Creation of a multipurpose sign language lexical resource: The GSL lexicon database

Athanasia-Lida Dimou[1], Theodore Goulas[1], Eleni Efthimiou[1], Stavroula-Evita Fotinea[1], Panayiotis Karioris[1], Michalis Pissaris[1], Dimitris Korakakis[1], Kiki Vasilaki[2]

[1]ILSP - R.C "Athena", [2]Aristotle University of Thessaloniki – Philology Department
[1]Artemidos 6 & Epidavrou, Maroussi, 15125 Athens, Greece
E-mail: {ndimou, tgoulas, eleni_e, evita, pkarior}@ilsp.gr, pissarakia@gmail.gr, korakakis79@gmail.com, kikivasilaki@yahoo.gr

Abstract

The GSL lexicon database is the first extensive database of Greek Sign Language (GSL) signs, created on the basis of knowledge derived from the linguistic analysis of natural signers' data. It incorporates a lemma list that currently includes approximately 6,000 entries and is intended to reach a total number of 10,000 entries within the next two years. The design of the database allows for classification of signs on the basis of their articulation features as regards both manual and non-manual elements. The adopted information management schema accompanying each entry provides for retrieval according to a variety of linguistic properties. In parallel, annotation of the full set of sign articulation features feeds more natural performance of synthetic signing engines and more effective treatment of sign language (SL) data in the framework of sign recognition and natural language processing.

Keywords: GSL, SL lexicon, manual feature, non-manual features, sign articulation, SL technologies, SL data acquisition

1. Introduction

Here we present the methodology followed in creating a multipurpose lexical data base of the Greek Sign Language (GSL) which currently incorporates approximately 6,000 sign entries and it is intended to reach a content of 10,000 entries in the next two years. The main effort is been placed on creation of an extensive resource of sign lemmas which may serve a variety of goals, including extraction of bilingual dictionaries/ glossaries, incorporation of lexical information in natural language processing (NLP) systems as in the case of machine translation (MT) from and into sign language, creation of training material for sign recognition technologies, and input to sign synthesis tools enabling signing by virtual signers (avatars).

Given the scope of the resource and the range of usability cases it is intended to serve, design criteria which had to be satisfied extend from naming conventions to coding of manual and non-manual elements of each sign for representation via synthetic signing and retrieval purposes.

The GSL lexicon database in its current status has been created by integrating two different available lexical resources after careful content evaluation and thorough revision of the previously available database structure design.

In the rest of the paper, we report on the methodological milestones and undertaken actions that the reported attempt required, as well as the procedures that are planned to be carried out next in order to extend the database content. In this framework, an initial study of available data has revealed considerable participation of non-manual features in GSL sign formation, while in many cases, non-manuals disambiguate the meaning of lemmas articulated by means of the same manual activity (see also Section 4 below). Thus, annotation of non-manual elements of signs becomes a central task in the current attempt, given the need to fully code articulation features of sign lemmas to equally support SL data computing and synthetic signing needs, parallel to the al use of the lexicon in communication and education context.

2. Exploited resources for the GSL lexicon database

The main resources used for the creation of the GSL lexicon data base derive from two different sources, i) the content of the bilingual (GSL-Modern Greek) multimedia dictionary NOEMA [1], and ii) the lemmatized GSL DICTA-SIGN[2] corpus. We provide next information on the structure of these two sources, which influenced the design of the GSL lexicon database.

2.1 Multimedia dictionary NOEMA

The NOEMA dictionary is the first electronic dictionary of GSL signs and contains 3,000 video lemmas of general language falling within the definition of basic lexicon content (Efthimiou & Katsoyannou, 2001). NOEMA is a bilingual dictionary which is aimed to provide structured knowledge of GSL lexicon to a large non-specialized audience. It is equally addressed to natural deaf GSL signers and to hearing individuals who are interested in learning GSL as a second language. Thus, the dictionary organization is intended to serve both groups of end users; to this end every sign has been categorized according to the thematic group it belongs to and is associated with a Greek translation, as well as synonyms and antonyms in

[1] http://www.ilsp.gr/en/services-products/products/item/item/2-noema

[2] http://www.dictasign.eu/

GSL. The NOEMA dictionary has actually been constructed as a tool to support an introductory course in GSL, providing paradigms of all handshapes recorded to be used by the language in basic vocabulary concepts.

