομογραφίας, πχ. ομόγραφο μεταξύ ουσιαστικού και απαρεμφάτου
- σημασιολογικές πληροφορίες, π.χ μέλος σώματος, άνθρωπος, επάγγελμα, ασθένεια κλπ
- συντακτικές πληροφορίες, π.χ. έμψυχο υποκείμενο, πρόθεση με την οποία συντάσσεται, μεταβατικό ρήμα κλπ

Με τον καιρό ο αριθμός των σημασιολογικών και συντακτικών κωδίκων αυξάνεται με στόχο την κάλυψη όλων περιπτώσεων. Ιδιαίτερη σημασία δίνεται στους κώδικες των ομογράφων για την αποτελεσματικότερη επίλυση των αμφισημιών. Το πλήθος των κωδίκων εξαρτάται από την ανάλυση της γλώσσας πηγής.

Παράδειγμα κωδικοποίησης ουσιαστικού στο STEM

```
F  -A00CARTOONIST
F  AG10
F  BA 1010  11
F  CA HU,CON,CT,
F  DA PROFES
F  FAH00N1  1  SHEDJASTI'S KJNOYME'NWN SHEDJ'WN
```

Ερμηνεία κωδίκων στο παραπάνω παράδειγμα:
Γενικά, κάθε γραμμή κωδικοποίησης αριθμείται με ένα γράμμα A,B,C,D κλπ όπως φαίνεται παραπάνω. Η κωδικοποίηση μπορεί να ερμηνευτεί ως εξής για κάθε σειρά.

A00 = εισαγωγή νέας λέξης στο λεξικό. Ο κώδικας 00 αντιπροσωπεύει τη συχνότητα εμφάνισης της λέξης. Στην περίπτωση που η λέξη ανήκει σε περισσότερες από μία γραμματικές κατηγορίες, τότε 00 παίρνει η συχνότερα εμφανιζόμενη κατηγορία, 01 η αμέσως επόμενη σε συχνότητα κ.ο.κ.

G11= πρόκειται για το κλιτικό παράδειγμα του ουσιαστικού.

BA 1010 = είναι ο κώδικας που αντιπροσωπεύει το μέρος του λόγου στο οποίο ανήκει η λέξη. Οι άλλοι κώδικες δίπλα του (1 1) αντιπροσωπεύουν το γένος και τον αριθμό.

CA HU, CON, CT = πρόκειται για συντακτικούς κώδικες, π.χ. ο κώδικας HU δηλώνει ότι η λέξη αναφέρεται σε άνθρωπο, ο CON ότι πρόκειται για κάτι συγκεκριμένο, και ο CT ότι είναι μετρήσιμο.

DA PROFES = ο κώδικας αυτός είναι σημασιολογικός και δηλώνει επάγγελμα.

Τέλος, στην Ελληνική σειρά κωδικοποίησης, εκεί όπου δίνεται η ερμηνεία της λέξης καθώς και τα χαρακτηριστικά της στα Ελληνικά, υπάρχει το κλιτικό παράδειγμα, το γένος και η πτώση. Επισημαίνεται ότι η γραφή των Ελληνικών γίνεται με λατινικούς χαρακτήρες για τη διευκόλυνση του συστήματος. Προς χάριν του τελικού αποτελέσματος και για να μπορεί ο χρήστης να διαβάζει Ελληνικά, υπάρχει πίνακας μετατροπής των λατινικών αυτών χαρακτήρων σε Ελληνικούς, ο οποίος ενεργοποιείται κατά το στάδιο της εξαγωγής του μεταφρασθέντος κειμένου.

Παράδειγμα κωδικοποίησης ρήματος στο STEM

```
F  -A00DERANGE
F  AG31
F  BA 0404      001A -  1
F  CA ATRAN, ,PACOM,PRCOM,PRDPAS,HUOBJ
F  FAH00N3105   1   DJATARA'SSW
F  FAH01N1   1      DJATARAHI'
F  FAH02N2          DJATARAGME'NOS
```

Ερμηνεία κωδίκων στο παραπάνω παράδειγμα:
Οι κώδικες ανήκουν στις ίδιες κατηγορίες όπως και εκείνοι που εξηγήθηκαν στο προηγούμενο παράδειγμα. Αυτό που σημειώνεται στην κωδικοποίηση ρημάτων είναι ότι στο Ελληνικό μεταφραστικό αντίστοιχο συμπληρώνεται όχι μόνο η σημασία του ρήματος, αλλά και η σημασία του γερουνδίου που προκύπτει από αυτό, καθώς και της μετοχής του. Για τα μεταφραστικά αντίστοιχα στα Ελληνικά των ουσιαστικών και των επιθέτων η αναγνώριση του κλιτικού τους παραδείγματος γίνεται αυτόματα από το σύστημα, ενώ για τα ρήματα πρέπει να συμπληρωθεί ο κώδικας του κλιτικού τους παραδείγματος χειρωνακτικά.

Β. Το IDLS

Το IDLS, περιέχει καταχωρήσεις εκφράσεων που αποτελούνται από περισσότερες από μια λέξεις καθώς και κανόνες για τη μετάφραση κάποιων λέξεων μέσα σε ένα συγκεκριμένο περιβάλλον. Αυτό το λεξικό είναι ειδικό για κάθε γλωσσικό ζεύγος, δηλ σε κάθε κανόνα υπάρχει το μεταφραστικό αντίστοιχο για μία μόνο γλώσσα.

Για να κωδικοποιηθεί μια λέξη στο IDLS πρέπει να υπάρχει κατ' αρχήν στο STEM.

Επίσης, υπάρχουν και ειδικά ορολογικά λεξικά, τα οποία είναι υποδιαιρέσεις των δύο βασικών λεξικών που δηλώνονται με έναν κώδικα, ανάλογα με το είδος της ορολογίας.

Όλες οι πληροφορίες που χρειάζεται το σύστημα για την συντακτική και τη γραμματική αναγνώριση βρίσκονται στα λεξικά του υπό μορφή κωδίκων.

Κάθε μήνα γίνεται η νέα έκδοση των λεξικών του EC-SYSTRAN, όπου ενσωματώνονται όλες οι βελτιώσεις που έχουν κάνει οι γλωσσολόγοι που εργάζονται για την ανάπτυξη του συστήματος κατά το συγκεκριμένο χρονικό διάστημα.

Παράδειγμα κωδικοποίησης στο IDLS

```
F   -A41HOUSE $C-UC=PYR $C-B1,E,10 $C-ADNOM20 ENGINE
F   FA100N1   2   STACMO'S
F   FA101N2       PYROSVESTJKO'S
F   FA100SWD4
```

Εδώ η κωδικοποίηση γίνεται με μακροεντολές που είναι μια απλοποιημένη μορφή assembler ή με bytes, που περιγράφουν συντακτικές σχέσεις μεταξύ των λέξεων ή δίνουν τις γραμματικές/σημασιολογικές πληροφορίες. 'Ετσι όταν το συγκεκριμένο περιβάλλον αναγνωρίζεται από το σύστημα, δίνεται διαφορετική σημασία από αυτήν που υπάρχει στο STEM.

Η κωδικοποίηση στο IDLS χρησιμοποιεί κανόνες οι οποίοι μπορούν να επέμβουν σε συγκεκριμένα στάδια της ανάλυσης (βλ. σχήμα με στάδια μετάφρασης), ανάλογα με τον κώδικα προτεραιότητας που τους αποδίδεται.

Στο παραπάνω παράδειγμα έχει κωδικοποιηθεί το εξής περιβάλλον για τη λέξη *house*: στην περίπτωση που:

α) η λέξη *house* συναντηθεί σε κείμενο το οποίο έχει υποβληθεί για μετάφραση με τον κώδικα PYR (κώδικας που δίνεται για τη μετάφραση κειμένων της Πυροσβεστικής Υπηρεσίας στην Ελλάδα),

β) είναι ουσιαστικό, μέσα σε μια ονοματική φράση και

γ) προσδιορίζει το ουσιαστικό *engine*, τότε δίνεται η εντολή η λέξη *house* να μεταφραστεί ως *"σταθμός"*, η δε λέξη *engine* να πάρει επιθετική σημασία και να μεταφραστεί *"πυροσβεστικός"*.

Με τις μακροεντολές μπορούν να περιγραφούν και πιο πολύπλοκα περιβάλλοντα, αλλά δεν μπορούν να αντιμετωπισθούν όλες οι περιπτώσεις. Για γενικότερα φαινόμενα, όπως για ημερομηνίες, για απαρέμφατα κ.λπ. χρησιμοποιούνται ρουτίνες.

Επίλογος

Το EC-SYSTRAN είναι ένα από τα μακροβιότερα συστήματα Μηχανικής Μετάφρασης και στην πορεία του έχει αξιολογηθεί αρκετές φορές, αλλά και συνεχίζει να αξιολογείται κυρίως από την Ευρωπαϊκή Επιτροπή έτσι ώστε να εντοπίζονται τα προβλήματα και να προτείνονται βελτιώσεις για την καλύτερη απόδοση του συστήματος. Όλα τα γλωσσικά ζεύγη δεν έχουν τα ίδια περιθώρια βελτίωσης καθώς αυτό εξαρτάται από το χρόνο τον οποίο έχει αφιερωθεί στην ανάπτυξή τους (π.χ. τα ζεύγη Αγγλικά<->Γαλλικά αναπτύσσονται εδώ και 20 περίπου χρόνια και έχουν φτάσει σε πολύ υψηλό επίπεδο απόδοσης) και τη γραμματική/συντακτική συνάφεια των γλωσσών (π.χ. Ισπανικά -> Γαλλικά). Για την περίπτωση των Ελληνικών υπάρχει σαφώς περιθώριο βελτίωσης καθώς το ζεύγος είναι σχετικά νέο και δεν υπάρχει μεγάλη γραμματική/συντακτική συνάφεια μεταξύ Ελληνικών και Αγγλικών.

