From a Sign Lexical Database to an SL Golden Corpus – the POLYTROPON SL Resource

Efthimiou, E.[1], Fotinea, S-E.[1], Dimou, A-L.[1], Goulas, T.[1], Karioris, P.[1], Vasilaki, K.[2], Vacalopoulou, A.[1], Pissaris, M.[1], Korakakis, D.[1],

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

Abstract

The POLYTROPON lexicon resource is being created in an attempt i) to gather and recapture already available lexical resources of Greek Sign Language (GSL) in an up-to-date homogeneous manner, ii) to enrich these resources with new lemmas, and iii) to end up with a multipurpose-multiuse resource which can be equally exploited in end user oriented educational/communication services and in supporting various SL technologies. The database that hosts the newly acquired resource, incorporates various SL oriented fields of information, including information on compounding, GSL synonyms, classifier qualities, lemma related senses, semantic groupings etc, and also lemma coding for their manual and non-manual articulation activity. It also provides linking of GSL and Modern Greek equivalent(s) lemma pairs to serve bilingual use purposes. A by-product of considerable value is the parallel corpus which derived from the GSL examples of use accompanying each lemma entry in the dictionary and their translations into Modern Greek. The annotation of the corpus for the entailed signs and assignment of respective glosses in combination with data capturing by both HD and Kinect cameras in three repetitions, allowed for the creation of a golden parallel corpus available to the community of SL technologies for experimentation with various approaches to SL recognition, MT and information retrieval.

Keywords: SL data acquisition, SL lexicon resource, GSL-Greek bilingual dictionary, deaf accessibility services, SL technologies, SL-text parallel golden corpus

1. Introduction

In the framework of research activities undertaken within the POLYTROPON project[1], significant effort has been placed in maintaining and extending a Greek Sign Language (GSL) lexicon dataset which consisted of lemmas captured by means of diverse capturing devices, lemma list construction methodologies and approaches for verification of acceptance by the local deaf community, covering a time space of approximately fifteen years of acquisition phases.

In (Dimou et al., 2014), the rational and methodological principles for revisiting and recapturing the existing GSL lexicon resource have been justified, similarly to the goal of extending the already created lemma list and the adopted scheme of fields of information.

As regards the content, among the crucial issues that had to be faced was the verification of acceptance by the general deaf public of the sign content, as well as the presentation protocol regarding the different sign categories entailed in the lexicon (i.e. adaptation of the pronoun neutral predicate form, classification of classifier based lemmas according to the classifier generating them, and handling of compound sign lemmas as well as special expressions included in the lemma list). In parallel, a number of decisions on data acquisition methodology were related to the purpose of creating a multiuse resource.

To sum up, the POLYTROPON resource is being created in an attempt: i) to gather and recapture already available lexical resources of Greek Sign Language (GSL) in an up-to-date homogeneous manner, ii) to enrich these resources with new lemmas, and iii) to end up with a multipurpose-multiuse resource which is equally exploitable in end user oriented educational/ communication services and in supporting various SL technologies, including information extraction, Web accessibility tools, incorporation of lexical information in natural language processing (NLP) systems for sign language processing 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), along with simpler tasks which are depending on availability of language resources such as creation of bilingual dictionaries and glossaries.

We refer next to the various aspects of the POLYTROPON resource acquisition process and its current implementation.

2. Content Definition and Acquisition Methodology

The main sources for the POLYTROPON lexicon content are two pre-existing GSL lexicon databases: i) the lemma list of the bilingual (GSL-Modern Greek) multimedia dictionary NOEMA[2] (set to circulation in the form of DVD-ROM in 2001), and ii) the list deriving from the lemmatized GSL segment of the Dicta-Sign corpus[3].

However, both these lemma lists were not adopted in full in the new database. While on technical grounds, the rationale behind recapturing the data was based on the two following facts:

i) a considerable bulk of the available resources was

[1]http://www.ilsp.gr/el/infoprojects/meta?view=project&task=show&id=198

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

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

captured in the late ’90s, thus being subject to capturing devices limitations of that time &

ii) the DICTA-SIGN lemmas bulk was extracted from a corpus annotation procedure as a result of glosses annotation, meaning that the extracted lemmas were articulated in the context of a phrase, thus being formed away from the typical lemma presentation scheme adopted in a dictionary

There have been recognized a number of serious issues related to language research and language data acquisition methodology, that have dictated the revision of the available lexicon content and the adoption of an acquisition methodology which would exclude any interference from the oral language environment.

