Annotation and Maintenance of the Greek Sign Language Corpus (GSLC)

Eleni Efthimiou, Stavroula - Evita Fotinea
Institute for Language and Speech Processing (ILSP) / R.C. Athena
Artemidos 6 & Epidavrou,
GR 151 25 Athens
Greece
E-mail: eleni_e@ilsp.gr, evita@ilsp.gr

Abstract

This paper presents the design and development of a representative language corpus for the Greek Sign Language (GSL). Focus is put on the annotation methodology adopted to provide for linguistic information and annotated corpus maintenance and exploitation for the extraction of a linguistic model intended to support both sign language recognition and creation of educational content.

1. Introduction

The Greek Sign Language (GSL) has developed as a minority non-written language system -in a socio-linguistic environment similar to those holding for most other known sign languages- used as the mother language of the Greek deaf community.

Video recordings of GSL have been produced for various reasons but, the development of the Greek Sign Language Corpus (GSLC) is the first systematic attempt to create a re-usable electronic language corpus organised and annotated according to principles deriving from requirements put by specific application demands (Mikros, 2004). The GSLC is being developed in the framework of the national project DIANOEMA (GSRT, M3.3, id 35) that aims at optical analysis and recognition of both static and dynamic signs, incorporating a GSL linguistic model for controlling robot motion. Linguistic analysis is a sufficient component for the development of NLP tools that, in the case of sign languages, support deaf accessibility to IT content and services. To effectively support this kind of language intensive operations, linguistic analysis has to derive from safe language data -defined as data commonly accepted by a specific language community- and also provide for an amount of linguistic phenomena, which allow for an adequate description of the language structure. The GSLC annotation features have been, however, broadly defined to serve multipurpose exploitation of the annotated part of the corpus. Different instantiation of corpus reusability are provided by measurements and data retrieval, which serve various NLP applications along with creation of educational content.

2. Development and maintenance of GSLC

2.1 Corpus development

A definition of corpus provided by Sinclair (1996) in the framework of the EAGLES (http://www.ilc.cnr.it/EAGLES) project, runs as follows: "A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language". Furthermore, the definition of computer corpus in the same document crucially states that: "A computer corpus is a corpus which is encoded in a standardised and homogenous way for open-ended retrieval tasks…".

Here we will use the term corpus as always referring to an electronic collection of pieces of language, also adopting the classification by Atkins et al. (1991), which differentiates corpus from a generic library of electronic texts as a well defined subset that is designed following specific requirements to serve specific purposes. Among the most prominent purposes for which oral language (written) electronic corpora are created, lies the demand for knowledge management either in the form of information retrieval or in the form of automatic categorisation and text dispatching according to thematic category. Electronic corpora differentiate as to intended use and the design requirements that they fulfil.

The design of GSLC content has been led by the demand to support sign language recognition as well as theoretical linguistic analysis. In this respect, its content organisation makes a distinction between three parts on the basis of the utterance categories to be covered.

The first part comprises a list of lemmata which are representative of the use of handshapes as a primary sign formation component. This part of the corpus is developed on the basis of measurements of handshape frequency of use in sign morpheme formation, but it has also taken into account the complete set of sign formation parameters. In this sense, in order to provide data for all sign articulation features of GSL, the corpus also includes characteristic lemmata with respect to all manual and non-manual features of the language.

The second part of GSLC is composed of sets of controlled utterances, which form paradigms capable to expose the mechanisms GSL uses to express specific core

grammar phenomena. The grammar coverage that corresponds to this part of the corpus is representative enough to allow for a formal description of the main structural-semantic mechanisms of the language.

The third part of GSLC contains free narration sequences, which are intended to provide data of spontaneous language production that may support theoretical linguistic analysis of the language and can also be used for machine learning purposes as regards sign recognition.

All parts of the corpus have been performed by native signers under controlled conditions that guarantee absence of language interference from the part of the spoken language of the signers' environment (DIANOEMA Project, 2006a; 2006b), whereas quality control mechanisms have been applied to ensure data integrity.

2.2 Content selection

The initial target of sign recognition imposed the demand for the collection of lists containing representative lemmata, capable to exhibit the articulation mechanisms of the language. These lists may provide a reliable test bed for initial recognition of single articulation units. Lemmata lists comprising the first part of the GSLC involve two categories, (i) commands related to robot motion control and (ii) simple and complex sign morphemes, representative of the basic vocabulary of GSL.