References

1. Piggot, I. M.
   "Systran development at the EC Commission: 1976 to 1992", 1992, Luxembourg

2. Hutchins W.J & Somers Harold
   An Introduction to Machine Translation.
   Academic Press 1992. ISBN 012 362830 x

3. *"Linguistic description of Systran",*
   Luxembourg, April 1993

4. Literature and Information Technology: a few comments

Assist. Professor Maria Tsoutsoura
Department of Modern Languages, Translation & Interpreting
Ionian Univerity
Megaro Kapodistria, GR-49100 Corfu HELLAS
tel.: & fax: +30 661 22033 E-mail: career@ionio.gr

Up to now, information technology in literature is applied in two fields: databases and publishing. Databases usually record works according to period, genre or subject and only rarely according to other literary aspects. However, data banks should not end up being mere museum collections of works nor lead to normative appreciation of reading choices, but they should depict the dynamic evolution of literary events in language, time and space and thus essentially contribute to research work.

Electronic publishing, on the other hand, allows the text to be enriched with annotation or illustration and offers the opportunity of parallel texts. It might contribute to genetic criticism through a virtual concept of texts and lead to a better perception of writing itself, usually thought of as pragmatic activity far from the idea of lasting texts.

Although literary criticism is enriched with new terms - sometimes as ambiguous and vague as "hypertext" - which have their origin in information technology, it is only starting to look for the appropriate reading interface, so that readers use electronic publications and the wide public discovers in them something more than a simple fashion to follow.

Λογοτεχνία και Ηλεκτρονικά Εργαλεία: πρώτες διαπιστώσεις

Επίκ. Καθηγήτρια Μαρία Τσούτσουρα
Τμήμα Ξένων Γλωσσών, Μετάφρασης και Διερμηνείας
Ιόνιο Πανεπιστήμιο
Μέγαρο Καποδίστρια, Κέρκυρα 49100

Από την δεκαετία του ’80, με την εμφάνιση του προσωπικού υπολογιστή, σημειώθηκε το ενδιαφέρον και των πρώτων φιλολόγων για την αξιοποίηση των ηλεκτρονικών εργαλείων σε εφαρμογές πέρα από την απλή επεξεργασία των κειμένων σε αντικατάσταση της γραφομηχανής. Σήμερα έχουν πια υλοποιηθεί αρκετές πιλοτικές δοκιμές στον τομέα αυτό, ώστε να είναι θεμιτή μια πρώτη αξιολόγησή τους, ως οδηγός για μελλοντικά σχέδια. Θα διατυπώσουμε λοιπόν συμπερασματικά κάποιες παρατηρήσεις σχετικά με τις συστηματικές εργασίες που έχουν γίνει μέχρι στιγμής στο πλαίσιο των ακαδημαϊκών και ερευνητικών ιδρυμάτων, ώστε να αναδειχθεί η γενικότερη συμπεριφορά, η δυναμική, αλλά και οι αδυναμίες των εφαρμογών ηλεκτρονικού περιβάλλοντος για την πλαισίωση και τη μελέτη της λογοτεχνίας.

Οι εφαρμογές των ηλεκτρονικών μέσων στη φιλολογία ακολούθησαν μέχρι σήμερα δύο κυρίως διαδρομές: τη σύσταση βάσεων δεδομένων και την εκδοτική των κειμένων.

Η σύσταση βάσεων δεδομένων αναφέρεται συνήθως σε βιβλιομετρική καταγραφή των έργων μιας ορισμένης περιόδου, ενός μορφολογικού τύπου κειμένων ή των έργων που περιστρέφονται γύρω από έναν θεματικό άξονα. Στις περιπτώσεις αυτές είναι συχνά δυσχερής η διατύπωση των κριτηρίων που προσδιορίζουν το σώμα των στοιχείων που πρέπει να καταγραφούν και να ευρετηριασθούν. Στα πρώτα στάδια βρίσκεται άλλωστε η περιγραφή και συγκρότηση των λειτουργικών εργαλείων πρόσβασης στη βάση από τον χρήστη.

Οι ενστάσεις που διατυπώνονται στις εργασίες αυτής της μορφής αναφέρονται σε δύο κυρίως σημεία:

- την αποτύπωση της ιστορίας της λογοτεχνίας ως μουσειακής συλλογής κειμένων και όχι ως εξέλιξης φαινομένων,
- τη δημιουργία ενός κανονιστικού μηχανισμού αξιολόγησης που μπορεί να οδηγήσει σε μια επικίνδυνη, περιοριστική πρόβλεψη της αναγνωσιμότητας.

Σπανιότερα, η σύσταση των βάσεων δεδομένων αναφέρθηκε σε άλλες μορφές ευρετηρίασης, όπως η καταγραφή λογοτεχνικών θεμάτων και θεματικών τόπων. Και στις περιπτώσεις όμως αυτές μοιάζει συχνά σαν μια πεπερασμένη, εικονοποιημένη απόδοση μιας ιδέας που προϋπήρχε με μια σειρά εργαλείων που δεν χαρακτηρίζονται για την δυναμική τους συγκρότηση, ούτε προβλέπουν ερευνητική, πέραν της ευρετικής αξιοποίησης των βάσεων.

Η εκδοτική των κειμένων σε ηλεκτρονική μορφή παρουσιάζει άλλωστε ιδιαίτερο ενδιαφέρον, για τον κύριο λόγο ότι καθιστά δυνατή την χειραφέτηση του κειμένου από την δυσδιάστατη μορφή, καθώς επιτρέπει τον ελεύθερο εμπλουτισμό του με κάθε λογής συμπληρώσεις, σημειώσεις, υπομνήματα και εικονογράφηση, αξιοποιεί τα χειρόγραφα και τις προκείμενες παραλλαγές και παρέχει την δυνατότητα συγκειμενικής παρουσίας παραλλήλων κειμένων, όπου πολλαπλασιάζονται οι διαδραστικές τους λειτουργίες.

Από αυτή την άποψη η ηλεκτρονική έκδοση, με την έννοια του ιδεατού κειμένου (virtual text), μπορεί να οδηγήσει σε νέες μεθόδους την γενετική κριτική, που αναζητά την πηγή και την διαδρομή της δημιουργίας των λογοτεχνικών έργων και προσβλέπει στον εντοπισμό της τελειότερης μορφής των υψηλά αξιολογημένων κειμένων και στον καλύτερο τρόπο της εκδοτικής τους παρουσίασης. Η ανάλυση του λόγου ως προϋπόθεση της ηλεκτρονικής έκδοσης μπορεί από αυτή την άποψη να αναδείξει την διαβάθμιση αξιολόγησης και λειτουργικότητας των κειμένων, που ανάγεται συνήθως χονδρικά στη διαφορά της χρηστικής γραφής (pragmatic writing) και των διαχρονικών έργων (lasting texts).

Οι εφαρμογές της ηλεκτρονικής τεχνολογίας στις φιλολογικές μελέτες έχουν ήδη προικίσει την κριτική με έναν ενδιαφέροντα νέο οπλισμό όρων που προέρχονται από την μεταφορική χρήση τους στα υπολογιστικά εργαλεία και παραμένουν συχνά ασαφείς, όπως το υπερκείμενο (hypertext). Ζητούμενο είναι να συμβάλλουν δημιουργικά στην έρευνα και να προτείνουν νέες διευρυμένες αναγνωστικές προτάσεις, ακόμα και για εκείνους που ήδη ξέρουν να διαβάζουν - όχι απλοϊκούς μηχανισμούς πρόσβασης στα λογοτεχνικά κείμενα για ένα κοινό που τις χρησιμοποιεί ως άλλοθι του συρμού για να μην διαβάσει ποτέ.

IV. Παρουσίαση Νέων Βιβλίων / *Presentation of New Books*

G. Babiniotis, Dictionary of Modern Greek Language, Athens 1998, Lexicology Center, pages 2064

Καθηγητής Χριστόφορος Χαραλαμπάκης
Πανεπιστήμιο Αθηνών

Abstract

In this article the Dictionary of Modern Greek Language by G. Babiniotis is presented. As it is mentioned in the preface of the "Lexicology Center", this Dictionary is indeed "the first extensive, complete and comprehensively compiled- in terms of linguistics specifications- Dictionary of Modern Greek. The three grate novelties applied in the Dictionary, according to the editors of the project, are: a) The comments on the correct use of words b) the structure and the blueprint appearance of the Dictionary and c) the etymology of all the words of Modern Greek language. However, the fact that the vernacular Modern Greek language has been described to such an extend for the first time and both its liveliness as well as wealth of expression illustrated must be regarded as the primary virtue of this work.

Γ. Μπαμπινιώτη, Λεξικό της νέας Ελληνικής γλώσσας, Αθήνα 1998: Κέντρο Λεξικολογίας, σελίδες 2064.G.