One of the assets of NOEMA has been the search option in the dictionary content by means of a selected handshape or a combination of handshapes (Figure 1). The latter has been accomplished by annotation on the dictionary database for main as well as secondary handshape(s) used in sign formation for all its lemmas.

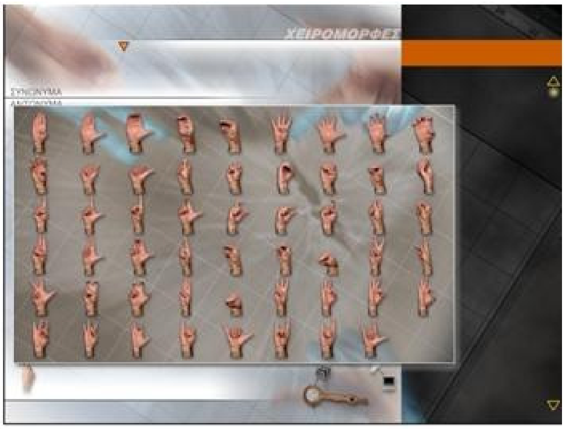


Figure 1: Interface for handshape based search option in the NOEMA dictionary

The video lemmas that comprise the NOEMA dictionary provided the content substructure of the GSL lexicon database; the 3,000 signs from the domain of general language constituted a significant core for the creation of the new lexicon. However, prior to transfer of the lemmas to the new database, a thorough evaluation study took place which pointed out a number of significant improvements needed to take place in order to optimize the data acquisition process, currently under development.

The list of enhancements in respect to the content available in NOEMA incorporates re-acquisition of lemmas by means of HD and Kinect cameras, corrections in lemma representation where necessary, addition of paradigms of use and coding of the manual and non-manual articulation elements of each sign.

As regards lemma correction, this involves two sets of corrections: i) while recording predicative lemmas any indication of declination (as to first person singular), which was often met with predicate GSL lemmas representation in NOEMA, is strictly avoided, and ii) a small number of lemmas which have been recognized to derive via interference from oral/written Greek but are not recognized as an integral part of the GSL vocabulary have been omitted from inclusion in the new vocabulary list.

All lemmas are acquired alongside with paradigms of use which aim at clarifying the represented concept and demonstrating all possible contexts of use of a specific lemma. Lemmatization of the utterances which serve as paradigms of use adds new lemmas to the initial lexicon which is significantly augmented via this process.

Another category of lemmas which is not transferred in the new database as it used to appear in NOEMA, involves association of classifiers with a specific equivalent lemma in Greek, as it has been i.e. the case of associating classifier C with the Greek lemma for PIPE. In the current framework, classifiers are treated as a class of entities associated with specific semantic properties and only those cases which are identified by native GSL signers as related to a specific concept without the need for associating their interpretation with information previously provided in their linguistic context, are treated as autonomous lemmas. Thus, in the currently adopted design, classifiers which have not been lexicalized are studied within their signed context and are treated in the lexicon either as bound morphemes or as semantic indicators with pronominal function.

2.2 Lemma extraction from an annotated corpus

Complementary to lemmas deriving from NOEMA, the GSL lexicon database has also been enriched by lemmas extracted from the annotated GSL segment of the Dicta-Sign corpus[3].

The corpus created during the Dicta-Sign project (Matthes et al., 2010; 2012) made available natural discourse productions in four SLs: Greek, German, French and English, to a significant extend fully annotated for the entailed lemmas in the ilex[4] (Hanke & Storz, 2008) annotation environment by means of the HamNoSys notation system (Hanke, 2004; Prillwitz et al., 1989). Lemma annotation of the Greek segment of the corpus enriched the GSL lexicon database with approximately 2,000 lemmas.