Morpheme selection was based on the minimum requirement of handshape frequency of occurrence, that imposed use of at least the 15 most frequent handshapes, which are responsible for the formation of a 77% of the whole amount of lemmata met in the environment of primary school education (unpublished measurement, V. Kourbetis: personal communication). Both categories contained simple and complex signs, taking into account the use of either one, or two hand formations. Except for handshapes, all other articulation parameters have been taken into account in lemma content design. These parameters include the sets of manual and non-manual features of sign formation and involve location, palm orientation, movement of the hand as well as facial expressions and head and body movement (Stokoe, 1978).

Internal organisation of lemmata lists includes categorisation according to motion commands, location indicators, number formation, finger spelling, temporal indicators, various word families, GSL specific complex sign roots and the standard signing predicate categories.

The video-corpus contains parts of free signing narration, as well as a considerable amount of elicitated grouped signed phrases and sentence level utterances, reflecting those grammar phenomena of GSL that are representative for the structural organisation of the language. Theoretical linguistic analysis of such data allows for extraction of safe assumptions as regards the rule system of the language and also provides a safe ground for the use of phrase level annotation symbols.

When structuring the phenomena list that are represented by controlled sentence groups in the video-corpus, a number of GSL specific linguistic parameters were taken into account, with the target to capture the main multi-layer articulatory mechanisms the language uses to produce phrase/sentence level linguistic messages, along with distribution within utterances of a significant number of semantic markers for the expression of quantity, quality and schema related characteristics. The two parts of the video-corpus (free narration and controlled sentences per grammar phenomenon) function complementarily as regards the target of rule extraction for annotation purposes and machine learning for sign recognition.

The phenomena for which GSLC provides extensive paradigms (Efthimiou, Fotinea & Sapountzaki, 2006) include the GSL tense system with emphasis on major temporal differentiations as regards present, past and future actions in combination with various aspectual parameters, multi-layer mechanisms of phrase enrichment for the expression of various adverbial values in phrase or sentence level, the use of classifiers, affirmation with all types of GSL predicates, formations of negation, WH- and Yes/No question formation, various control phenomena and referential index assignment.

In order to receive unbiased data, a strict procedural rule was to avoid any hint to natural signers as to preference in respect to sentence constituents ordering. In cases of deviation from neutral formations as when expressing emphasis, instructions to informants focused on the semantic dimension of the tested sentence constituent, rather than on possible structural arrangements of the relevant utterances. Furthermore, with the general aim to eliminate external destructions (such as environment language interference), the use of written Greek was excluded from communication with the natural signers.

2.3 Evaluation of the video-corpus

In order to ensure prosodic and expressive multiplicity, it has been decided to use at least 4 signers for the production of GSLC in all three parts of the corpus content. The selection of natural signers has been based on theoretical linguistics criteria related to mother language acquisition conditions (White, 1980;.Mayberry, 1993. Signers chosen to participate in GSLC production should, hence, be deaf or bilingual hearing natural GSL signers, raised in an environment of deaf natural signers. This selection criterion strictly excludes the use of deaf signers that are not natural GSL signers, in order to ensure the highest degree of linguistic integrity of the data, and, at the same time, eliminate –if not completely make vanish of– the language interference effects from Greek to GSL throughout the development of the video-corpus.

Upon completion of the GSLC video recording, uninformal quality control procedures have been followed targeting at high degrees of acceptance of the video-recorded signing material. Each part of the video-corpus had to be evaluated by natural signers, on the basis of peer review, with respect to intelligibility of the linguistic message. In case a video segment was judged poorly, the segment had to be re-collected and re-evaluated, hence, ensuring that only highly judged video segments are included in the GSLC.