Όπως αναφέρεται στον Πρόλογο του "Κέντρου Λεξικολογίας", το Λεξικό αυτό είναι πράγματι "το πρώτο εκτενές, πλήρες και με γλωσσολογικές προδιαγραφές συντεταγμένο λεξικό της Νέας Ελληνικής". Οι τρεις μεγάλες καινοτομίες του Λεξικού, κατά τους εκδότες του έργου, είναι: α). *Τα Σχόλια για τη σωστή χρήση των λέξεων*, β). *Η δομή και η τυπογραφική εμφάνιση του Λεξικού* και γ). *Η ετυμολογία όλων των λέξεων της σύγχρονης γλώσσας*. Πρώτιστη όμως αρετή πρέπει να θεωρηθεί το γεγονός ότι για πρώτη φορά περιγράφεται σε τόση έκταση η κοινή νέα ελληνική, και αναδεικνύεται η ζωντάνια και ο εκφραστικός της πλούτος.

Είναι προς τιμήν του Γιώργου Μπαμπινιώτη ότι *δεν ρυθμίζει αλλά περιγράφει τη γλώσσα*, όπως χρησιμοποιείται στις ποικίλες εκφάνσεις της, στο γραπτό και προφορικό λόγο, στην καθημερινή ζωή, στον τύπο, τη λογοτεχνία και σε επιστημονικά κείμενα. Βασικό κριτήριο της λημματογράφησης υπήρξε η χρήση. Έτσι δεν διστάζει να συμπεριλάβει στο Λεξικό του όλες τις ξένες λέξεις που χρησιμοποιούνται ευρέως από τους νεοέλληνες, έστω και αν ενοχληθούν ορισμένοι που επιδιώκουν την "καθαρότητα" της γλώσσας. Στο Λεξικό μπορεί κανείς χιλιάδες ξενισμούς, από το *κολγκερλ* (το οποίο σεμνότυφα δεν υπάρχει σε κανένα άλλο νεοελληνικό λεξικό) και το *μόντεμ* ως το επίθετο *σέξι*, το *ντόπινγκ*, το *φάιναλ φορ* και το *χάμπουργκερ*. Με την ίδια νηφαλιότητα αντιμετωπίζει τη γλώσσα των νέων (*άπαικτος* στη σημασία ασυναγώνιστος, *κυριλέ, φλιπάρω, κολλητός*), τη αργκό: *το βλέπω χλομό να πάρουμε αύξηση* = "δύσκολο, απίθανο"), τη λαϊκή και τις λεγόμενες "χυδαίες" λέξεις, πολλές από τις οποίες με τη συχνή χρήση τους έχασαν την αρχική σημασία τους. Έμφαση δίνεται και στις λόγιες λέξεις οι οποίες παρουσιάζονται σε απαιτητικά κείμενα και αποτελούν ιδιαίτερο στοιχείο ύφους: *αμελλητί, βότρυς, ηώς* (= η αυγή), *συνδαιτυμόνας*. Καταγράφονται αρχαιοπρεπείς λέξεις (λ.χ. *πεφυσιωμένος*) και πολλές λόγιες φράσεις που βρίσκονται σε ευρύτατη χρήση: *ο κύβος ερρίφθη, τα προς το ζην, γηράσκω αεί διδασκόμενος*.

Το Λεξικό είναι επίκαιρο, ζωντανό και ενημερωμένο ως την τελευταία στιγμή της εκτύπωσής του, καθώς περιέχει λέξεις που βρίσκονται στο επίκεντρο της πολιτικής, κοινωνικής, πνευματικής, οικονομικής αλλά και της καθημερινής μας ζωής, όπως: *βουλή των εφήβων, διασωλήνωση,*

κλωνοποίηση, μετροπόντικας, Μηχανισμός Συναλλαγματικών Ισοτιμιών (Μ.Σ.Ι.) *ξενοφοβία, οικονομικός πρόσφυγας, ρατσισμός, τρανσέξουαλ* κ.ά. Οι ετυμολογίες είναι πειστικές και υπεύθυνες. Πολλές φορές έχουμε μέσα σε πλαίσιο υπέροχες μικρές πραγματείες,. Δεν επαναλαμβάνονται φυσικά οι αφελείς ετυμολογίες που έχουν όλα τα σύγχρονα λεξικά, όπως λ.χ. ότι ο *βατραχάνθρωπος* ετυμολογείται από το *βάτραχος + άνθρωπος*, ενώ πρόκειται για μεταφραστικό δάνειο. (Πβ. αγγλ. frogman, γερμ. Froschmann, γαλλ. homme - grenouille, ιταλ. uomo rana, ισπαν. hombre rana).

Τα πλούσια Σχόλια του Λεξικού (που υπάρχουν μέσα σε ειδικό πλαίσιο) ως προς τη σωστή χρήση των λέξεων, την ορθογραφία, την ετυμολογία, τις σημασιολογικές τους διαφοροποιήσεις, κ.ά. αποτελούν βασική αρετή του έργου και ένα από τα σημαντικά στοιχεία που το κάνουν να ξεχωρίζει από όλα τα άλλα λεξικά. Τα ερμηνεύματα είναι άψογα, χωρίς περιττολογίες και πλατειασμούς. Η ορθογράφηση των ξένων λέξεων κινείται πολύ σωστά προς απλουστευτική κατεύθυνση που θα ξαφνιάσει ορισμένους (π.χ. *ερκοντίσιον, ερμπάς, οτοστοπ*). Η καταγραφή συνωνύμων και αντιθέτων είναι προσεκτικά μελετημένη με εντυπωσιακά σχόλια που δείχνουν οξύ γλωσσικό αισθητήριο. Εντυπωσιάζει η πληθώρα των κυρίων ονομάτων και τοπωνυμίων, η προσεκτική επιλογή χιλιάδων επιστημονικών όρων και η καταγραφή εκατοντάδων ακρωνυμίων που έχουν μπει με ταχύτατους ρυθμούς στη ζωή μας.

Υποδειγματική είναι η κατάταξη των σημασιών οι οποίες χαρακτηρίζονται με βάση το ύφος, δηλ. το *πώς*, (αρχαιοπρεπές, λόγιο, καθημερινό, οικείο, κ.ά.), το *που*, το είδος της γλωσσικής επικοινωνίας και της χρήσης (λογοτεχνία, λαϊκή γλώσσα, διάλεκτοι, αργκό) και το *γιατί*, το είδος της χρήσης: ειρωνική, σκωπτική, υβριστική, καταχρηστική, εκφραστική κ.ά. Σε αρκετά λήμματα μπαίνει τάξη στο χάος της πολυσημίας, όπως λ.χ. στο λήμμα *κόβω*, στο οποίο καταγράφονται 52 σημασίες και δεκάδες φράσεις. Για το ρήμα *κάνω* εντοπίζονται 34 σημασίες και για το *βάζω* 30 σημασίες και 24 φράσεις.

Για τη δημιουργική χρήση της γλώσσας είναι πολύτιμη η καταγραφή των συνδυαστικών δυνατοτήτων των λέξεων (σύμπλοκα: collocations). Βλ. λ.χ. τη λέξη θάλασσα: *ανοιχτή/μανιασμένη/ακύμαντη/αγριεμένη/ζεστή/γαλανή/γαλήνια/σκοτεινή/αφιλόξενη*. Η θάλασσα *φουρτουνιάζει, σκάβει τα βράχια, φουσκώνει, βουίζει, λυσσομανάει, αγριεύει, γαληνεύει* κ.ά.

Τα πολυλεκτικά σύνθετα λημματοποιούνται για πρώτη φορά σε νεοελληνικό λεξικό (λ.χ. *κινούμενα σχέδια*), σε περιορισμένη όμως, κλίμακα. Σε πολλά λεξικά έμειναν εντελώς απαρατήρητα επί σειρά ετών. Έτσι καταγράφονται εδώ για πρώτη φορά φραστικά ονόματα του τύπου: *φαινόμενο του θερμοκηπίου, ολυμπιακή φλόγα* κ.ά. Η προβολή των φράσεων με ιδιαίτερα τυπογραφικά στοιχεία και η ένταξή τους μέσα στις σημασίες του λήμματος, είναι κάτι το καινοφανές σε λεξικό. Διευκολύνει αφάνταστα τον χρήστη να εντοπίσει τις παγιωμένες εκφράσεις του τύπου: *ζώνη ασφαλείας, φακοί επαφής, δημόσια διοίκηση, πιάνω το Μάη, κάτι τρέχει στα γύφτικα* κ.ά.

Χωρίς μεγάλα λόγια, αλλά με έργα, ο Γ. Μπαμπινιώτης, με ένα επιτελείο έμπειρων λεξικογράφων, μας έδειξε με μαεστρία και τέχνη τη μαγεία και τη γοητεία, τον δυναμισμό και τη ζωντάνια της μικρής σε ομιλητές αλλά μοναδικής σε ιστορικό βάθος και ανεπανάληπτης σε πλούτο και εκφραστική πληρότητα Ελληνικής γλώσσας.

Σύγχρονο Ελληνοδανικό Λεξικό
Greek - Danish Dictionary

Μαρία Γαβριηλίδου
Ινστιτούτο Επεξεργασίας του Λόγου

Abstract
The presentation to the public of the Greek -

Danish Dictionary which has just been published by Patakis Publications took place on Wednesday 1 April 1997 at the Danish Institute of Athens. This dictionary project was funded by the LINGUA / SOCRATES program, the Cultural and Education Ministries of the two countries and the Eleni Nakou Foundation. It was a Greek - Danish joint effort (partners were the Center for Business Studies of the Aarhus Business School (CEF), ILSP, and the Danish Institute of Athens); chief editor was Rolf Hesse, who was assisted (besides the Institutes) by a team of translators and domain specialists. The dictionary contains 28,000 headwords and covers the currently written and spoken language. It was based on electronic corpora, both for the selection of the words to be included (using the measure of frequency) and for the selection of examples of actual use. It offers the user information about the part of speech, translational equivalent, examples of usage, idioms, collocations, domain and register. The publication of the Greek - Danish Dictionary is highly important, given that it is the first dictionary of this size and coverage between the two languages.