2.1 Revision of pre-existing Lexicon Resources and Lemma List Enrichment

As already mentioned, the out-of-date video quality or the lemma in context articulation, were only partially reasons for recapturing the GSL lemmas. Planning of the new capturing procedure provided the opportunity for an in-depth evaluation of existing lemmas against SL linguistic criteria and the rethinking of lemma lists formation. These criteria included parameters such as whether the items already classified as lemmas were real lexical items or classifier constructions to express a concept imported from the environment spoken language, decisions about how sign lemmas have to be presented in the lexicon data base (i.e. the pronoun neutral predicate representation against signing the first person singular pronoun when forming a predicative sign, which is directly influenced from Modern Greek (MG) that lacks a morphology neutral form such as the Gerund form of English, or organizing treatment and presentation of compound lemmas in line with (Liddell & Johnson, 1986) and (Sandler & Lillo–Martin, 2006)), but also critical, whether all already captured lemmas provided formations connected with specific concepts and recognized as such by the (majority of) GSL signing community or were ad hoc formations improvised by informants of the early capturing.

Although such lemma formations were of limited number, this latter case of “mistakes” or “unknown” lemmas turned to be a source error factor in the early NOEMA dataset, which was only noticed through the actual use of the dictionary and active communication with the native GSL signers’ community. Serious consideration has been dedicated to the definition of what kind of new signed data should be included in the lemma list, the main issue being to guarantee that SL grammar principles are met and also the sign(s) representing a given concept are widely accepted by the GSL community. As a consequence of the above, in a first phase the 3.000 video lemmas of general language domain falling within the definition of basic lexicon content (Efthimiou & Katsoyannou, 2001), which formed the lemma list of the NOEMA dictionary, were thoroughly revisited in order to identify needed enhancements or corrections both in respect to content formal representation issues and wide acceptance of the video lemmas. As a result of this work, the NOEMA lemma list has been filtered in respect to GSL wise “peculiar” content and sign forms not widely accepted have either been removed or replaced by more appropriate ones. Similarly, all not appropriately performed entries in respect to morphological markers have been spotted and received a commentary to guide their proper acquisition during the new capturing sessions. In the same line of evaluation, the lemmas extracted from the GSL Dicta-Sign corpus have also been filtered for sense representation in the corpus and sense disambiguation, as well as against all check parameters holding for the NOEMA lemma list. The merge of the two originally available lists formed the initial content of the POLYTROPON lexicon, while GSL synonyms and antonyms of the entailed lemmas provided the first round of lemma list expansion.

Figure. 1: GSL synonyms linked with one sense and one corresponding lemma in Greek.

The structure of information accompanying sign lemmas in the database has partially adopted the scheme followed in the NOEMA database, enriched with information fields which serve the purposes of the POLYTROPON resource. Thus, each sign lemma is associated with one or more equivalents in MG, a gloss, its GSL synonym(s) if any (Fig. 1), and one semantic sense, where possibly different senses of a single form are disambiguated via linking with different entries in the lexicon database (Fig. 2) and different examples of use. The database incorporates a number of further fields for information which become visible only when a specific application needs to exploit this kind of information. Among the information inserted in the database is the category of “special” or “fixed expressions” in both GSL and MG (Fig. 3). Participation of individual lemmas in the formation of special expressions has become visible, in order to allow for the retrieval of such expressions as a result of searching by means of the lemmas involved in their construction. Here the main issue is that lemmas in principle lose their initial sense when they appear in a special expression, which usually expresses a metaphor.

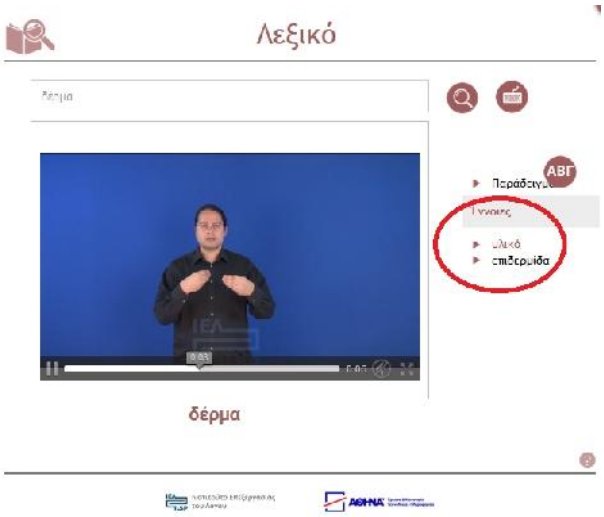

Figure 2: Multiple senses of a single form are visible in the lexicon database.