3. Corpus annotation

3.1 Morpheme level annotation

. Technological limitations regarding annotation tools often impeded the use of data synchronised with video. The situation has slowly started to change as, at an experimental level, open tools have been started to develop to suit the needs of sign language annotation. Research projects, as the European ECHO (http://www.nmis.isti.cnr.it/echo) (2000-2004) and the American SignStream (http://www.bu.edu/asllrp/SignStream/) of the National Center for Sign Language and Gestures Resources (Boston University, 1999-2002) (Neidle, 2002) produced video-corpora that complied to a common set of requirements and conventions. Tools such as the iLex (Hanke, 2002) attempt to solve issues related to convention integrity of data, arising from the lack of a writing system which follows orthographic rules. In the same context, the Nijmegen Metadata Workshop 2003 (3. Crasborn, & Hanke, 2003) proposed a common set of metadata for use by sign language video-corpora.

The definition of annotation features assigned to a given signing string, reflects the extent of the desired description of grammatical characteristics allotted to the 3-dimensional representation of the linguistic message. Basic annotation fields of GSLC involve glosses for Greek and English, phrase and sentence boundaries, dominant and non-dominant hand information, eye-gaze, head and body movement and facial expression information, as well as grammar information such as tags on signs and grammar phenomenon description to facilitate data retrieval for linguistic analysis.

Starting from the need for theoretical linguistic analysis of minimal grammatically meaningful sign units, as well as the description of articulation synthesis of basic signs, the term sign morpheme has been adopted to indicate the level of grammatical analysis of all simple sign lemmata.

For the annotation of the video-corpus at the morpheme level, the basic phonological components of sign articulation, for both manual and non-manual features, have been marked on a set of representative simple morphemes and complex signs. For the representation of the phonological characteristics of the basic morphemes the HamNoSys (Hamburg Sign Language Notation System, http://www.sign-lang.uni-hamburg.de/projects/HamNoSys.html) annotation system is used (Prillwitz et al., 1989)

The characteristics of sign articulation are (sometimes dramatically) modified when moving from lemma list signing to phrase construction, where prosodic parameters and various grammar/agreement markers (i.e. two-hands plural) impose rendering of lemma formation, subject to phrase articulation conditions. Hence, recognition systems have to be taught to correctly identify the semantics of lemmata incorporated in phrase formations. Furthermore, accurate morpheme level annotations serve sign synthesis systems that have to produce utterances with the highest possible level of naturalness.

3.2 Sentence level annotation

Fully aligned with the phenomena list composing the controlled sentence groups of GSLC content, phrase level annotation focuses on coding the basic mechanisms of multi-layer articulation of the sign linguistic message and distribution of the most important semantic markers for the indication of qualitative, quantitative and schematic values. Both multi-layer articulation and semantic deixis are major characteristics of sign phrase articulation, whereas in the context of free narration, one major demand is the correct assignment of phrase boundaries.

Some of the most representative phrase level phenomena of GSL concern multi-layer articulation over one temporal unit that results in modification of the basic components of the sign phrase (Efthimiou, Fotinea & Sapountzaki, 2006). In the context of a nominal phrase, this is related to i.e. adjectival modification. The same holds for the articulation of predicative and nominal formations, which incorporate classifiers, or when providing tense indicators. A different type of phrasal annotation is adopted to indicate topicalisation of a phrase, irrespective of its grammatical category.

Sentence level annotation aims at providing for reliable extraction of sentence level structure rules, incorporating basic multi-layer prosodic articulation mechanisms, question formation and scope of quantification and negation.

For the safe use of GSLC, a subset of sentences, which are representative for all phenomena contained in the corpus, have been manually annotated. In free narration parts, sign utterance boundaries are manually marked according to generally accepted temporal criteria (segmentation boundary is set at the frame where the handshape changes from the last morpheme of the current signing string to the first morpheme of the next) and according to annotators' language feeling.

