Transforming Spontaneous Telegraphic Language to Well-Formed Greek Sentences for Alternative and Augmentative Communication

Georgios Karberis and Georgios Kouroupetroglou

University of Athens,
Department of Informatics and Telecommunications,
Panepistimiopolis, Ilisia, Athens, Greece
{grad0350, koupe}@di.uoa.gr

Abstract. The domain of Augmentative and Alternative Communication (AAC) studies appropriate techniques and systems that enhance or accomplish the retaining or non-existing abilities for interpersonal communication. Some AAC users apply telegraphic language, as they attempt to speed up the interactive communication or because they are language impaired. In many AAC aids, a "sentence" is formulated by combining symbols of an icon-based communication system. To be accepted by the communication partner, the output should be a correct oral sentence of a natural language. In this paper we present our effort to develop a novel technique for expanding spontaneous telegraphic input to well-formed Greek sentences, by adopting a feature-based surface realization for Natural Language generation. We first describe the general architecture of the system that accepts compressed, incomplete, grammatically and syntactically ill-formed text and produces a correct full sentence. The NLP techniques of the two main modules, named preprocessor and translator/ generator, are then analyzed. A prototype system has been developed using Component Based Technology (CBT) which is under field evaluation by a number of speech-disabled users. Currently it supports fully the BLISS and MAKATON icon based communication systems. Some limitations of the module are also discussed along with possibilities for future expansions.

1 Introduction

Computer Mediated Interpersonal Communication (CMIC) (or its equivalent term e-communication) launches an important societal role for all citizens. In CMIC either voice or text is commonly used to achieve synchronous (i.e. in real time) or asynchronous (e.g. messaging, mailing) communication between two or more individuals. In some cases an alternative symbolic communication system (such as BLISS, PIC, PCS, MAKATON, SIGSYM, LEXIGRAMS, OACKLAND and REBUS) can be also utilized [1]. Traditionally, interpersonal communication is referred in the context of the assistive technology and the communication aids. However recently, general solutions have been proposed allowing communication between able-bodied and the disabled [2], [3], [4]. The domain of Augmentative and Alternative Communication (AAC) studies the appropriate techniques and the

I.P. Vlahavas and C.D. Spyropoulos (Eds.): SETN 2002, LNAI 2308, pp. 155–166, 2002.

systems that enhance or accomplish the retaining or non-existing abilities for interpersonal communication.

AAC constitutes a highly multilingual communication environment as almost an infinite number of vocabulary sets from various orthographic languages or symbol systems can be created or adapted. In the majority of the AAC aids the two partners in a communication session apply different meaning representations chosen among text and symbols. Some AAC users apply telegraphic language, because either they are language impaired or they attempt to speed up the interactive communication. Telegraphic language is very brief and concise and many words are omitted. Substitutionally, in many AAC aids, one combines elements of an icon-based system to formulate a "sentence". Nevertheless, to be accepted by the communication partner, the output should be a correct oral sentence of a natural language. In this case, as an intermediate step, the icon–based "sentence" is lexical translated to telegraphic language. Research work concerning lexical knowledge in the AAC field has focused primarily on knowledge-based rate enhancement techniques for natural languages, such as COMPANSION [5], [6] and co-generation [7]. The formal description, processing and translation of symbol systems (e.g. BLISSYMBOLICS) have also been investigated: The COMPANSION system [5] uses a statistical model for syntax analysis in English and it accepts input from the keyboard. The PVI system [9] uses Tree Adjoining Grammars (TAG) to generate French sentences. KOMBE produces also French sentences [8] from input in SODI-GRACH, which is not a widely accepted symbolic system compared to BLISSYMBOLICS. Whereas morphological treatment in some languages, like the English, seems to be relatively simple, it can become a central issue for highly inflectional languages, such as Hungarian [17], Basque [18] and Greek. In the literature there are also proposals concerning the exploitation of already existing large-scale lexical resources in AAC [10]. The role of multilingual lexical linguistic information and lexical translation relations for orthographic languages and symbolic systems has been discussed in depth in [11], [12].