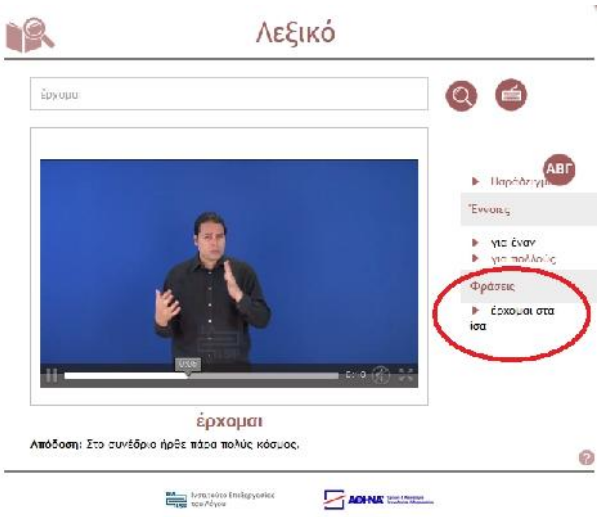

Figure 3: Special expressions –either of GSL or Greek origin– linked with entailed lemmas for easy search.

Furthermore, a number of information fields not visible to external users such as the HamNoSys[4] coding of sign lemmas, feature coding for the non-manual activity involved in sign formation (Pfau & Quer, 2010), GSL grammar markers for i.e. classifier constructions, plural formation and compounding are also available.

Especially in respect to classifier constructions, only those items which are identified by native GSL signers as lexicalized forms representing specific concepts without the need for associating their interpretation with information previously provided in their linguistic context, are treated as autonomous lemmas. Thus, in the adopted lexicon design, classifiers which have not been lexicalized are classified within their signed context and are treated in the lexicon either as bound morphemes adding semantic values or as semantic indicators with pronominal function.

However, the decision to include paradigms of use for each sign lemma (Fig. 4) has proved to be a turning point in the resource development. Since acceptance of the signs has been a priority issue, a significant criterion for the identification of generally accepted sign lemmas was their association with examples of use that would be indicative natural signers' productions clarifying the use of the examined lemma in context.

In order to best serve this goal, the addition of examples to lemma related information was organised as a corpus acquisition task following the good practice developed within the Dicta-Sign project for SL data acquisition (Matthes et al., 2010; 2012), taking measures to eliminate interference from the spoken or written forms of Modern Greek (MG) to the wider possible extend.

Fulfilment of this task was planned and executed by a group of experts where the major presupposition was the strict use of GSL in group discussions. The working group consisted of six signers, including native GSL signers, GSL codas and SL linguists. The native GSL signers of the team were the persons who had reviewed the NOEMA and the GSL Dicta-Sign lemma lists. Discussion of each lemma was based on meaning (=concept) representation, while the examples of use were decided after discussion among the group members while probing the best candidate phrases for each lemma. In the finalization phase and prior to each recording session, lemmas and their examples to be acquired next, were collected and archived in recording material dedicated sessions, where the content of each recording was fixed.

The new acquisition of lemmas and their examples of use was performed by means of one HD and one Kinect camera, while the whole of material was acquired in three repetitions of each item, in order to create a resource appropriate to be exploited in sign recognition.

The recorded example phrases are annotated for the included lemmas and translated into MG, while each lemma is assigned a gloss. This procedure allowed for checks and a remedy of possible inconsistencies with respect to the lemma list, thus ensuring that all signs used in the example of use phrases find an equivalent lemma in the lemma list of the lexicon. Through this procedure, the example phrases provided a further source for lemma enrichment, since a general convention has been that all signs used in the example phrases, need to be searchable and retrieved in the lemma list.