Creation of the Dicta-Sign corpus intended to elicit naturally produced signing, hence the elicitation procedures were carefully designed so as to promote naturalness of the acquired data. The outcome of the related data acquisition process was a corpus rich in continuous signing information markers, incorporating in-context lemma productions. In terms of the currently developed GSL lexicon, the sign lemmas deriving from the Dicta-Sign corpus need to be enhanced in respect to speed of production and co-articulation effects during the new acquisition process.

However, searching in the corpus for lemma extraction proved to be valuable for also providing a wide spectrum of use cases related to each lemma.

Furthermore, the study –currently in progress– on extraction and classification of the GSL classifiers system beyond the set of lexicalized classifier items referred to above, is heavily based on annotated data deriving from the same corpus.

3. Compilation of the GSL lemma list

In order to provide content to a common database, both

[3] http://www.sign-lang.uni-hamburg.de/dicta-sign/portal/lang_inform.html.

[4] www.sign-lang.uni-hamburg.de/ilex

sets of NOEMA and Dicta-Sign data had to be unified. The required lemmatization procedure ended with identification of 5,500 unique GSL lemma entries. The derived lemma list needed to be checked for corrections and undergo enhancements as indicated in 2.1 and 2.2 above, in order to ensure homogeneity during the new video recordings and the previously completed evaluation as to inclusion/exclusion criteria, applied to each lemma before its addition to the GSL lexicon database. Other decisions relate to the way compounds are treated depending on whether they are formed via combination of only free or free and bound morphemes, the provisions made with respect to GSL vs. oral Greek synonyms for the representation of a specific concept, and the coding of non-manual articulation features. Compounding has been decided to initially be addressed on the basis of a continuum from productive to lexical compounds approach (Liddell & Johnson, 1986), also adopbted by (Sandler & Lillo–Martin, 2006). Lemma corrections against intuitive GSL linguistic knowledge and selection of paradigms of use have been undertaken by two GSL natural signers, members of the development team supported by a team of three SL linguists.

Compilation of the “GLOSS” field of the database against a lemma list of Modern Greek revealed several one to many GSL to Greek alignments. Since within the scope of this lexicon is to provide for a wider semantic association of concepts and representations between GSL and Modern Greek, the need for the development of a linking mechanism that will enable proper lemma association in the two languages and will also effectively support lexical retrieval and sign language NLP applications has become obvious and related on-going experiments will be published in the next period.

4. Non-manual features

Work is SL linguistics has long recognised the importance of non-manual markers in the articulation of a sign. Non-manuals are considered to be an integral part of sign articulation when they participate along with manual activity in sign formation, and for this reason they have to be specified in the lexical entry of a sign (Pfau & Quer, 2010).

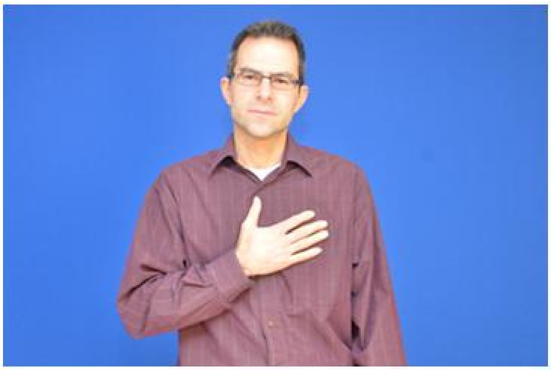


Figure 2: GSL sign LOVE –non-manual neutral articulation

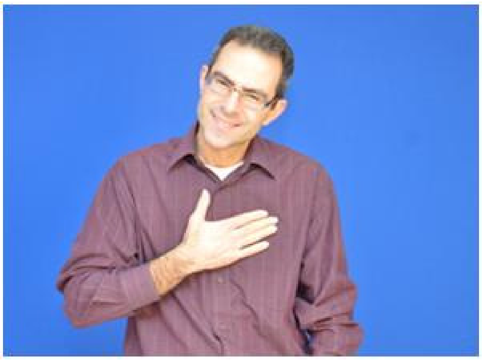


Figure 3: GSL sign THANK-YOU – head movement and facial features differentiate the signed concept from the flat, with respect to non-manuals sign LOVE