This paper presents a new approach for expanding spontaneous telegraphic input in AAC to well-formed sentences for the Greek language. The adopted method consists of a feature-based surface realization for Natural Language generation. We first describe the general architecture of the system that accepts compressed, incomplete, grammatical and syntactically ill-formed text and produces a correct full sentence. The NLP techniques of two main modules, named preprocessor and translator/generator, are then analyzed. Furthermore, a prototype of such a system developed using Component Based Technology (CBT) is given. Some limitations of the module are also discussed along with possibilities for future expansions.

2 General Architecture

The general architecture of a typical AAC system is given in Fig. 1. A language impaired user, through an appropriate to his/her abilities input device (such as switch {either mechanical or infrared or acoustic}, touch screen, trackball, mouse, head-stick, touch tablet) selects a number of symbols from the selection set in order to form a message [13]. This icon-based "sentence" is then lexically translated to telegraphic form [11]. Non-language impaired users can formulate telegraphic sentences directly.

The Telegraphic-to-Full Sentence (TtFS) module produces a well-formed complete sentence to drive either a Text-to-Speech system or to reach the communicator partner through alternative output forms (e.g. chat, e-mail, print).

We concentrate our work on developing the TtFS module, which will work as a language text translator/generator. The input is a compressed, incomplete, grammatical and syntactically ill-formed Greek text. The output of the module is a full Greek sentence, grammatical and syntactically correct. The module uses a database with morphological, syntactic and semantic knowledge.

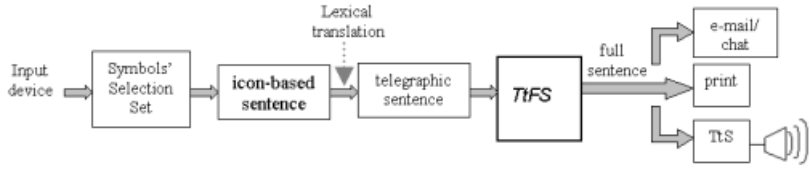

Fig. 1. General architecture of an AAC system

Two modules can accomplish the whole processing of the TtFS:

1. The *preprocessor* module, which splits the sentence into words and identifies the part of speech for each of them. The preprocessor also adds any missed words, such as articles. The output of the module is a full sentence, but grammatical incorrect.
2. The *translator/generator* module, which applies syntactic and grammatical rules as well as semantic information to the output of the preprocessor and generates a well-formed complete Greek sentence.

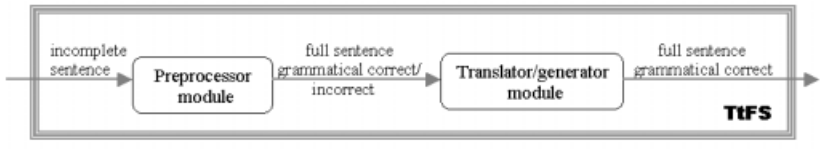

Fig. 2. The architecture of the TtFS

3 The Preprocessor

The preprocessor module has a linear, step-by-step architecture (Figure 3). Initially, the preprocessor module splits the input sentence into words. For every word a small chunk is created. One part of the chunk is the word class of the given word. The above information is retrieved from a computational morphological and syntactic lexicon of Modern Greek. Currently the module includes 3000 lemmas, in order to support the full icon based communication system BLISSYMBOLICS [14], [15],

[16], as well as MAKATON. Then, the preprocessor checks each word and applies the following syntactic rules:

1. If there is not any conjunction in the sentence, it assumes that there is only one main sentence. If there is at least one conjunction, it creates two or more different sentences. The first sentence is the main and the others are subordinate clauses. The remaining processing is taking place for each of them.
2. If the user omits to add an article before a noun, or before an adjective, then a new chunk, which contains the article, is added. The order of this chunk in the list is just before the chunk of the noun or the adjective, correspondingly. This processing is taking place only for nouns and adjectives before the verb of the sentence.
3. It is examined if for each verb a noun follows. If this is the case, the semantics of the noun are checked and the omitted words are added in a chunk between them.

The following cases are checked for the semantics of a noun: place, person, food, drink, object, vehicle. The semantic information for the verbs and the nouns is encoded in the database of the system. In this database there is also stored the lexicon and morphological information for every lemma.