Given that the lexicon entries are constantly enriched, the POLYTROPON lexicon resource has become an expanded database, which is relatively difficult to check for possible mismatches and omissions. In order to facilitate cross-checks and also provide a tool for lexicon inspection by end users who are not necessarily familiar with database structure, a simple interface has provided visualization of various pieces of information related with each lemma as depicted in Figures 1, 2, 3 and 4.

In order to increase the visibility of the POLYTROPON resource by making it known to those interested in using parts of it in SL research and education, a bilingual dictionary for the language pair GSL-MG is already

[4] HamNoSys: the Hamburg Notation System (Hanke, 2004; Prillwitz et al., 1989).

extracted from the lexicon database while the resource is documented for its content and metadata within the *clarin:el* repository, the Greek sector of CLARIN[5], the European infrastructure for language resources and technology. The dictionary content has become available free of charge but subject to Creative Commons (CC) licensing[6].
For its identification in clarin:el, the POLYTROPON resource has received the persistent identifier (PID): http://hdl.grnet.gr/11500/ATHENA-0000-0000-42D5-5 (Fig. 7).

2.2 The POLYTROPN Parallel Corpus

The GSL phrases captured to serve as examples of use of lemmas in context along with their translations to MG constituted a parallel corpus of considerable length and richness, which is available in HD and Kinect for Linux captures in three repetitions for each signed utterance.

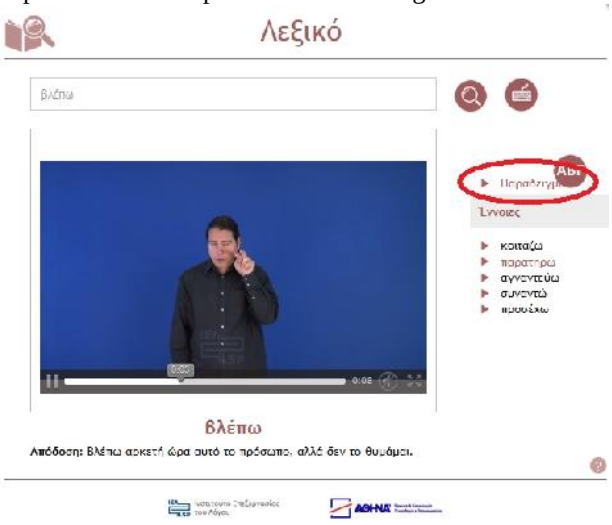

Figure 4: Example of use in GSL connected with its translation in Greek.

As data acquisition progressed, it became obvious that the lexicon database could be further exploited as an independent resource to serve technologies which crucially depend on a "golden" (parallel) corpus such as sign recognition, information retrieval directly from SL video and machine translation. To serve this goal, the POLYTROPON parallel corpus is being annotated in the iLex[7] (Hanke & Storz, 2008) environment. Annotation tiers provide further information on lemma timestamps, glosses, HamNoSys coding for manual activity, non manual features on the sign and utterance level to indicate morpho-phonological, semantic and/or syntactic functions, and also classifier related information. Upon completion of the annotation work, the corpus, which currently entails 3.400 signed utterances, will become available to the research community for experimentation, via the CLARIN infrastructure.

3. The POLYTROPON Database: a Resource for Sign Language Technologies

The parallel corpus created as reported above, is currently exploited in testing an approach to machine translation (MT), while the bilingual MG-GSL dictionary that has been produced as a by-product of the GSL lexicon database structure, -incorporating approximately 10,000 entries at the time of writing- has already been adopted in the official educational content platform of the Greek Ministry of Education to support accessibility of written content by Deaf end users, while a subset of it is incorporated in *e-class* the Greek platform for University level curriculum content, to support accessibility of educational information by deaf students.
Similarly, the HamNoSys coded information of the database is exploited in a simple interface for dynamic SL phrase formation to be used by both L1 and L2 users, while a Web based text accessibility tool is also supported by the POLYTROPON lexicon database.
These lexicon based technologies also exploit a suite of written language technologies including a lemmatizer and a morphological analyzer for MG, necessary to correctly identify and link the various tokens relating to a specific lemma in MG texts.
Such tools form the necessary background to allow successful bilingual connections and search retrieval results in the database underlying the interfaces to be presented next (Efthimiou et al., 2015).