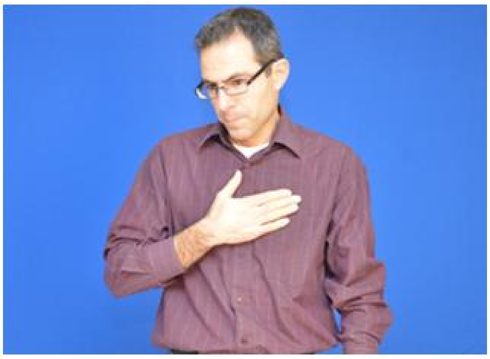


Figure 4: GSL sign MINE (1-Sg-Poss) – head movement and facial features identify the signed concept as differing from concepts THANK-YOU and LOVE

There are two kinds of non-manual markers: facial and non facial. Facial non-manuals occur entirely on the face, while non-facial markers take the form of a particular head or body movement (Neidle et al., 2000).

When they form part of sign phonology, there is a strong tendency for non-manual markers to be synchronized with the manual part of the sign. For example, in articulating the GSL signs HAPPY, SAME and GET-BORED the signs’ manual articulation is obligatorily accompanied by a particular facial expression performed in parallel. Moreover, non-manual markers in GSL may distinguish two (or more) otherwise identical signs, i.e. they can define minimal pairs. For instance, the signs LOVE, THANK-YOU and the first person singular of the possessive pronoun (MINE) are all identified by the different non manual signals accompanying the same hand activity as indicated in Figures 2, 3 and 4. Similarly, pairs of signs are very often distinguished by non-manual articulation elements, like the signs BE-CRAZY ABOUT and COMMIT SUICIDE which are minimally distinguished by facial expression.

Non-manual features are systematically addressed in respect to the lexicon under development as according to the design specifications of the GSL lexicon database. They are dedicated a separate section in which the presence or absence of facial and body features are annotated and accordingly demonstrate critical alternations in the meaning of a manually signed or a classifier entity (Efthimiou et al., 2010).

Furthermore, coding of signs in respect to non-manuals is ranked as equally important for synthetic signing as manual features coding. Incorporation of non-manuals is directly related to the degree of achieved naturalness and related acceptance of synthetic signing by Deaf communities in general. In our case, it is a prerequisite for exploiting the reported resource in teaching and communication environments which consume language technologies.

Enrichment of lemmas with annotations for both manual and non-manual features is facilitated by a dedicated section in the Sis-Builder[5] tool (Goulas et al., 2010).

For the facilitation of assignment of HamNoSys notation symbols to manual activity involved in formation of a specific sign, the environment provides virtual keyboards for the marking of symmetries, handshape, hand position, hand location and motion actions, partly shown in Figure 5.