Then, the preprocessor module merges all the words into a sentence. This sentence is a full one, as there are not missed words, but it is grammatical incorrect.

4 The Translator/Generator

The translator/generator module inputs a full sentence from the preprocessor and applies syntactic and grammatical rules to generate a well-formed sentence. Figure 4 presents the architecture of this module.

Initially, it splits the sentence and creates a list of chunks. Every chunk corresponds to a specific word of the sentence. Each chunk has five fields: number, gender, tense, person and case. The default value for each field is:

tense	:	present
case	:	nominative
gender	:	masculine
number	:	singular
person	:	first

After the creation of the chunks list, the word class for each word is retrieved from the lexicon. Then, depending on the word class of each word, the fields' values on its corresponding chunk changes properly. For example, if one word recognized to have a plural number, then the field "number" changes to *plural*, or if a pronoun is "εσύ" ("you") then the "person" changes to *second* and if there is a pronoun like " αυτός" ("he") it changes to *third*. Next, the sentence is checked for conjunctions. Then it is separated in one main and subordinate clauses, if there are any.

After these initializing steps, the module acts out the syntactic analysis. It uses syntactic rules to determine the subject and the object complement for each sentence.

First, it examines each sentence to identify the subject. The identification is accomplished with the use of syntactic patterns. The patterns of subjects that are used in the Greek language are stored in a table of the database (see Table 1). The sentence is checked from the start to its verb, if there is any. The check is performed

sequentially using all the patterns of the database. The order of storing them in the database is from the larger to the smaller. If there is no subject in the sentence, the module assumes that the subject is the pronoun "εγώ" ("I"). For the special case a subordinate clause has no subject, the subject of the main sentence is assumed.

The examination for object complements is follwing. The technique used is the same as described above. The rest of the sentence (without the subject) is checked with the patterns for object complements stored in the database. In the case the verb is a transitive one, the object complement is labeled as a predicate.

Table 1. The patterns used to identify the subject and the object complement. Abbreviations: PRE=Ppreposition, ART=article, PAR=particle, NN=noun, PRO=pronoun, ADJ=adjective, CON=conjuction and NUM=number

Subject	Object Complement
PRE+ART+PAR	PRE+ART+NN+CON+ART+NN
ART+NN+CON+ART+ADJ+NN	PRE+ART+ADJ+NN
ART+ADJ+NN+CON+ART+NN	ART+NN+CON+ART+NN
ART+NN+CON+ART+NN	PRE+ART+NN
ART+PAR+NN	PRE+ART+PAR+NN
ART+ADJ+NN	PRE+ART+NN+ART+ADJ
PRO+CON+ART+ADJ+NN	ART+ADJ+NN
ART+ADJ+NN+CON+PRO	ART+ADJ
ART+NN+PRO+CON+PRO	PRE+ART+ADJ
ART+NN+CON+PRO	ART+PAR+NN
PRO+CON+ART+NN+PRO	ART+NN
ART+NN+CON+PRO	PRE+ART+PAR
PRO+CON+ART+NN+PRO	PRE+ART+NUM
PRO+CON+ART+NN	ART+NUM
ART+ADJ	ART+PAR
PRO+CON+PRO	ART+ADJ
ART+NN+PRO	PRE+PRO
ART+NUM	PRE+NN
ART+PRO	PRO
ART+NN	ADJ
ART+PAR	PRE+ADJ
PAR	NN
NN	#
ADJ	#

After the assessment for subjects, object complements and predicates, the module applies some grammatical rules of the Greek language, to the chunks list. During the application of the rules, the field values of the chunks list are changed. The goal of this process is label each chunk with the appropriate number, gender, tense and case, according to the grammatical rules of the Greek language, (described in 4.1).

The final step of the module is to inflect every word to its correct tense, gender and number, according to the values of the labels of its corresponding chunk. The inflection is accomplished with the use of morphological knowledge. The inflection technique is described in section 4.2.

After the generation of the inflected words, the module merges them to produce a well-formed sentence, which represents the output of the whole module.