Την Τετάρτη 1 Απριλίου 1998 έγινε στο Ινστιτούτο της Δανίας στην Αθήνα η παρουσίαση του μεγάλου **Σύγχρονου Ελληνοδανικού Λεξικού** που μόλις εκδόθηκε από τις εκδόσεις Πατάκη. Το λεξικό αυτό είναι αποτέλεσμα ενός διακρατικού προγράμματος που πραγματοποιήθηκε με την οικονομική υποστήριξη του Προγράμματος LINGUA / SOCRATES της Ευρωπαϊκής Ένωσης, του Ιδρύματος "Ελένη Νάκου", του Ελληνικού Υπουργείου Πολιτισμού, των Υπουργείων Παιδείας της Δανίας και της Ελλάδας, καθώς και του Ερευνητικού Συμβουλίου Ανθρωπιστικών Σπουδών της Δανίας.

Συντονιστής του προγράμματος ήταν το Κέντρο Επαγγελματικών Ερευνών (Center for Erhvervsforskning, CEF) της Ανωτάτης Εμπορικής Σχολής του Aarhus. Αρχισυντάκτης του λεξικού ήταν ο κ. Rolf Hesse, ορκωτός μεταφραστής και καθηγητής φιλολογίας, και συμμετείχαν το Ινστιτούτο Επεξεργασίας Λόγου και το Ινστιτούτο της Δανίας στην Αθήνα. Επίσης στο πρόγραμμα συμμετείχαν Δανοί και Έλληνες γλωσσολόγοι, μεταφραστές καθώς και ειδικοί επιστήμονες ως σύμβουλοι ορολογίας.

Με την έκδοσή του ολοκληρώθηκε ένα εξαετές λεξικογραφικό πρόγραμμα, τα πρώτα τρία χρόνια του οποίου είχαν οδηγήσει στην έκδοση του Δανοελληνικού λεξικού.

Τι περιλαμβάνει το λεξικό

Το Σύγχρονο Ελληνοδανικό Λεξικό περιλαμβάνει 28.000 λήμματα και απευθύνεται τόσο σε Δανούς όσο και σε Έλληνες. Το γεγονός αυτό οδήγησε στην οργάνωση του συμπληρωματικού υλικού (πίνακες, επεξηγήσεις, σχόλια) και στις δύο γλώσσες.

Έτσι, το λεξικό έχει δύο εισαγωγές (Ελληνική και Δανική) συμμετρικά δομημένες: η Ελληνική Εισαγωγή περιλαμβάνει μία πολύ σύντομη παρουσίαση της Δανικής γλώσσας και ιστορίας, αναλυτική παρουσίαση του Δανικού αλφαβήτου με οδηγίες προφοράς και παραδείγματα και τέλος μικρή γραμματική της Δανικής γλώσσας (σε επίπεδο κλιτικής μορφολογίας). Αντίστοιχα, η Δανική Εισαγωγή περιλαμβάνει (όσο μπορούμε να καταλάβουμε!) σύντομη παρουσίαση της Ελληνικής γλώσσας, παρουσίαση του Ελληνικού αλφαβήτου με οδηγίες προφοράς και παραδείγματα και μικρή γραμματική της Ελληνικής γλώσσας (σε επίπεδο κλιτικής μορφολογίας). Ακολουθεί πίνακας συντομογραφιών και συμβόλων με επεξήγηση και στις δύο γλώσσες και σύντομες οδηγίες χρήσης του λεξικού.

Με την ίδια λογική (δηλαδή της παρουσίασης και στις δύο γλώσσες) είναι καταρτισμένοι οι Πίνακες στο τέλος του λεξικού. Το λεξικό περιλαμβάνει:

- πίνακες Δανικών και Ελληνικών ομαλών και ανωμάλων ρημάτων που συνοδεύονται από τις ιδιαιτερότητες της κάθε γλώσσας στην

κλίση (π.χ. εσωτερική αύξηση, ανώμαλος σχηματισμός χρόνων στα Ελληνικά, κτλ.),

- πίνακες αριθμητικών,
- ονόματα μηνών και ημερών, και
- πίνακες γεωγραφικών ονομάτων (Κρατών και Πόλεων της Ευρώπης).

Μακροδομή του λεξικού

Για τη μακροδομή του λεξικού (λημματολόγιο) χρησιμοποιήθηκαν ως πηγές το λημματολόγιο του Δανοελληνικού λεξικού (αντεστραμμένο), και πίνακας συχνοτήτων της Ελληνικής γλώσσας που προέκυψε από επεξεργασία ηλεκτρονικού σώματος κειμένων του ΙΕΛ (λημματοποίηση και στατιστική επεξεργασία). Δεδομένου ότι ο στόχος της κατάρτισης του λημματολογίου ήταν να περιλάβει το σύγχρονο Ελληνικό λεξιλόγιο, το Σώμα Κειμένων περιελάμβανε *σύγχρονα* κείμενα, τα οποία προέρχονταν από τους τομείς του Τύπου (εφημερίδες ΕΛΕΥΘΕΡΟΤΥΠΙΑ και ΒΗΜΑ, φύλλα των ετών 1995-96), της Λογοτεχνίας και της Επιστήμης (με χρονολογία συγγραφής μεταγενέστερη του 1992).

Το λεξικό ως συστατικά του σύγχρονου λεξιλογίου καταγράφει και

- λέξεις λόγιας προέλευσης που είναι σε χρήση στο σύγχρονο Ελληνικό λόγο: *απαρτίζω, όθεν, κλινήρης, αντιολισθητικός*
- νεολογισμούς (λέξεις και εκφράσεις): *πουρό, ξενερώνω, τη βρίσκω, καμικάζι*
- λέξεις ξένης προέλευσης που χρησιμοποιούνται ευρύτατα στο σύγχρονο Ελληνικό λόγο: *βινύλιο, βίντεο, φεϊγβολάν, προσπέκτους, σικέ*
- σύγχρονους επιστημονικούς / τεχνικούς όρους με ευρεία χρήση στη γενική γλώσσα: *αφθώδης, νωτιαίος, ενάγων, υβρίδιο*. Πρέπει να σημειωθεί ότι, δεδομένου ότι το λεξικό είναι γενικής γλώσσας, οι όροι που περιλαμβάνονται είναι προσεκτικά επιλεγμένοι.

Μικροδομή του λεξικού

Η μικροδομή του λεξικού (πληροφορία που συνοδεύει το λήμμα) περιλαμβάνει

- το μέρος του λόγου
- το γένος των ουσιαστικών
- λεπτομερή χωρισμό σημασιών (πολλές φορές ο χωρισμός υποβοηθείται από συνώνυμα, π.χ. αγαθός (*καλός*) god, (*αφελής*) naiv
- τη μεταφραστική αντιστοιχία στη Δανική
- χρηστικά παραδείγματα
- ιδιωματικές εκφράσεις/παροιμίες/collocations (μάτια που δεν **βλέπονται** γρήγορα λησμονιούνται, **ζωή** και κότα, **ξεράθηκα** στα γέλια, μου ξηγήθηκε **σπαθί**)
- πληροφορία για τον τομέα στον οποίο ανήκει ένα λήμμα (*ανατομία, γεωλογία, νομική*)
- πληροφορία για το επίπεδο λόγου (*λόγια, καθομιλουμένη, χυδαία*)
- σε περιπτώσεις λέξεων πολιτισμικά καθορισμένων και άρα προβληματικών στην αντιστοιχία τους στην Δανική, το λεξικό χρησιμοποιεί σχέδια και περιγραφικό ορισμό (*το τσαρούχι* ορίζεται ως *traditionel bondesko* και συνοδεύεται από το σχετικό σχέδιο).

Αποτιμώντας το συνολικά: πρόκειται για ένα πολύ προσεγμένο λεξικό με συνεπείς αρχές, το οποίο σέβεται τις ιδιαιτερότητες της σύγχρονης Ελληνικής γλώσσας, καλύπτει ικανοποιητικό μέρος του εύρους του λεξιλογίου της σε σχέση με το μέγεθός του και με τους στόχους του, και αποτελεί όντως **σύγχρονο** Ελληνοδανικό λεξικό.

V. Γλωσσάριο Όρων Γλωσσικής Τεχνολογίας και Πληροφορικής / *Language Technology and Informatics Terminology Forum*

Το γλωσσάριο όρων Γλωσσικής Τεχνολογίας και Πληροφορικής εμπλουτίζεται συνεχώς με νέους όρους. Στο παρόν τεύχος εκτός από τις προτάσεις για νέους όρους γίνεται και σχολιασμός για όρους που έχουν δημοσιευθεί σε προηγούμενα τεύχη. Ο σχολιασμός και η περαιτέρω συζήτηση για την καταλληλότητα των όρων πιστεύουμε ότι θα συμβάλει στη βελτίωση της απόδοσης των όρων από την Αγγλική στην Ελληνική. Σημειώνεται ότι οι ορθές αποδόσεις όρων θα πρέπει να βασίζονται σε σημασιολογικές, πραγματολογικές και εξωγλωσσικές γνώσεις, καθώς και στη γνώση του μικρόκοσμου στον οποίο χρησιμοποιείται ο κάθε όρος.