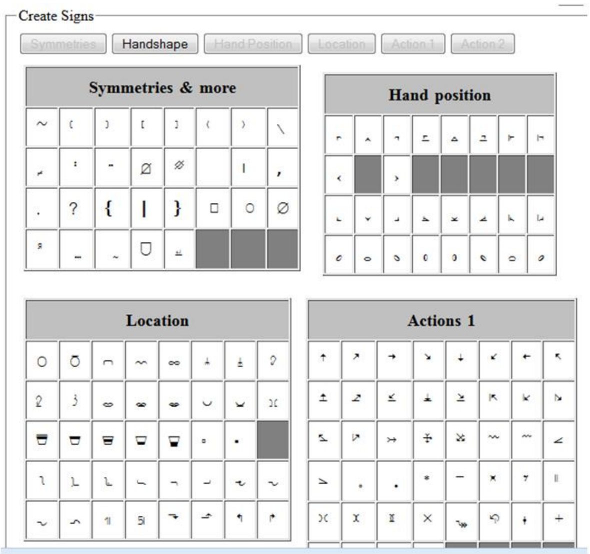
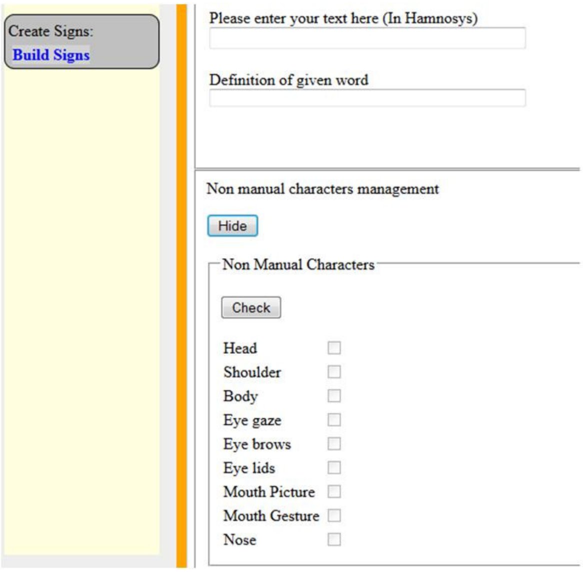


Figure 5: Virtual keyboards for the annotation of manual activity in sign formation

Non manual elements of sign formation are coded in SiS-Builder by selection from a drop-down menu of values for all possible facial and body features which participate in sign articulation parallel to manual activity. Figures 6a and 6b depict the set of non manual features taken into account for coding and the way coding takes place via selection from the available drop-down menus.

Annotation of signs in respect to both their manual and non manual articulation parameters (Figure 7) provides the necessary information for their more natural synthetic representation. In fact, this information is crucial for a range of applications in the area of SL processing, focussing on improvement of retrieval and sign recognition results. Nonetheless, completeness in representation of articulation features of signs is also crucial in SL linguistics research and SL learning environments equally in the framework of treating SL as first or second language.



Create Signs:
Build Signs

Please enter your text here (In Hamnosys)

Definition of given word

Non manual characters management

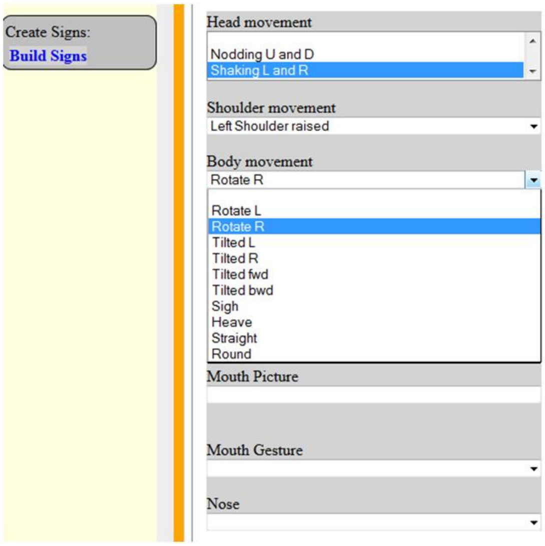
Hide

Non Manual Characters

Check

Head ☐
Shoulder ☐
Body ☐
Eye gaze ☐
Eye brows ☐
Eye lids ☐
Mouth Picture ☐
Mouth Gesture ☐
Nose ☐

Figure 6a: The set of non manual features handled by the SiS-Builder environment



Create Signs:
Build Signs

Head movement
Nodding U and D
Shaking L and R

Shoulder movement
Left Shoulder raised

Body movement
Rotate R

Rotate L
Rotate R
Tilted L
Tilted R
Tilted fwd
Tilted bwd
Sigh
Heave
Straight
Round