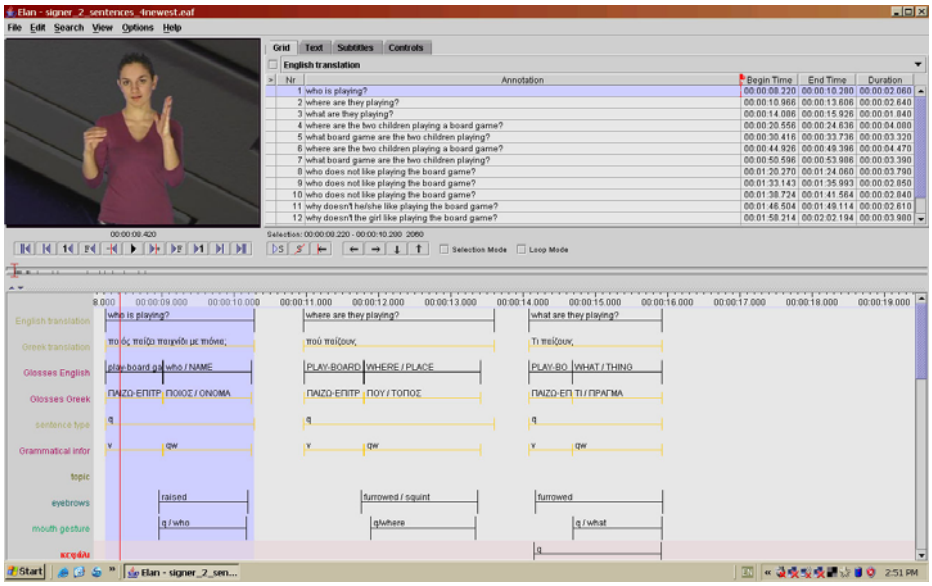

Figure 1: Annotation and retrieval of WH-question data in GSL.

The chosen annotation system is ELAN (Eudico Linguistic Annotator) the key characteristics of which are in a nutshell summarised next. ELAN (version 2.6) is an annotation tool that allows creation, editing, visualisation and retrieval of annotations for video and audio data, aiming at providing a sound technological basis for the annotation and exploitation of multi-media recordings. Figure 1 provides an instantiation of the GSLC annotation and retrieval procedure.

3.3 Evaluation of the annotated corpus

Assignment of annotations to GSLC involves two expert GSL annotators with expertise in sign language linguistics and sign language technological issues.

Annotation quality control is based on peer-review with annotation control on sample video-corpus parts, on a mutual basis by the expert annotators. Additionally, one external GSL expert annotator executes peer sample quality control on the whole annotated video-corpus. The parts of the annotated video-corpus for which conflicting evaluation reports are provided, are discussed among the three evaluators resulting in a commonly approved annotation string that is finally taken into account.

4. Exploitation of the annotated corpus

4.1 Extraction of measurements for sign recognition

In the context of DIANOEMA project, a linguistic model had to be extracted from GSLC, aiming to enhance recognition results as regards possible ambiguity or misclassified components. The linguistic model was the result of various measurements and of those parameters which formulate them as, for example, the total duration of annotated video with signing data, the set of annotation tiers, the number of lemmata which have been assigned some feature, or the set of features been assigned.

The phenomena of interest were identified and various retrieval procedures were applied in the annotated corpus in order to collect a representative sample of their instantiations. Measurements of occurrences of the different instantiations of a phenomenon allowed for mapping conditions, which rule its different realizations. As a consequence, it was possible to evaluate most productive mechanisms of utterance and incorporate them to the linguistic model intended to perform smoothing of the recognition outcome.

The various retrieval operations performed on the total duration of the annotated corpus, took into account the whole set of annotation parameters (27 ELAN tiers) and assigned features. Files of occurrences of phenomena were created which often provided a demonstration of their realization, significantly deviating from commonly accepted options, the latter usually based on a limited set of data. Valuable use demonstrations were provided for phenomena such as the use of pronominal indices, negation, question and plural formation.

An example of how the linguistic model was constructed, is provided by the measurements output, which defined the options for plural formation in GSL. The vast majority of plural signs made use of classifiers to indicate plurality. The next most common option was to exploit location indices, where two-handed plural and repetition for plural formation (appreciated among the standard rule options) were left far beyond the top, followed only by the very rare occurrences of numeral and index based plural formations.