1. Προτεινόμενοι όροι

Τους παρακάτω όρους προτείνει ο Καθηγητής Γεώργιος Κουρουπέτρογλου.

aliasing	**αλλοίωση**	το φαινόμενο της (πιθανής) φασματικής αλλοίωσης κατά την ψηφιοποίηση αναλογικών σημάτων
alternative communication	**εναλλακτική επικοινωνία**	μέθοδοι επικοινωνίας που χρησιμοποιούνται από πρόσωπα χωρίς καμιά φωνητική δυνατότητα
anti-aliasing	**αντι-αλλοίωση**	τεχνική (συνήθως φίλτρο) που εμποδίζει το φαινόμενο της (πιθανής) φασματικής αλλοίωσης κατά την ψηφιοποίηση αναλογικών σημάτων
augmentative communication	**επαυξητική επικοινωνία**	η χρήση βοηθημάτων ή τεχνικών που ενισχύουν ή συμπληρώνουν τις υπάρχουσες φωνητικές ή προφορικές δεξιότητες
average word branching factor	**μέσος παράγοντας διακλάδωσης λέξης**	μέσος παράγοντας διακλάδωσης λέξης κατά την αναγνώριση λέξεων όταν η αναγνώριση βασίζεται σε ένα μοντέλο γλώσσας, ή η ποσότητα που εκφράζει τη μέση δυσκολία ή αβεβαιότητα αναγνώρισης λέξεων [**Βλ. και και perplexity**]
bank-of-filters	**Βλ. filter bank**	
code word	**κωδικολέξη**	παράμετρος εισόδου μιας διαδικασίας ανυσματικής κβάντωσης που αντιπροσωπεύει ένα σύνολο από ανύσματα (π.χ. ένα σύνολο φασματικών συνιστωσών)
codebook	**κωδικολεξικό**	λεξικό κωδικολέξεων
concatenation	**συρραφή**	π.χ. συρραφή προεκφωνημένων μονάδων ομιλίας
contour spectrogram	**φασματογράφημα ισοϋψών**	
difference limen	**Βλ. just noticeable difference**	
filter bank	**τράπεζα φίλτρων**	σύνολο ζωνοπερατών φίλτρων που καλύπτουν μια ζώνη συχνοτήτων [αναφέρεται και ως **bank-of-filters**]
just noticeable difference	**μόλις διακρίσιμη διαφορά**	η διαφορά ακουστικής στάθμης ή η διαφορά συχνότητας που μπορεί να αναγνωριστεί από έναν ακροατή με αβεβαιοτητα 50%
masking	**απόκρυψη**	η κατάσταση κατά την οποία η παρουσία ενός ήχου καθιστά αδύνατο το άκουσμα ενός άλλου ήχου
perplexity	**περιπλοκή**	ποσότητα που εκφράζει τη μέση δυσκολία ή αβεβαιότητα αναγνώρισης λέξεων, όταν η αναγνώριση βασίζεται σε ένα μοντέλο γλώσσας. Ονομάζεται και μέσος παράγοντας διακλάδωσης λέξης του μοντέλου γλώσσας [**Βλ. και average word branching factor**]
phonation	**φώνηση**	αναφέρεται κύρια στην ταλάντωση των φωνητικών χορδών, αλλά μπορεί να περιλάβει όλους τους τρόπους με τους οποίους λειτουργεί ο φάρυγγας σαν πηγή ήχων [**Βλ. και voicing**]
phone	**(φωνητικός) φθόγγος**	μια συγκεκριμένη εκφώνηση ή μια συγκεκριμένη περίπτωση ενός φωνήματος
pitch	**μουσικός τόνος, ύψος ήχου**	πως ένας ακροατής αντιλαμβάνεται αν ένας ήχος είναι χαμηλός (βαθύς) ή υψηλός (οξύς) με βάση μια υποκειμενική κλίμακα ύψους, χωρίς δηλαδή να λαμβάνει υπόψη του τις φυσικές ιδιότητες του ήχου. Ακουστικά συσχετίζεται με τη **βασική συχνότητα**
quasi-sinusodial	**ψευδοημιτονοειδές**	οιονεί ημιτονοειδές, σαν ημιτονοειδές, κατά κάποιον τρόπο ημιτονοειδές

reference pattern	**Βλ. template**	
speaker adaptation	**προσαρμογή ομιλητή**	τεχνικές που προσπαθούν να τροποποιήσουν ένα σύνολο υποδειγμάτων σε ένα σύστημα αναγνώρισης ομιλίας ανεξάρτητης του ομιλητή, χρησιμοποιώντας νέα δεδομένα εκμάθησης από ένα συγκεκριμένο ομιλητή
spectrogram	**φασματογράφημα**	
spectrograph	**φασματογράφος**	ειδική συσκευή παραγωγής φασματογραφημάτων
speech aid	**βοήθημα ομιλίας**	επαυξητικά συστήματα που ενισχύουν ή συμπληρώνουν τις υπάρχουσες δεξιότητες ομιλίας ή που χρησιμοποιούνται εναλλακτικά για επικοινωνία από πρόσωπα χωρίς καμιά δυνατότητα ομιλίας
speech enchancement	**βελτίωση ομιλίας**	επεξεργασία του σήματος ομιλίας πριν την ακρόασή του, με σκοπό τη βελτίωση ενός ή περισσοτέρων παραγόντων αντίληψης ομιλίας, όπως της συνολικής ποιότητας, της κατανοητότητας, του βαθμού κόπωσης ακροατή, κλπ
speech restoration	**αποκατάσταση ομιλίας**	επεξεργασία του σήματος ομιλίας που έχει υποστεί υποβιβασμό της ποιότητάς του, με σκοπό να καταστεί όσο το δυνατόν ίδιο με το αρχικό
speech visualization	**οπτικοποίηση ομιλίας**	μέθοδοι και τεχνικές για την οπτική αναπαράσταση του σήματος ομιλίας (περιλαμβάνουν από το κοινό φασματογράφημα μέχρι τις χρονοσυχνοτικές κατανομές και τα μοντέλα ακρόασης)
talking faces	**ομιλούντα πρόσωπα**	συνθετικές, μη στατικές αναπαρατάσεις του προσώπου ενός ομιλητή στην οθόνη του υπολογιστή, οι οποίες συγχρονίζουν τις κινήσεις του κατά την παραγωγή συνθετικής ομιλίας
template	**υπόδειγμα**	αντιπροσωπευτικό πρότυπο αναφοράς που χαρακτηρίζει τις παραλλαγές των εκφωνήσεων ήχων ομιλίας της ίδιας κλάσης ή ήχων με τα ίδια χαρακτηριστικά [**αναφέρεται και ως reference pattern**]
template adaptation	**προσαρμογή υποδείγματος**	τεχνικές που προσαρμόζουν τις τιμές των παραμέτρων του υποδείγματος σε αλλαγές του σήματος εισόδου (π.χ. νέο περιβάλλον ομιλίας, νέος ομιλητής), έτσι ώστε η αναγνώριση ομιλίας να μπορεί να αντιμετωπίζει σήματα ομιλίας που είναι κάπως διαφορετικά από εκείνα που χρησιμοποιήθηκαν κατά την εκμάθηση του υποδείγματος
template matching	**ταίριασμα υποδείγματος**	διαδικασία μιας μεθόδου αναγνώρισης ομιλίας, κατά την οποία οι παράμετροι μιας άγνωστης εκφώνησης συγκρίνονται με τις αντίστοιχες παραμέτρους ενός συνόλου υποδειγμάτων που δημιουργήθηκαν μετά από εκμάθηση, για να διαπιστωθεί άν ταιριάζει η άγνωστη εκφώνηση με ένα μέλος του συνόλου των υποδειγμάτων
template training	**εκμάθηση υποδείγματος**	τεχνικές που επιτρέπουν τη δημιουργία ενός συνόλου υποδειγμάτων σε ένα σύστημα αναγνώρισης ομιλίας που βασίζεται στη μέθοδο "ταίριασμα υποδείγματος"
text normalization	**κανονικοποίηση κειμένου**	επεξεργασία κειμένου, πριν τη χρησιμοποίησή του ως εισόδου σε ένα συνθέτη ομιλίας [**Βλ. και text preprossening**]
text preprossening	**προεπεξεργασία κειμένου**	επεξεργασία κειμένου, πριν τη χρησιμοποίησή του ως εισόδου σε ένα συνθέτη ομιλίας [**Βλ. και text normalization**]
time-frequency	**χρονοσυχνοτικός**	
voicing	**φώνηση**	η διαδικασία παραγωγής των ηχηρών και των άηχων ήχων [**Βλ. και phonation**]

Καθηγητής Γεώργιος Κουρουπέτρογλου
Τμήμα Πληροφορικής
Τομέας Επικοινωνιών και Επεξεργασίας Σήματος
Πανεπιστήμιο Αθηνών
Πανεπιστημιούπολη, Ιλίσια, 15781 Αθήνα

Professor Georgios Kouroupetroglou
Department of Informatics
Division of Communication and Signal Processing
University of Athens
Panepistimioupolis, Ilisia, GR-15781 Athens, Greece
E-mail: koupe@di.uoa.gr • http://www.di.uoa.gr/

2. Σχόλια σε όρους που προτάθηκαν σε προηγούμενα τεύχη

Από το Γραφείο Ορολογίας του Παιδαγωγικού Ινστιτούτου μας εστάλησαν προτάσεις με: Α) σχολιασμό προτάσεων για όρους Πληροφορικής και Γλωσσικής Τεχνολογίας των Καθηγητών Γ. Καραγιάννη, Ι. Κόντου και Γ. Παπακωνσταντίνου, στο περιοδικό "Λογοπλοήγηση", τεύχος 3, Δεκέμβριος '97, σελ. 30-31 και Β) διάφοροι όροι σχετι-

κά με τους οποίους αναμένονται προτάσεις και απόψεις. Τα παραπάνω έχουν δημοσιευθεί στο Εβδομαδιαίο Δελτίο Ορολογίας- Γλώσσας του Παιδαγωγικού Ινστιτούτου 18/3/98 και 1/4/98. (Δρ. Σ.Η. Διάμεσης, Σύμβουλος Π.Ι.). Τους όρους επιμελήθηκαν οι κ.κ. Β. Παπανδρέου και Μ. Επιφανίδης.