Mouth Picture

Mouth Gesture

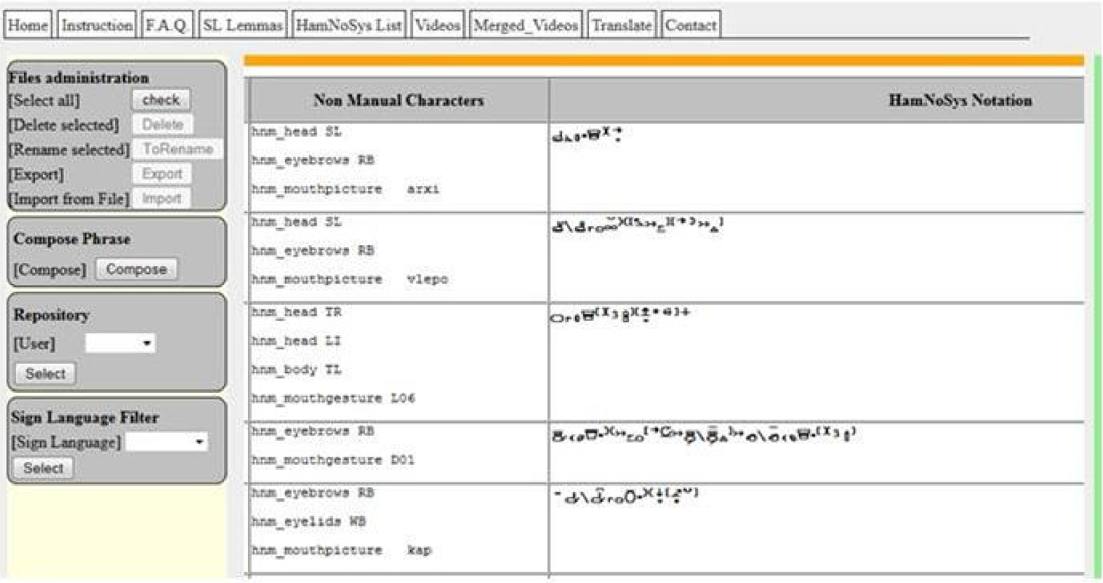
Nose

Figure 6b: Non manual feature values assignment via drop-down selection

5. Data acquisition methodology and set-up

Recording sessions follow a predefined script which includes the lemmas to be acquired each time along with the set of usage examples accompanying each lexical entry, which are selected on the basis of linguistic-lexicographic criteria to satisfy demonstration of semantic/syntactic properties of the lemmas.

[5] http://speech.ilsp.gr/sl



Non Manual Characters	HamNoSys Notation
hnm_head SL hnm_eyebrows RB hnm_mouthpicture arxi	[illegible]
hnm_head SL hnm_eyebrows RB hnm_mouthpicture vlepo	[illegible]
hnm_head TR hnm_head LI hnm_body TL hnm_mouthgesture L06	[illegible]
hnm_eyebrows RB hnm_mouthgesture D01	[illegible]
hnm_eyebrows RB hnm_eyelids WB hnm_mouthpicture kap	[illegible]

Figure 7: Manual and non-manual activity annotation on sign lemmas of GSL in the SiS-Builder environment

The data acquisition team is composed of the engineer who controls data flow from the acquisition devices, the studio officer who is in charge of the studio set-up and cameras control, an interpreter/facilitator who supports the informant, and a native signer who performs the scheduled lemmas and their paradigm of use phrases in three (3) repetitions each. Prior to recording, the team members need to study the lemmas to be captured and decide on their representative paradigms of use, if such paradigms are not already available in the GSL corpus. During capturing, the predefined list of lemmas which falls within the session's schedule is projected to the informant by means of a monitor.

The examples which are associated with each lemma are noted down in a note taking environment in the form of "written GSL", completely avoiding the use of subtitles in Greek language, in order to provide an easy to check list of all signs that are contained in the usage examples and also diminish oral language interference effects in the grammar of the paradigm utterances. Lemmatization of the newly produced paradigm of use utterances is intended to ensure that all signs used in the example phrases are also incorporated in the lemma list, thus using this qualitative control also as a means of augmenting the lexicon with new lemmas.