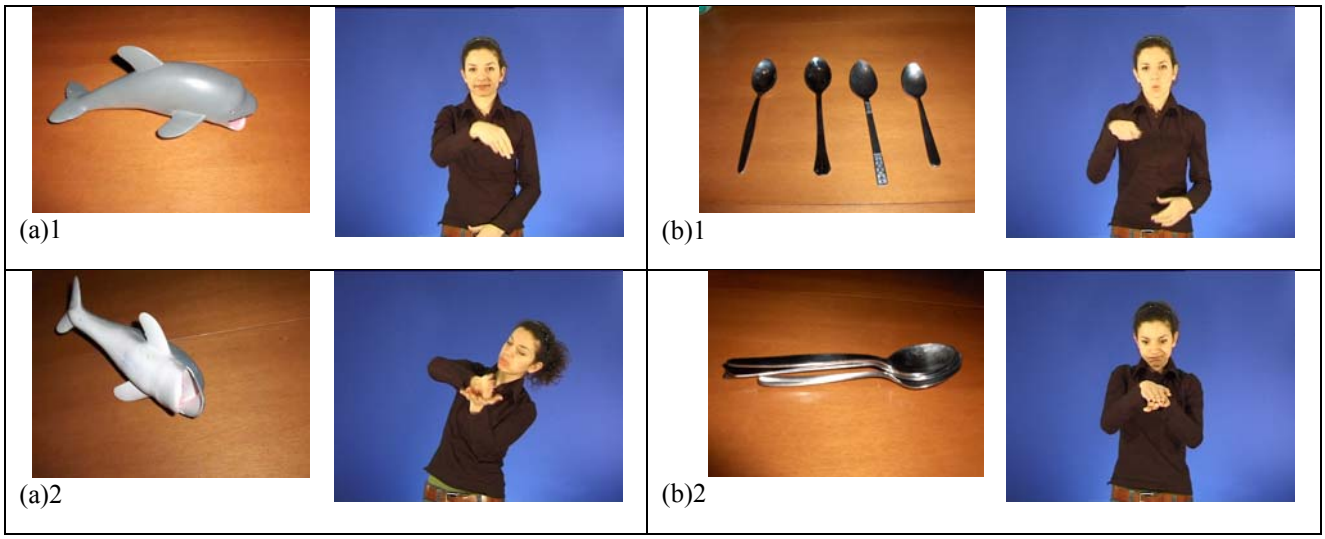

Figure 2: Icon driven classifier productions of GSL: (a) dolphin swimming (1), dolphin lying on flat surface (2); (b) spoons in a row (1), stacked spoons (2).

4.2 Linguistic model for GSL classifiers

A specific part of the elicitated corpus was devoted to the use of classifiers in GSL. In order to drive the informants to use a wide range of classifiers, different sets of stimuli were organised so as to cover the range of semantic properties assigned to base signs by use of appropriate classifiers. Elicitation focused on quantity, quality and spatial properties. The means to derive linguistic data were appropriate sets of icons, free discussion and story telling stimulated by film display.

The so derived data have been classified according to semantic indicator and are further elaborated in order to be incorporated in an educational environment as GSL grammar content. In Figure 2 it is demonstrated how icon driven classifier productions were derived. Example (a) demonstrates the use of flat B classifier to indicate the surface onto which a dolphin lies (2) opposite to the use of the sign for dolphin in the default case (1). Example (b) arranges spoons in a row repeatedly locating the handshape for spoon in the signing space (1), while in (2) a stack of spoons is indicated by a two hand formation of the flat B classifier.

5. Concluding remarks

The current state-of-the-art on technological advances and the open scientific issues related to sign language technologies have brought about the significance of annotated corpora for decoding the various aspects of sign language articulation message.

An appropriately annotated sign language corpus may provide a re-usable source of linguistic data to be exploited in the environment of sign language technologies but also in diverse situations as incorporation of SLs in various Natural Language Processing (NLP) environments or the creation of language teaching educational content. In this sense, an annotated corpus is essential to the development of sign recognition systems and also to the creation of adequate language resources such as lexical databases and electronic grammars needed in the context i.e. of Machine Translation. Language resources being equally crucial for the development of sign synthesis machines and conversion tools from spoken to sign language that often drive sign synthesis machines, underline the usability of a corpus which supports extraction of both reliable measurements and linguistic data.

GSLC design and implementation have equally focused on sign recognition support and on the extraction of a linguistic model for GSL. GSLC extensibility is intrinsically foreseen as regards both its content and adopted annotation features. This allows for corpus re-usability in linguistic research and sign language technology applications.

6. Acknowledgements

This work has been partially funded by the European Union, in the framework of the Greek national project DIANOEMA (GSRT, M3.3, id 35).