Α)

- **model** υπόδειγμα, πρότυπο, μοντέλο
- **pattern** ίχνος, αχνάρι, πρότυπο
- **standard** πρότυπο

Ερωτήματα:

Ποιά νομίζετε ότι είναι η πιο σωστή μετάφραση των παραπάνω αγγλικών λέξεων;

Πώς μπορεί να αποδοθεί πιό σωστά η λέξη **πρότυπο** στα αγγλικά και πώς η λέξη **υπόδειγμα**;

- **unification grammar** ενοποιητική γραμματική, γραμματική ενοποίησης

Ποιά μετάφραση νομίζετε ότι αποδίδει ορθότερα την έννοια του όρου;

Σύμφωνα με την άποψη των γραφόντων είναι ορθότερος ο όρος Ενοποιητική Γραμματική (βλ. Παράδειγμα Λατινική Γραμματική ή Παθητική Φωνή κ.λπ.)

- **declarative knowledge**
  διακηρυκτική γνώση, δηλωτική γνώση (declaration: διακήρυξη, statement: δήλωση)
- **interaction**
  διάδραση, διεπίδραση, αλληλεπίδραση, διαντίδραση
- **aligned parallel corpora**
  στοιχισμένα παράλληλα σώματα κειμένων ή ευθυγραμμισμένα παράλληλα σώματα;
- **mapping**
  απεικόνιση ή χαρτογράφηση, σχεδίαση;
- **matching**
  ταίριασμα ή συνταίριασμα, προσαρμογή, αντιπαραβολή;
- **recursion**
  αναδρομή ή αναπόληση, επιστροφή;
- **supervised learning**
  εκμάθηση με επίβλεψη ή επιβλεπτική ή επιβλεψιακή εκμάθηση, επιστασιακή μάθηση;

Β)

- **Intercity (train**)
  διαστικό (τρένο), ιντερσίτυ
- **design**
  τεχνοσύνθεση "σύνθεση", ντιζάϊν
- **functionalism**
  λειτουργισμός, "φονκσιοναλισμός"
- **Ενιαίο Λύκειο**
  Unified, Joint, Comprehensive (?) Lyceum
- **κυκλοφοριακός**
  traffic-(π.χ. traffic control- έλεγχος κυκλοφορίας)
- **κυκλοφορικός**
  circulatory-(π.χ. circulatory pressure- κυκλοφορική πίεση)
- **ηλεκτρονιακός**
  ο αναφερόμενος στο/α ηλεκτρόνιο/α (electron movement)
- **ηλεκτρονικός**
  ο αναφερόμενος στην Ηλεκτρονική (Electronic circuits)

3. Απόψεις

Η επιστολή που ακολουθεί είναι του Καθηγητή Γ. Καραγιάννη και απευθύνεται στον Δρα Σ. Διάμεση και στον Καθηγητή Γ. Κουρουπέτρογλου.

Αγαπητέ κ. Διάμεση, Αγαπητέ κ. Κουρουπέτρογλου,

Θέλω να σας ευχαριστήσω θερμά για την συνεισφορά σας στο “forum” όρων της γλωσσικής τεχνολογίας και τον προβληματισμό που αναπτύξατε.

Σχετικά με τους προταθέντες όρους από τον καθηγητή κ. Κουρουπέτρογλου καθώς και σχετικά με τους προταθέντες σε προηγούμενο τεύχος όρους που σχολιάστηκαν από το Παιδαγωγικό Ινστιτούτο ήθελα να σημειώσω τις εξής παρατηρήσεις:

aliasing: Ο όρος "αναδίπλωση" ο οποίος χρησιμοποιείται πολλά χρόνια τώρα (στο ΕΜΠ από το

1984) φαίνεται να είναι πιο εκφραστικός γιατί αποδίδει το φαινόμενο αυτό καθ' εαυτό, δηλαδή αν δεν έχει σεβασθεί κανείς το θεώρημα του Shannon κατά την δειγματοληψία έχει μία φασματική συμπεριφορά σαν να αναδιπλούται το φάσμα και οι υψηλές συχνότητες να εμφανίζονται συμμετρικά σαν χαμηλές. Φυσικά το φάσμα αλλοιούται, αλλά μήπως ο όρος "αναδίπλωση" βοηθάει περισσότερο τον σπουδαστή στην κατανόηση της έννοιας και είναι περισσότερο παραστατικός.

anti-aliasing (filter): "Αντι-αναδιπλωτικό (φίλτρο)". Φίλτρο κατά την διαδικασία της δειγματοληψίας χρήσιμο για να ορισθεί (περιορισθεί) το εύρος ζώνης ενός σήματος ώστε να μην υπάρξει αναδίπλωση του φάσματος. Η λέξη "αντι-αναδιπλωτικό" είναι καλύτερη από την "αντι-αλλοιωτικό" και ο όρος σαφέστερος.

codebook: Ίσως "κωδικοβιβλίο" είναι επίσης καλό. Υπάρχει μικρή ειδοποιός διαφορά. Πρόκειται για μία συλλογή από κώδικες (βιβλίο) που χρησιμοποιείται σε διαδικασίες διέγερσης συστήματος και συσταδοποίησης (clustering). Το αποτέλεσμα της χρήσης του "κωδικοβιβλίου" δεν είναι σαν την χρήση της γλώσσας για να παραπέμπει σε έννοια σαν του λεξικού (παράθεση λέξεων) αλλά κάτι πιο πολύπλοκο.

Vector quantization: Διανυσματική κβαντοποίηση ή κβαντισμός αντί ανυσματική κβάντωση. Η λέξη "άνυσμα" χρησιμοποιείται περισσότερο στα μαθηματικά και λιγότερο στην τεχνολογία.

pitch: Για τη φωνή αξίζει να σημειωθεί ότι "pitch" είναι η θεμελιώδης περίοδος. Όπως το σήμα της φωνής είναι ψευδοπεριοδικό, η μεταβολή του pitch αντιστοιχεί στην λεγόμενη "μικρομελωδία".

enhancement- restoration: Οι δύο εκφράσεις έχουν μία βασική διαφορά. Το "enhancement" χρησιμοποιείται πιο πολύ για την περίπτωση καλυτέρευσης της ποιότητας σε ότι αφορά τα θέματα θορύβου κάθε είδους ενώ το "restoration" χρησιμοποιείται για την περίπτωση που το σήμα είναι ελλιπές και αποκαθίστανται τα σημεία που λείπουν. Οι ίδιοι όροι χρησιμοποιούνται και στις εικόνες. Έτσι μία καλή λύση είναι να αντιστοιχίσει κανείς το "ανάδειξη" στο "enhancement" και το "αποκατάσταση" στο "restoration".

template: Στο ΕΜΠ χρησιμοποιείται ο όρος "ίχνος". Οι όροι "template", "matching", "training" κ.λπ. είναι γενικοί όροι της αναγνώρισης προτύπων και δεν χρησιμοποιούνται μόνο στην φωνή.

model: Είναι δόκιμος όρος στην Ελληνική ο όρος "μοντέλο". Οι λέξεις "υπόδειγμα" η "πρότυπο" χρησιμοποιούνται με άλλη έννοια. Μοντέλο είναι κάτι τεχνητό που χρησιμοποιείται για να πλησιάσει την πραγματικότητα. Συνήθως πρόκειται για μαθηματικά δημιουργήματα (mathemetical models) τα οποία υπολογίζονται με βάση τα δεδομένα του προβλήματος ώστε να είναι όσο κοντύτερα γίνεται στα πραγματικά συστήματα. Στην θεωρία συστημάτων υπάρχει ειδική μεθοδολογία και τα μοντέλα είναι μοντέλα φυσικών συστημάτων. Στην επεξεργασία φυσικής γλώσσας χρησιμοποιούνται για να εκφράσουν την "μοντελοποίηση" των γλωσσικών φαινομένων. Οι λέξεις "υπόδειγμα" και "πρότυπο" πρέπει να αποφεύγονται να χρησιμοποιούνται αντί του "μοντέλου" γιατί μπορούν να οδηγήσουν σε σοβαρή σύγχυση.

pattern: Αυτή η λέξη προέρχεται από την γαλλική patron η οποία έχει ελληνική ρίζα (πατέρας). Το "πρότυπο" δεν έχει καμμία σχέση με "μοντέλο". Είναι μία βάση αναφοράς για σύγκριση σε συστήματα μηχανών που μαθαίνουν. Τα πρότυπα που φθάνουν στη μηχανή από τον εξωτερικό κόσμο συγκρίνονται με τα πρότυπα από τον χώρο προτύπων (χώρος μάθησης της μηχανής) ώστε να ταξινομηθούν σωστά στις κατάλληλες κατηγορίες. Δεν είναι μαθηματικά οικοδομήματα όπως τα μοντέλα, αλλά απλώς αποδεκτά δεδομένα που προέρχονται από μετρήσεις. Η λέξη ίχνος (=αχνάρι) χρησιμοποιείται για την Αγγλική λέξη "template" η οποία αντιστοιχεί σε μία ειδική κατηγορία προτύπων που προσφέρονται στο ένα - προς - ένα ταίριασμα.

standard: Και εδώ πρέπει να χρησιμοποιείται δυστυχώς ή ευτυχώς η λέξη πρότυπο που έχει έτσι δύο έννοιες. Ενώ η λέξη "standard" μεταφράζεται πρότυπο και η λέξη "quality standard" σαν "πρότυπο ποιότητας", η λέξη "standardisation" μεταφράζεται σαν "τυποποίηση" και όχι "προτυποποίηση". Το "πρότυπο" στην περίπτωση της "τυποποίησης" έχει αντίστοιχη λειτουργία με αυτό των μηχανών που μαθαίνουν γιατί αποτελεί "αναφορά για σύγκριση" και μάλλον είναι ευτύχημα η χρήση της ίδιας λέξης με δύο έννοιες. Γενικά το "πρότυπο" στα Ελληνικά θα μπορούσε να έχει σαν γενικό ορισμό: "Αποδεκτές αναφορές για σύγκριση" καλύπτοντας όλες τις έννοιες της λέξης.