4.1 Syntactic and Grammatical Rules

The translator/generator module applies the following grammatical and syntactic rules of the Greek language (all the rules are being applied before the inflection of each word):

Agreement between the subject and the verb of a sentence.
The object complement in accusative.
The predicate in nominative.

Agreement between the subject and the verb of a sentence
The verb of a sentence must have the same number and the same person with the subject of the sentence where it belongs. If a subordinate clause doesn't have a subject, then the module assumes that the subject is the same with that of the main sentence. For example:

Input sentence:	{αυτός}+{έχω}	({he}+{have})
Output sentence:	"Αυτός έχει"	("He has")

At the starting point the default values of the labels for the verb "έχω" ("have") are *singular* for the number and *first* for the person. The label person of the word "αυτός" ("he") is *third*. Before the inflection of each word, the corresponding labels of the verb "έχω" ("have"), changes to become the same as the subject "αυτός". Thus, the inflected word for the verb "'έχω" ("I have") is "έχει" ("he has") and the output sentence is well formed.

The object complement in accusative
The object complement of a sentence must always be in accusative. Thus, for all the words that consist the object complement, the label for case changes to *accusative*. For example, consider the following sentence to be handled by the TtFS system:

Input:	{μητέρα}+{κάθομαι}+{καρέκλα}
	{mother}+ {sit} +{chair}
Output:	"Η μητέρα κάθεται στην καρέκλα"
	"The mother is sitting on the chair"

The preprocessor module adds: a) an article before the noun "μητέρα" ("mother") and b) the prepositional article "στο" ("on+the") before the noun "καρέκλα" ("chair"). The word "στο" ("on+the") suits for the verb "κάθομαι" ("sit") if a noun follows with the meaning of an item, like "καρέκλα" . The input sentence for the second module is:

{άρθρο}+{μητέρα}+{κάθομαι}+{στο}+{καρέκλα}. In the second module, every word, by default, has its case label in *nominative*. The object complement of the sentence is: {στο}+{καρέκλα}, thus, the label for the case changes to *accusative* and, after the inflection, the output sentence is well formed.