7. References

Atkins, S., Clear, J. & Ostler, N. (1991). Corpus design criteria. *Literary and Linguistic Computing*, Vol. 7 pp.1--16.

Bowden, R., Windridge, D., Kadir, T., Zisserman, A. & Brady, M. (2004). A Linguistic Feature Vector for the Visual Interpretation of Sign Language. In *Tomas Pajdla, Jiri Matas (Eds), Proc. 8th European Conference on Computer Vision, ECCV04. LNCS3022,* Springer-Verlag, Volume 1, pp. 391--401.

Bellugi, U. & Fischer, S. (1972). A comparison of Sign language and spoken language: rate and grammatical mechanisms. *Cognition: International Journal of*

Cognitive Psychology, 1, pp. 173--200.

Crasborn, O., Hanke, T. (2003). Metadata for sign language corpora. *Available on-line at*: http://www.let.ru.nl/sign-lang/echo/docs/ECHO_Metadata_SL.pdf.

DIANOEMA Project (2006a), Composition features of the GSL Corpus, WP1-GSL Corpus, *Technical Report* (in Greek).

DIANOEMA Project (2006b). Annotation Features of GSL, WP3– Language model / video annotations of GSL, *Technical Report* (in Greek).

Efthimiou, E., Sapountzaki, G., Karpouzis, C. & Fotinea, S-E. 2004. Developing an e-Learning platform for the Greek Sign Language. *Lecture Notes in Computer Science* 3118: pp. 1107--1113. Springer.

Efthimiou, E., Fotinea, S-E. & Sapountzaki, G. 2006. Processing linguistic data for GSL structure representation, In *Proceedings of the Workshop on the Representation and Processing of Sign Languages: Lexicographic matters and didactic scenarios, Satellite Workshop to LREC-2006 Conference*, May 28, pp. 49--54.

ELAN annotator, Max Planck Institute for Psycholinguistics, available at: http://www.mpi.nl/tools/elan.html

Fotinea, S-E., Efthimiou, E., Karpouzis, K. & Caridakis, G. 2005. Dynamic GSL synthesis to support access to e-content, In *Proceedings of the 3rd International Conference on Universal Access in Human-Computer Interaction (UAHCI 2005)*, 22-27 July 2005, Las Vegas, Nevada, USA.

HamNoSys Sign Language Notation System: www.sign-lang.uni-hamburg.de/projects/HamNoSys.html

Hanke, T. (2002). iLex - A tool for sign language lexicography and corpus analysis. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation. Las Palmas de Gran Canaria*, Spain. Paris : ELRA pp. 923--926

Karpouzis, K. Caridakis, G., Fotinea, S-E. & Efthimiou, E. 2007. Educational Resources and Implementation of a Greek Sign Language Synthesis Architecture, *Computers and Education*, Elsevier, Volume 49, Issue 1, August 2007, pp. 54--74, electronically available since Sept 05.

Kraiss, K.-F. (Ed.), 2006. Advanced Man-Machine Interaction - Fundamentals and Implementation. *Series: Signals and Communication Technology*, Springer.

Mayberry, R. (1993). First-Language acquisition after childhood differs from second language acquisition: The case of American Sign Language. *Journal of Speech and Hearing Research*, Vol. 36 , pp. 51--68

Mikros, G. (2004). Electronic corpora and terminology. In *Katsoyannou, M., and Efthimiou, E,, (eds) Terminology in Greek: Research and Implementation Issues*. Kastaniotis publications, Athens (in Greek).

Neidle, C (2002). SignStream™ Annotation: Conventions used for the American Sign Language Linguistic Research Project. Boston, MA: *American Sign Language Linguistic Research Project Report* No. 11, Boston University.

Prillwitz, S., Leven, R., Zienert, H., Hanke, T. and Henning, J. (1989). HamNoSys. Version 2.0. Hamburg Notation System for Sign Language. An Introductory Guide.

Sinclair, J. (1996). Preliminary recommendations on corpus typology. EAGLES Document EAG--TCWG--CTYP/P, electronically available at: http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html.

Stokoe, W. (1978). Sign Language Structure (Revised Ed.). *Silver Spring*, MD: Linstok

White, L. (1980). Grammatical Theory and Language Acquisition. *Indiana University Linguistic* Club.