3.1 The POLYTROPN Lexicon: a Resource for Synthetic Signing

Phonological coding of the POLYTROPON lemma list has enabled the development of a simple interface for dynamic synthetic signing, which can be equally used by GSL knowledgeable and non-knowledgeable end users in order to facilitate communication via GSL language productions.
A search box allows retrieval of constituents to compose the wished utterance. Phrase components may be reordered via drag-and-drop actions, while the users not familiar with GSL can advice template based instructions for structuring the phrases they create, according to GSL grammar. Special provision is taken for lexical items not included in the lexicon as well as for proper nouns to be visualized by means of fingerspelling.
The database copy to serve synthetic signing is hosted in the Cloud, while the related interface is currently attached to the set of deaf accessibility tools incorporated in the "Photodentro" platform that hosts the official educational content of primary and secondary levels of the Greek educational system (Fig. 5) [8].

[5] CLARIN: Common Language Resources and Technology Infrastructure (www.clarin.eu).
[6] https://creativecommons.org.
[7] www.sign-lang.uni-hamburg.de/ilex.

[8] The experimental implementation of all integrated tools to "fotodendro" can be reached for experimentation via: sign.ilsp.gr/jas/dev/demo.html.

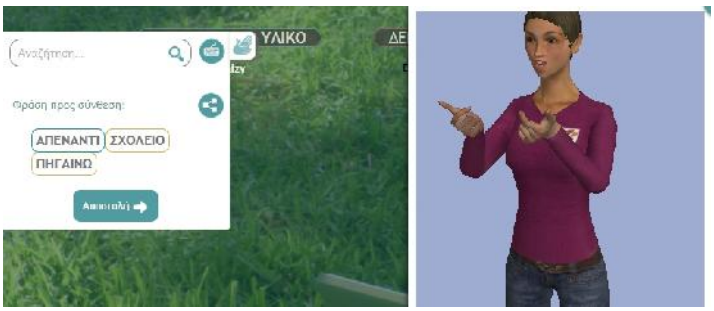

Figure 5: Input to synthetic signing.

Figure 6: Lexicon content linked with Web text accessibility tool.

3.2 The POLYTROPN Bilingual Dictionary Applications

Current on-line bilingual dictionaries based on material from the lexicon database are incorporated in two educational content platforms as deaf accessibility aids. They incorporate language technology tools which facilitate query entry and result retrieval, irrespective of the morphological complexity of the token form of MG used as the source for search (Fig. 8).

In both environments, deaf users may also prefer to insert search items by means of a virtual keyboard for fingerspelling as an alternative input device.

3.3 The POLYTROPN Lexicon as an Accessibility Tool for WEB Content

An especially well accepted application of the POLYTROPON resource in the educational context, is the direct linking of unknown words found in a text in the Web with their GSL equivalents. When activated, this option allows the user to view the GSL sign linked to a given word in a text by simply double clicking on the unknown item as depicted in Fig. 6.

Again language technology tools that run in the background enable retrieval of the proper pair in the lexicon resource irrespective of the morphological form of the search item in the text.

4. Conclusion

The POLYTROPON lexicon database has been created to mainly address SL processing needs in the framework of human language technologies applications and also in service of SL technologies with focus on sign recognition and synthetic signing. 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 of video-lemmas to coding of manual and non-manual elements of each sign for representation via synthetic signing and retrieval purposes. Within a time span of three years, the database has become the richest resource for GSL lexicographic data, its enrichment being steadily in progress. The so far acquired data are already exploited in a number of Web based applications supporting deaf education and communication needs. However, the collection of the resource has also triggered new challenges on technological and SL linguistic grounds. 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 by Deaf individuals in a variety of communication environments. Furthermore, the content of the resource allows for experimentation of new approaches in the framework of the standard and new SL technologies including SL recognition, dynamic synthetic signing, machine translation and information retrieval from video sources.