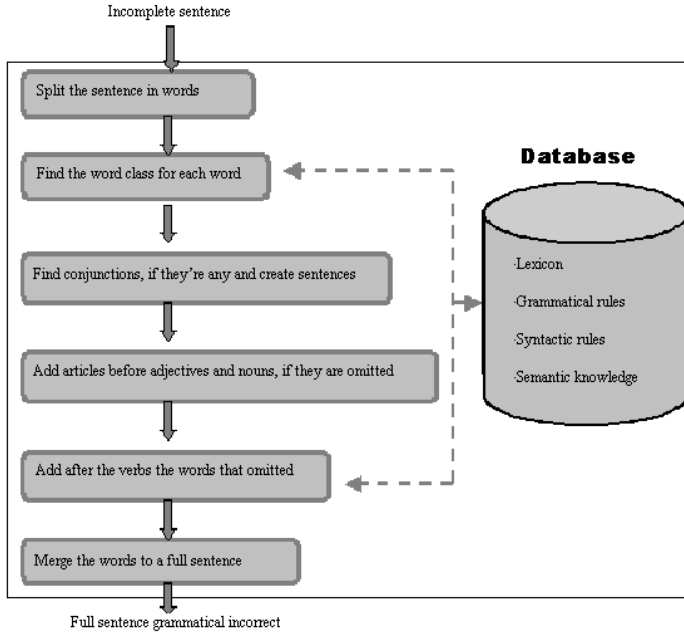

Fig. 3. The architecture of the *Translator/Generator* module

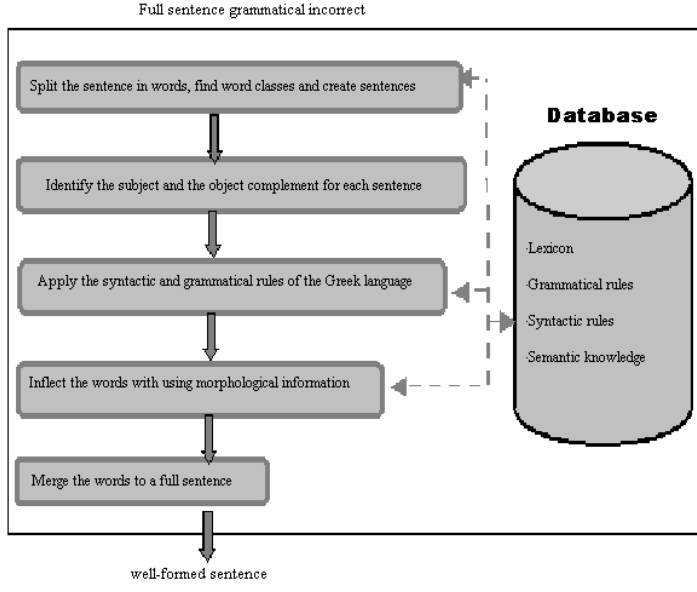

Fig. 4. The architecture of the *Preprocessor* module

The predicate in nominative
The predicate of a sentence must always be in *nominative*. We notice that the predicate gives an attribute to the subject. Hence, it takes the person and the number of the subject. For example:

Input sentence : {γονείς} + {είμαι}+{νέος}
{parents}+ {I am} +{young}
Output sentence: "Οι γονείς είναι νέοι"
"The parents are young"

The preprocessor module adds the word "άρθρο" (article) before the noun "γονείς" ("parents"). The input sentence for the second module is: {άρθρο} + {γονείς}+{είμαι}+{νέος} ({article}+{parents}+{I am}+{young}). The subject of the sentence is: {άρθρο}+{γονείς} ({article}+{parents}) and the word "νέος" ("young") is a predicate, because the verb "είμαι" ("I am") is a transitive one. The number of the subject is found to be *plural* because after the word "γονείς" ("parents") the article doesn't give any information. Hence, the labels of the verb "είμαι" ("I am") and the predicate "νέος" ("young") change according to the labels of the word "γονείς" ("parents").

4.2 Knowledge Needed

The module uses a specially designed database, which stores the lexicon with appropriate syntactic and morphological knowledge.

Lexicon
The lexicon consists of a database table and it is open to include any Greek word. The module is independent from the symbol selection set.

Syntactic information
The module uses the following syntactic knowledge:

- Patterns for the identification of the subject, e.g. ART+ADJ+NN (where ART is an article, ADJ is an adjective, and NN is a noun).
- Patterns for the identification of the object complement, e.g. ART+NN (where ART is an article and NN is a noun).
- Special patterns for the omitted words after the verbs, e.g. in the case of the verb "πηγαίνω" ("going") the stored information is:

Word	πηγαίνω	going
Item	στο	to+*article*
Person	μαζί+ με+*άρθρο*	with+*article*
Food	για	for+*article*
Drink	για	for+*article*

Place	στο	to+*article*
Game	για	for+*article*
Time	*άρθρο*	at+*article*
Vehicle	με+*άρθρο*	with+*article*
Verb	να	to

For example, if the noun is "σχολείο" ("school") which is a place, then the omitted Greek word is the prepositional article "στο" ("to"+ *article*). If the noun is a vehicle, like "λεωφορείο" ("bus"), then the words are "με"+*άρθρο* ("with"+ *article*)", and if it is a person, like "μητέρα" ("mother") then the words are "μαζί με"+άρθρο ("with"+ *article*)".

Morphological information/ word inflection

At the final step of the process, every word is inflected using the label values from its corresponding chunk. The morphological information for the inflection of each word is stored in the database. In reality, the word class and an inflection code along with its stem are specified for the members of the lexicon. This code corresponds to a pattern of specific endings. We have encoded the complete ending patterns for all the Greek words. For example, all the nouns in the Greek language have one out of 56 different endings. These 56 endings are stored only once. For each noun, adjective, article or verb we stored only the inflection code that corresponds to the correct pattern ending and the stem of the word. The module generates all the inflectional morphemes using the inflection code and the stem of the word. Thus, by combining the stem with the corresponding ending, it can inflect a word in every number, any tense, any gender and any case, according to the label values and the corresponding word class. Thus, the following values are retrieved for each word class:

- Verb tense, person, number
- Noun number, case
- Adjective number, gender, case
- Participle number, gender, case
- Pronoun number case
- Article number, gender, case
- Number number, case
- Adverb null
- Preposition null
- Conjunction null