υπόδειγμα: Η λέξη υπόδειγμα δεν έχει θέση ούτε στα θέματα της τυποποίησης, ούτε στα θέματα της εκμάθησης μηχανών. Έχει πιο πολύ να κάνει με θέματα συμπεριφοράς π.χ. "υπόδειγμα χαρακτήρος", "υποδειγματική συμπεριφορά", "υπόδειγμα μαθητού, δασκάλου κ.λπ.". Συνώνυμο του υποδείγματος θα μπορούσε κανείς να θεωρήσει το "παράδειγμα προς μίμηση".

unification grammar: Ενοποιητική γραμματική είναι πράγματι ο καλύτερος όρος.

interaction: "Αλληλεπίδραση" είναι ο καλύτερος και ο πιο κατανοητός όρος και μάλλον έχει επικρατήσει.

aligned parallel corpora: Προτιμούμε το "στοιχισμένα" γιατί είναι πιο εκφραστικό από το "ευθυγραμμισμένα". Το "ευθυγραμμισμένα" δεν αποδίδει την λειτουργία και είναι αποτυχία και στα Αγγλικά.

mapping: Προτιμούμε το "απεικόνιση" που είναι η κυριολεκτική έννοια. Το "χαρτογράφηση" και "σχεδίαση" είναι λάθος στο πλαίσιο της γλωσσικής τεχνολογίας.

matching: Προτιμούμε τον όρο "ταίριασμα" που έχει επικρατήσει.

recursion: είναι σαφώς "αναδρομή". Έχει καθιερωθεί εδώ και είκοσι πέντε χρόνια. Το "αναπόληση" είναι πολύ ρομαντικό για μία παρόμοια μαθηματική διαδικασία ενώ το "επιστροφή" είναι λάθος.

supervised learning: "Εκμάθηση με επίβλεψη" είναι δόκιμο. Έχει σαν συνώνυμο το "εκμάθηση με δάσκαλο" (learning with teacher). Αντίθετο είναι το **unsupervised learning** "εκμάθηση χωρίς επίβλεψη" ή "εκμάθηση χωρίς δάσκαλο".

clustering: "συσταδοποίηση" είναι ένας επιτυχημένος όρος και “συστάδα” για το "cluster".

Ελπίζω σύντομα να μας δοθεί η δυνατότητα να ψηφίζουμε για τους όρους και να τους συγκεντρώσουμε σε ειδική βάση δεδομένων. Ελπίζω επίσης να τους υποβάλλουμε στον ΕΛΟΤ και στην ΕΛΕΤΟ για συζήτηση και έγκριση μόλις συγκεντρωθεί μεγαλύτερος αριθμός. Το ΙΕΛ προγραμματίζει μία ειδική υπηρεσία στο διαδίκτυο (internet) για τα θέματα της ορολογίας που είναι σχετικά με την γλώσσα και την πληροφορική.

Με τιμή
Γ. Καραγιάννης

Το ΙΕΛ προτείνει για συζήτηση και σχολιασμό τους ακόλουθους όρους:

- **expert systems**: έμπειρα, εμπειρογνώμονα ή εξειδικευμένα συστήματα
  (βλ. Σπ. Τζαφέστα, *Εισαγωγή στην Τεχνητή Νοημοσύνη και τα Έμπειρα Συστήματα*, Β’ Έκδοση, Αθήνα 1996).
  και εναλλακτικά: ειδήμονα συστήματα
- **annotation**: σχολιασμός, χαρακτηρισμός
  (Τμήμα Ηλεκτρονικής Λεξικογραφίας ΙΕΛ)
- **tagging**: γραμματικός χαρακτηρισμός
  (Τμήμα Ηλεκτρονικής Λεξικογραφίας ΙΕΛ)
- **tagging**: σημάδεμα, μαρκάρισμα
  (Ν. Νάσσος - Τμήμα Ηλεκτρονικής Λεξικογραφίας ΙΕΛ)
- **server**: διακομιστής, εξυπηρετητής
  (Ν. Νάσσος - Τμήμα Ηλεκτρονικής Λεξικογραφίας ΙΕΛ)
- **browser**: διαφυλλιστής
  (Ν. Νάσσος - Τμήμα Ηλεκτρονικής Λεξικογραφίας ΙΕΛ)
- **information supplier**: διαθέτης πληροφορίας αντί πάροχος ή παροχός πληροφορίας που δυστυχώς τείνει να επικρατήσει.
  (Γ. Καραγιάννης - ΙΕΛ)

VI. Ειδήσεις για τη Γλωσσική Τεχνολογία / *News Related to Language Technology and Informatics issues*

Συνέδρια / *Conferences*

First International Conference on Language Resources and Evaluation
Granada, Spain
28-30 May 1998

e-mail or Fax to:
LREC Secretariat
Facultad de Traduccion e Interpretacion
Dpto. de Traduccion e Interpretacion
C/ Puentezuelas, 55
18002 GRANADA, SPAIN
tel.: +34 58 24 41 00 - fax: +34 58 24 41 04
e-mail: reli98@goliat.ugr.es.
Web: http://ceres.ugr.es/~rubio/elra.html

Conference Addresses:
The Conference Chair is Antonio Zampolli (Istituto di Linguistica Computazionale del CNR and President of ELRA).
Antonio Zampolli - LREC
Istituto di Linguistica Computazionale del CNR
via della Faggiola, 32 - 56126 Pisa, ITALY
tel.:+39 50 560 481 - fax:+39 50 555 013
e-mail: pisa@ilc.pi.cnr.it

Exhibition:
An exhibition will be organised by ELRA. This exhibition is open to companies and projects wishing to promote, present and demonstrate their language resources products and prototypes to the wide range of experts and representatives from all over the world participating in the conference. For more information on this, please contact the ELDA office on elra-elda@calva.net.

ELRA
For more information about ELRA (the European Language Resources Association), please contact:

Khalid Choukri, ELRA CEO
55-57, rue Brillat Savarin
F- 75013 Paris, France
tel.: +33 1 43 13 33 33 - fax: +33 1 43 13 33 39
e-mail: elra@calva.net
Web: http://www.icp.grenet.fr/ELRA/home.html

AIMSA'98 8th International Conference on Artificial Intelligence:
Methodology, Systems, Applications
Sozopol, Bulgaria, September 21 - 23, 1998
http://boogie.cs.unitn.it/AIMSA-98/

Conference Chair
Fausto Giunchiglia Email: fausto@irst.itc.it
Istituto per la Ricerca Scientificae
Tecnologica (IRST) Via Sommarive, Loc. Pante' di Povo
Phone: +39 461 314517 (secr.), +39 461 314436 (off.)
Fax: +39 461 302040 / 314591
38050 Trento, Italy
http://www.cs.unitn.it/~fausto

Topics
AIMSA'98 will be mainly (but not only) centered around the use of REASONING in AI, and in particular: Case-based Reasoning, Multi-Agent Systems, Planning & Temporal reasoning, SAT & Decision Procedures, Inductive Reasoning. Other topics of interest will be: Abduction, Constraint Based Reasoning, Knowledge Acquisition, Knowledge Based Systems, Learning, Natural Language Processing, Temporal and Causal Reasoning. This is not an exhaustive list, and papers from other areas are also encouraged.