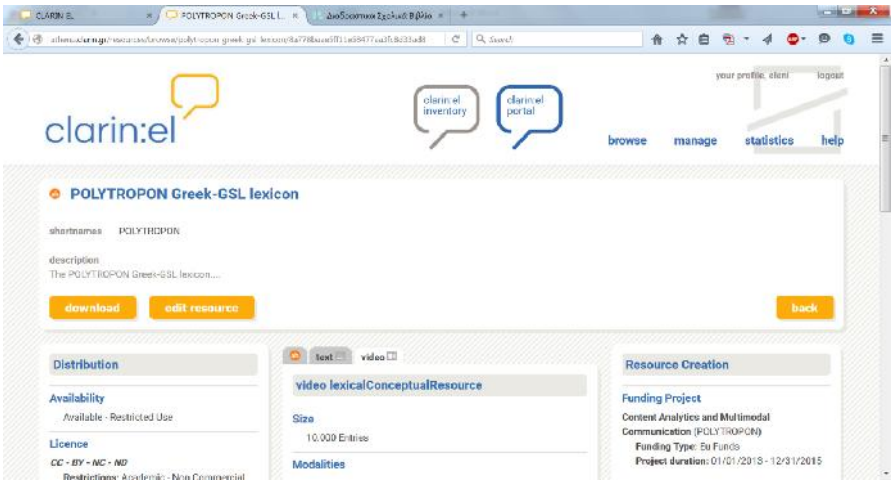

Figure 7: The POLYTROPON bilingual dictionary in the *clarin:el* repository.

In this framework, a new approach to corpus mining which is planned to be soon exploited on the basis of features relating to various parameters of sign articulation, classifier identification and features assigned to phrases as identifiers of sentence level properties, makes use of neural networks in combination to standard computer vision techniques already researched in the scope of SL technologies. Furthermore, the goal of the acquisition team is to provide the research community with a release of a golden corpus for machine learning in the areas of SL corpus mining and machine translation.

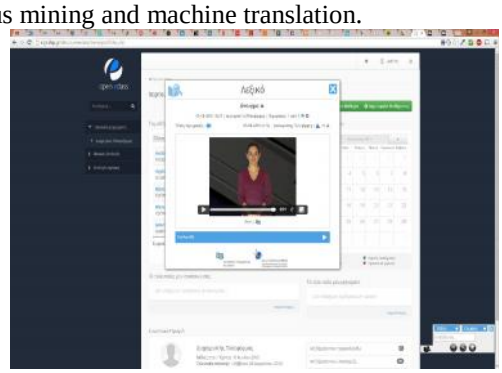

Figure 8: POLYTROPON resource use in *e-class* platform.

5. Acknowledgements

These research results have received funding from POLYTROPON project (KRIPIS-GSRT, MIS: 448306).

6. References

Dimou, A-L., Goulas, T., Efthimiou, E., Fotinea, S-E., Karioris, P., Pissaris, M., Korakakis, D., Vasilaki, K. (2014). Creation of a multipurpose sign language lexical resource: The GSL lexicon database. Proc. of *6th Workshop on the Representation and Processing of Sign Languages*, LREC 2014, Reykjavik, Iceland, pp. 37-42.

Efthimiou, E., Fotinea, S-E., Goulas, T., Kakoulidis, P. (2015). User friendly Interfaces for Sign Retrieval and Sign Synthesis. In M. Antona, M. & C. Stephanidis (Eds.). Proc. of *9th International Conference, UAHCI 2015, Part II*, held as Part of HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, LNCS vol. 9176, pp. 351-361, Springer, Heidelberg.

Efthimiou, E., Dimou, A-L., Fotinea, S-E., Goulas, T., Pissaris, M. (2014). SiS-builder: A Tool to Support Sign Synthesis. Proc. of *2nd Int'l Conference on the Use of New Technologies for Inclusive Learning*, pp. 26–36. York, UK.

Efthimiou, E. & Katsoyannou, M. (2001). Research issues on GSL: a study of vocabulary and lexicon creation. *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.), Proc. of *4th Workshop on Representation and Processing of Sign Languages*, LREC-2010, pp. 102-105.

Hanke,T. (2004). HamNoSys - representing sign language data in language resources and language processing contexts. Proc. of *1st Workshop on Representing and Processing of Sign Languages*, LREC-2004, pp. 1-6.

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

Klima, E., and Bellugi, U. (1979). The signs of language, Harvard University Press, USA, pp.205

Liddell, S. and Johnson, R. (1986). American Sign Language Compound Formation Processes and Phonological Remnants, In *Natural Language and Linguistic Theory*, nol.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. Proc. of *5th Workshop on the Representation and Processing of Sign Languages*, LREC-2012, 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, LREC 2010, Valetta, Malta.

Phau, R., and Josep, Q., (2010), Nonmanuals: their grammatical and prosodic roles., Sign Languages, In D. Brentari (ed). 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.