5 Implementation

The module TtFS described above has been developed using Component Based Technology (CBT) for effective integration, as an independent component, in thorough CMIC applications. As a first test, it has been incorporated under the ULYSSES framework [4], which facilitates the integration of multi-vendor

components into interpersonal communication applications. The module has not any user interface in a real AAC aid. Although, a special user interface has been developed to accomplished laboratory alpha tests. The output of the module possesses an appropriate format [19], [20] to drive even advanced modern Text-to-Speech systems [21]. Currently, the system supports fully the icon based languages, BLISS and MAKATON. The software implementation has been accomplished with Visual Basic of the MS-Visual Studio, version 6.0. The module was built as an ActiveX DLL (Dynamic Library Link) and it can be easily installed and used as an independent intermediate component in AAC aids.

6 Discussion

The TtFS prototype is under field evaluation by a number of speech-disabled users at different regions of Greece. Preliminary results are very positive regarding its usability in real time everyday spontaneous interpersonal communication sessions.

The following limitations has to be taken into account for the current version of TtFS:

Unlimited lexicon. TtFS requires a rather large amount of information to be associated with each word. This knowledge must be hand coded on the database of the module. An effort is under way to handle the problem of unrestricted vocabulary using automatic methods for deriving the necessary information from either on-line lexical resources or from corpus-based processing.

User Input Assumptions. The module assumes that the input would reflect the basic word order of the desired output. Additionally, the subject should be first in the sentence and the object complement second. Furthermore, function words may be left out, but content words must be included.

Future expansions of the TtFS may include:

- Extension of the module's functionality to support other major icon-based symbolic communication systems used in AAC.
- Text input from the keyboard.
- Support of random telegraphic input.
- Support the opposite function of TtFS: transforming well-formed Greek sentences to a corresponding symbol sequence "sentence" for a specific icon-based system.

7 Conclusion

In this paper we have presented a novel technique for expanding spontaneous telegraphic input to well-formed sentences for the Greek language by adopting a

feature-based surface realization for Natural Language generation. The general architecture of the system that accepts compressed, incomplete, grammatical and syntactically ill-formed text and can produce a correct full sentence has been described along with the NLP techniques of two main modules, named preprocessor and translator/generator. A prototype of such a system has been developed using Component Based Technology (CBT) as a part of a number of AAC aids designed for different user groups. The system supports currently the full icon based AAC systems BLISS and MAKATON. The system is under field evaluation by a number of speech-disabled users at different regions of Greece.

Acknowledgments. Part of the work reported in this paper was carried out within the framework of the AENEAS project (contract 98AMEA19), funded by the EPET II Programme of the Greek General Secretariat for Research and Technology.

References

1. von Tetzchner, S.: Use of Graphic Communication Systems in Telecommunication. In: von Tetzchner, S. (ed.): Issues in Telecommunication and Disability. CEC, DG XIII, Luxembourg (1992) 280-288
2. Kouroupetroglou, G., Viglas, C., Stamatis C., Pentaris, F.: Towards the Next Generation of Computer-based Interpersonal Communication Aids. In: Anogianakis, G., Buhler, C. and Soede, M. (eds.): Advancement of Assistive Technology. Assistive Technology Research Series, Vol. 3. IOS Press, Amsterdam Berlin Oxford Tokyo Washington (1997) 110-114
3. Viglas, C., Stamatis, C., Kouroupetroglou, G.: Remote Assistive Interpersonal Communication Exploiting Component Based Development. In: Edwards, A., Arato, A., Zagler, W. (eds.): Computers and Assistive Technology, Proceedings of the XV IFIP World Computer Congress, 31 August - 4 Sept. 1998, Vienna – Budapest, Congress, ICCHP'98, (1998) 487-496
4. Kouroupetroglou, G., Pino, A.: ULYSSES: A Framework for Incorporating Multi-Vendor Components in Interpersonal Communication Applications. In: Marincek C., Buhler, C., Knops, H., Andrich, R. (eds.): Assistive Technology – Added Value to the Quality of Life. Assistive Technology Research Series, Vol. 10. IOS Press, Amsterdam Berlin Oxford Tokyo Washington (2001) 55-59
5. McCoy, K. F., Pennington, C. A., Badman, A. L.: Compansion: From Research Prototype to Practical Integration. Natural Language Engineering **4** (1): Cambridge University Press, (1998) 73-95
6. McCoy, K., Demasco, P.: Some Applications of Natural Language Processing to the Field of Augmentative and Alternative Communication. Proceedings of the IJCAI '95 Workshop on Developing AI Applications for Disabled People, Montreal, (1995) 97-112
7. Copestake A.: Augmented and Alternative NLP Techniques for Augmented and Alternative Communication. In: Copestake. A., Langer, S., Palazuelos-Cagigas, S. E. (eds.): Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL) and the 8th Conference of the European Chapter of the ACL.. Morgan Kaufmann, San Francisco (1997) 37-42