Important Dates (extended):
Submission deadline: 24 April, 1998
Notification of acceptance: 5 June, 1998

Deadline for final papers: 19 June, 1998
Conference: September 21-23, 1998

EURISCON '98-MOBINET '98
Third European Robotics, Intelligent Systems & Control Conference
EURISCON '98
Divani Caravel Hotel Athens Grecce June 22-25 1998
General Chairman:
Prof. Spyros G. Tzafestas

EURISCON '98 will involve the following streams:
Stream 1: Robotics
Stream 2: Intelligent Systems
Stream 3: Control
Stream 4: Manufacturing
Stream 5: MobiNet-Workshop

Deadlines:
March 1, 1998 Receipt of abstracts
March 15, 1998 Receipt of invited sessions
April 15, 1998 Notification of acceptance
June 15, 1998 Receipt of camera-ready papers

Submission/Correspondence Addresses:
Professor Spyros G. Tzafestas
EURISCON'98
Intelligent Robotics & Automation Laboratory (IRAL)
Dept. of Electrical and Computer Engineering
National Technical University of Athens
GR-15773 Zographou, Athens, GREECE
tel.: +30-1-772 2489 (Office), +30-1-772 1527 (Lab.)
fax: +30-1-772 2490
e-mail: dkostis@robotics.ece.ntua.gr

Session:
"Towards Adaptive and Multilingual Information Extraction Systems"
Session date will be on Monday afternoon, June 22, 1998
Session Organiser: Dr. Constantine D. Spyropoulos
Research Director tel.: +30-1-6503196
Software and Knowledge Engineering Laboratory
Inst. of Informatics & Telecommunications
fax: +30-1-6532175
N.C.S.R. "Demokritos"
15310 Aghia Paraskevi, Greece
e-mail: costass@iit.nrcps.ariadne-t.gr¦

Twin Conference on "Professional Communication and Knowledge Transfer" ProCom '98
in Commemoration of the 100th Anniversary of Eugen Wuster
Vienna, 24/26 August 1998
Further details, including a program of events, exhibition details and registration information can be found at: http://www.mcs.surrey.ac.uk/AI/Wuster

Πανελλήνιο Συνέδριο Νέων Τεχνολογιών Πληροφόρησης
Αθήνα, 8 - 10 Οκτωβρίου 1998
http://www.ekt.gr/nit/events/nit98/

Οργάνωση: Εθνικό Κέντρο Τεκμηρίωσης (ΕΚΤ) και Ανθρωποδίκτυο Νέων Τεχνολογιών Πληροφόρησης
Συνδιοργάνωση με: Γεν. Γραμματεία Έρευνας & Τεχνολογίας (ΓΓΕΤ), Εταιρία Επιστημόνων Η/Υ & Πληροφορικής (ΕΠΥ)

Οι βασικές θεματικές ενότητες του Συνεδρίου περιλαμβάνουν:
- Συστήματα Πολυμέσων
- Γεωγραφικά Πληροφοριακά Συστήματα
- Σύγχρονα Θέματα Βάσεων Δεδομένων
- Το διαδίκτυο Internet και ο Παγκόσμιος Ιστός WWW
- Ηλεκτρονικές Εκδόσεις
- Ηλεκτρονικό Εμπόριο
- Ψηφιακές / Εικονικές Βιβλιοθήκες
- Θέματα Γλωσσικής Τεχνολογίας
- Ασύρματες Επικοινωνίες
- Ανάλυση, Σχεδίαση και Ανάπτυξη Συστημάτων

Επικοινωνία: Δρ. Γιάννης Θεοδωρίδης, ΕΚΤ,
e-mail: nit98@ekt.gr,
τηλ.: (1) 7210386, fax: (1) 7246824.

Data Semantics - 8 (DS-8):
"Semantic Issues in Multimedia Systems"
IFIP TC-2 Working Conference, organised by Working Group 2.6 (Database)
Rotorua, New Zealand, 5-8 January 1999
(http://zulu.cs.rmit.edu.au/~ds8)

Topics:
The topics of interest include, but are not limited to:

- Data modeling and query languages for databases media such as audio, video, computer image
- Semantic foundations for multimedia and hypermedia
- Methodological aspects of multimedia database design
- Information retrieval, knowledge discovery, and data mining
- Temporal and spatial issues in multimedia databases
- Semantic issues in standardization
- Interoperability of multimedia databases
- Relationships with high-level domain modeling approaches
- Multimedia user interfaces
- Industrial application challenges for multimedia databases

Dates:
June 30, 1998: Deadline for arrival of submissions
August 28, 1998: Notification of acceptance
September 30, 1998: Camera ready for final version to be published by Chapman & Hall
January 5-8, 1999: Conference

Conference Chair
Tharam Dillon
Dept. of Computer Science and Computer Engineering
LaTrobe University, Bundoora
VIC 3083 Australia
(fax) +61 3 9479 3060
(e-mail) tharam@latcs1.cs.latrobe.edu.au

More information about the conference and about IFIP WG2.6 can be found at URL: http://www.informatik.uni-ulm.de/dbis/IFIP-WG2.6.

Συμπόσια / *Symposia*

IMACS International Symposium on Soft Computing in Engineering
Applications SOFTCOM '98
Divani Caravel Hotel Athens Grecce
June 22-25 1998

General Chairman:
Spyros G. Tzafestas (ICCS-NTUA)

Topics include, but are not restricted to, the following:

- Fuzzy Logic and Fuzzy Computing
- Neural Networks and Neural Computing
- Genetic / Evolutionary Computing
- Hybrid Soft Computing
- Engineering Applications

Time Schedule:
Receipt of 1-page (A4) abstracts: March 1, 1998
Receipt of Invited Sessions: March 15, 1998
Acceptance notification: April 15, 1998
Receipt of full CR papers: June 15, 1998

All correspondence and submissions should be addressed to:
Professor Spyros G. Tzafestas
SOFTCOM '98
Intelligent Robotics & Automation Laboratory (IRAL)
Institute of Communication and Computer Systems
National Technical University of Athens
Zographou, Athens,
GR 15773, GREECE

tel.: +30-1-772 2489 (Office)
+30-1-772 1527 (Lab.)
fax: +30-1-772 2490
e-mail: dkostis@robotics.ece.ntua.gr

Συναντήσεις Εργασίας / *Workshops*

Towards a European Evaluation Infrastructure for NL and Speech

A workshop jointly organised by the European Network of Excellence in Language and Speech ELSNET and the EC Language Engineering-4 project ELSE to be held on Wednesday May 27, 9:00-13:00 at the

First International Conference on Language Resources and Evaluation
Granada, Spain
28-30 May 1998

The workshop is very timely as it takes place when the EC's 5th Framework Programme is taking shape. It is clear that the availability of a European evaluation infrastucture can be an important factor in European R&D activities, and that it can only be successful if it is organised and implemented on a European scale.

Contact:
Steven Krauwer
Trans 10,
3512 JK Utrecht,
The Netherlands

Phone: +31 30 253 6050
fax: +31 30 253 6000
e-mail: steven.krauwer@let.ruu.nl

The second international Workshop on Controlled Language Applications (CLAW98)
May 21-22 1998

Language Technologies Institute
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh,
PA. 15213
USA
http://www.lti.cs.cmu.edu/CLAW98/

ACL/COLING-98
Partially Automated Techniques for Transcribing Naturally Occuring, Continuous Speech

August 16, 1998
(following ACL/COLING-98)
University of Montreal,
Montreal
(Quebec, Canada)

Submissions:
Only e-mail submissions in LaTeX or Ascii will be accepted.

Authors should submit an abstract of no more than 800 words to:
trans98@cs.concordia.ca

Style files and templates for LaTeX submissions can be found at:
http://coling-acl98.iro.umontreal.ca/Styles.html

A copy of this call for papers can be found at:
http://coling-acl98.iro.umontreal.ca/Workshops.html

The official language of the conference is English.

Important Deadlines:

Submission Deadline:	April 15, 1998
Notification Date:	May 15, 1998
Camera ready copy due:	June 15, 1998

Information:
Any requests for information should be sent to:
trans98@cs.concordia.ca

Θερινά Σχολεία / *Summer Schools*

LOT Summer School 1998
Utrecht
June 15-26 1998

The LOTSummer School will take place in Utrecht. Host is UiL OTS, University Utrecht.

You can find course descriptions, enrollment forms and more information at:
http://wwwots.let.ruu.nl/LOT/zs98.html. You can also contact the LOT-secretariat,
(Christien Bok,
LOT, Trans 10, 3512 JK Utrecht,
The Netherlands,

tel.: +31(0)30-2536006,
fax: +31(0)30-2536000,
e-mail: LOT@let.ruu.nl).

Eleventh European Summer School in Logic, Language and Information
ESSLLI-99
August 1999
Utrecht, The Netherlands

The main focus of the European Summer Schools in Logic, Language and Information is the interface between linguistics, logic and computation.

Foundational, introductory and advanced courses together with workshops cover a wide variety of topics within six areas of interest: Logic, Computation, Language, Logic and Computation, Computation and Language, Language and Logic. ESSLLI-99 is organised under the auspices of the European Association for Logic, Language and Information (FoLLI).

Proposal Submission:
All proposals (subjset: ESSLLI-99) should be submitted by electronic mail to the program chair, at wansing@rz.uni-leipzig.de, in plain ASCII text as soon as possible, but no later than June 15, 1998. Authors of proposals will be notified of the committee's decision no later than September 1, 1998.

Timetable for Foundational Course Proposal Submission:

Jun 15, '98:	Proposal Submission Deadlines
Sep 1, '98:	Notification
Nov 15, '98:	Deadline for receipt of title, abstract, lecturer(s) information, course description and prerequisites
Jun 1, '99:	Deadline for receipt of camera-ready course material

Further Background Information:
To obtain further information, please visit the web site for ESSLLI-98
(http://www.coli.uni-sb.de/esslli/)
or FoLLI's home page on the web
(http://www.wins.uva.nl/research/folli/).

OSI/HESP International Summer School
Applications of Information Technologies to Biblical Textology Studies
27 July - 7 August 1998
Sofia, Bulgaria

For additional information and application forms please contact:
Summer School Applications of Information Technologies to Biblical Studies
(Attn: Milena Dobreva)
Institute of Mathematics and Informatics
Bl. 8, Acad. G. Bonchev St.
1113 Sofia
BULGARIA

tel.: (359-2) 713-2809
fax: (359-2) 971-3649
e-mail: aitbs@math.acad.bg