GSL lemmas are realized in isolations, in a clear, comprehensible manner. Examples of use are preferably small, simple phrases that demonstrate each sign's proper linguistic use. Examples need not be performed flat (in a dry manner), although non-manual markers that are related with a high degree of emotion demonstration on sentence level are advised to be left out for avoidance of confusion as to the proper sign articulation.

Recordings take place at a high-end technology studio (Figure 8) that provides all necessary facilities (lighting, storage media, microphones, cameras) for HD quality recording. In terms of data acquisition equipment, one HD camera (front view) and one Kinect camera (for depth information of sign articulation) are used. The synchronisation of these media is accomplished via clapping[6] as audio cue and flashing as visual cue.

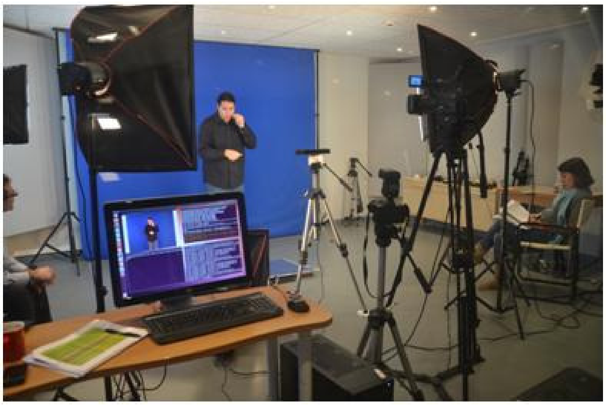


Figure 8: Lexicon acquisition studio set-up

6. Conclusion

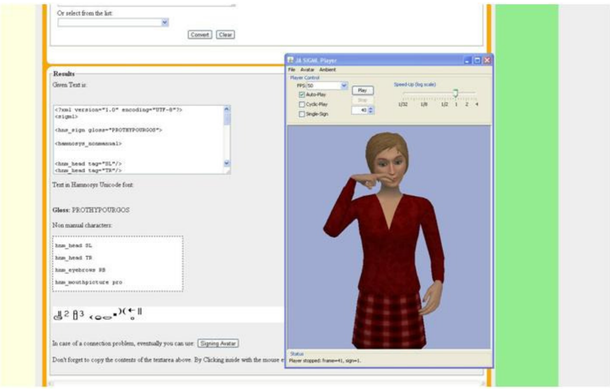
The GSL lexicon database is an ambitious project, opting for the creation of a multipurpose resource of at least 10,000 distinctive GSL lemma entries, mainly addressing SL processing needs in the framework of human language technologies applications and also in service of SL recognition and synthetic signing technologies. Thus, exhaustive coding of lemmas for their manual and non-manual features is a major task. In this context, association of lemmas within an appropriate ontology scheme is required to enable more efficient bilingual associations between GSL and Modern Greek, which will significantly augment accessibility of written Greek texts

[6] Microphones are typically used in multicamera data acquisition to capture clapping signals which are exploited in synchronization of the different video streams.

by Deaf individuals allowing for more effective language engineering solutions in a variety of communicative environments. These involve machine translation, meaning spotting and retrieval of information from a written text source, facilitation of visual processing and SL synthesis, the goal being to achieve proper linking of a sign with an equivalent word in Modern Greek, but also with all its available synonyms and the range of related hypernyms and/or hyponyms.

In parallel, systematic categorization of non-manual features of sign articulation is expected to lead to a more concrete definition of the linguistic function of non-manuals in GSL sign formation, as well as to higher acceptance of synthetic signing (Jennings et al., 2010), since sign synthesis engines which take non-manuals into account improve significantly in respect to naturalness of signing performance (Figure 9).

Finally, a resource providing the qualitative and quantitative range of information incorporated in the GSL lexicon, will be of value also to GSL language education both in respect to first and second language learning.