8. Guenthner F., Kruger-Thielmann, K., Pasero, R., & Sabatier, P.: Communication Aids for Handicapped Persons. In: Proceedings of the 2nd European Conference on Advances in Rehabilitation Technology (ECART2), May 26-28, 1993, Stockholm (1993) 1-4
9. Vaillant P.,: A semantics-based communication system for dysphasic subjects. Proc. Of the 6th Conf. On Artificial Intelligence in Medicine Europe, AIME '97, Grenoble, France (1997)
10. Zickus, M. W., McCoy, K. F., Demasco, P. W., Pennington, C. A.: A lexical Database for Intelligent AAC Systems. In: Langton, A. (ed.): Proceedings of the 18th Annual Conference of the Rehabilitation Engineering Society of North America (RESNA), Washington, DC: RESNA Press, (1995) 124-126
11. Antona M., Stephanidis C., Kouroupetroglou G.: Access to Lexical Knowledge in Modular interpersonal Communication Aids. Augmentative and Alternative Communication, **15**, (1999) 269-279
12. Antona M., Stephanidis C., Kouroupetroglou G.: Vocabulary Management in Modular Interpersonal Communication Aids. In: Anogianakis, G., Buhler, C. and Soede, M. (eds.): Advancement of Assistive Technology. Assistive Technology Research Series, Vol. 3. IOS Press, Amsterdam Berlin Oxford Tokyo Washington (1997) 200-205.
13. Kouroupetroglou G., Pino, A., Viglas, C.: Managing Accessible User Interfaces of Multi-Vendor Components under the ULYSSES Framework for Interpersonal Communication Applications. In: C. Stephanidis (ed) Universal Access in HCI. Lawrence Erlbaum Ass, (2001) 185-189
14. Bliss C. K.: Semantography-Blissymbolics, 3rd edition. Semantography-Blissymbolics Publications (1978)
15. McDonald, E. T.: Teaching and using Blissymbolics. Blissymbolics Communication Institute, Canada (1985)
16. McNaughton, S.: Communicating with Blissymbolics. Blissymbolics Communication Institute, Canada (1985)
17. Garay-Vitoria, N., Abascal, J. G.: PROFET. Word Prediction for Inflected Languages. Application to Basque Language. Proc. Of the Workshop on Natural Language Processing for Communication Aids. Madrid, Spain (1997) 29-36
18. Olaszi, P., Koutny, I., Kalman, S.: From BLISS Symbols to Grammatically Correct Voice Output: A Communication Tool for People with Disabilities. Int. Journal of Speech Technology, **5** (1) (2002) 49-56
19. Xydas, G., Kouroupetroglou, G.: Text-to-Speech Scripting Interface for Appropriate Vocalisation of e-Texts. Proc. of EUROSPEECH 2001, Aalborg, Denmark (2001) 2247-2250
20. Xydas, G., Kouroupetroglou, G.: Augmented Auditory Representation of e-Texts for Text-to-Speech Systems. In V. Matousek (eds.): Text, Speech and Dialogue. Lecture Notes in Artificial Intelligence Vol. 2166. Springer-Verlag, (2001) 134-141
21. Xydas, G., Kouroupetroglou, G.: The DEMOSTHeNES Speech Composer. Proc. of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Perthshire, Scotland (2001)