Results

Given Text is:

```
<?xml version="1.0" encoding="UTF-8"?>
<sigml>

<hns_sign gloss="PROTHYPOURGOS">

<hamnosys_nonmanual>


<hnm_head tag="SL"/>
<hnm_head tag="TR"/>
```

Text in Hamnosys Unicode font:

Gloss: PROTHYPOURGOS

Non manual characters:

```
hnm_head SL
hnm_head TR
hnm_eyebrows RB
hnm_mouthpicture pro
```

In case of a connection problem, eventually you can use:

Don't forget to copy the contents of the textarea above. By Clicking inside with the mouse e

FPS: 50

☑ Auto-Play

□ Cyclic-Play

□ Single-Sign

40

Status

Player stopped: frame=41, sign=1.

Figure 9: Avatar performance incorporating non manuals as annotated in the SiS-Builder environment for the GSL lemma PRIMEMINISTER

7. Acknowledgements

The research leading to these results has received funding from POLYTROPON project (KRIPIS-GSRT, MIS: 448306), based on initial data acquisition in the framework of the Dicta-Sign project (FP7/2007-2013 grant agreement n° 231135).

8. References

Efthimiou E., Fotinea S-E., Dimou A-L., Kalimeris C. (2010). Towards decoding Classifier function in GSL. In Dreuw, P. et al. (Eds.), LREC 2010, *Proceedings of 4th Workshop on Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, Valetta, Malta.

Efthimiou, E. & Katsoyannou, M. (2001). Research issues on GSL: a study of vocabulary and lexicon creation. In *Studies in Greek Linguistics, Computational Linguistics* 2: 42-50 (in Greek).

Goulas, T., Fotinea, S-E., Efthimiou, E. and Pissaris, M. (2010). SiS-Builder: A Sign Synthesis Support Tool. In Dreuw, P. et al. (Eds.), LREC-2010, *Proceedings of 4th Workshop on Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pp. 102--105.

Hanke,T. (2004). HamNoSys - representing sign language data in language resources and language processing contexts. In O. Streiter and C. Vettori (Eds.), LREC-2004. *Proceedings of 1st Workshop on Representing and Processing of Sign Languages,* pp. 1--6.

Hanke, T. and Storz, J. (2008). iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In *Construction and Exploitation of Sign Language Corpora. Proceedings of 3rd Workshop on the Representation and Processing of Sign Languages*. ELRA, Paris, pp. 64--67.

Jennings, V., Elliott, R., Kennaway, R. and Glauert, J. (2010). Requirements for a Signing Avatar. In Dreuw, P. et al. (Eds.), LREC 2010, *Proceedings of 4th Workshop on Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, Valetta, Malta, pp. 133--136.

Klima, E., and Bellugi, U., (1979). *The signs of language*. Harvard University Press, USA.

Liddell, S. and Johnson, R., (1986). American Sign Language Compound Formation Processes and Phonological Remnants. In *Natural Language and Linguistic Theory,* vol.4, Reideil Publishing Co, pp. 445--513.

Matthes S., Hanke T., Regen A., Storz J., Worseck S., Efthimiou E., Dimou A.-L., Braffort A., Glauert J. and Safar. E. (2012). Dicta-Sign – Building a Multilingual Sign Language Corpus. In Crasborn et al. (Eds.), LREC 2012, *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, Istanbul, Turkey.

Matthes S., Hanke T., Storz J., Efthimiou E., Dimou A-L, Karioris P., Braffort A., Choisier A., Pelhate J., Safar E. (2010). Elicitation tasks and materials designed for Dicta-Sign's multi-lingual corpus. In Dreuw, P. et al. (Eds.), LREC 2010, *Proceedings of 4th Workshop on Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, Valetta, Malta.

Neidle, C., Kegl, J., MacLaughlin, D., Bahan, B., et Lee, R.G. (2000). *The Syntax of American Sign Language. Functional Categories and Hierarchical Structure*. Cambridge, MA: MIT Press.

Pfau, R., & Josep, Q. (2010). Nonmanuals: their grammatical and prosodic roles. Sign Languages, In D. Brentari (Ed), pp. 381--402. Cambridge: Cambridge University Press.

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*. Signum Verlag, Hamburg.

Sandler, W. & Lillo – Martin, D. (2006) "*Sign Language and Linguistic Universals*", Cambridge University Press, UK, pp.72.