The POLYTROPON Parallel Corpus

Eleni Efthimiou, Kiki Vasilaki, Stavroula-Evita Fotinea, Anna Vacalopoulou,

Theodoros Goulas, Athanasia-Lida Dimou

ILSP / “Athena” R.C.
Artemidos 6 & Epidavrou, Maroussi, GR-15125, Greece
{eleni_e, kvasilaki, evita, avacalop, tgoulas, ndimou}@ilsp.gr

Abstract

Here we present the POLYTROPON parallel corpus for the language pair Greek Sign Language (GSL) – Modern Greek, which is created and annotated aiming to serve as a golden corpus available to the community of SL technologies for experimentation with various approaches to SL processing, focusing on machine learning for SL recognition, machine translation (MT) and information retrieval. The corpus volume incorporates 3,600 sentences performed by a single signer in three repetitions each, captured in front view by means of one HD and one kinect camera. Corpus creation was based on the validation procedure of a set of 2,000 lemmas deriving from the GSL segment of the Dicta-Sign corpus. Annotation of the corpus has provided interesting results in relation to all representation levels discussed within grammar theory, namely, lexicon, morphology, syntax, and semantics. Furthermore, it has allowed extraction of initial feature sets with the aim to reach a GSL level of abstraction close to the one currently available for Greek language representations, exploiting the inherent characteristics of the language. The POLYTROPON corpus is available to the SL research and SL technologies community via the CLARIN:EL infrastructure.

Keywords: SL data acquisition, GSL-Greek bilingual annotated resource, SL technologies, SL-Text parallel golden corpus, SL-based machine learning

1. Introduction

In the framework of research activities undertaken within the POLYTROPON project[1], significant effort was placed in maintaining and extending a Greek Sign Language 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. The aim of this venture was to unify GSL lexical resources acquired during a time span of approximately fifteen years of different acquisition phases. The methodological principles and rationale for revisiting and recapturing the existing GSL lexicon resources have been reported in Dimou et al. (2014) based on the usability plan of the database designed to accommodate the new lexicon resource.

Thus, the POLYTROPON lexical database (POLYTROPON Bilingual Dictionary, 2015) was created with a threefold goal: i) to gather and recapture already available lexical resources of 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 but also in developing various SL technologies, including information extraction, Web accessibility tools, incorporation of lexical information in natural language processing (NLP) systems for SL processing as in the case of machine translation from and into sign language, creation of training material for sign recognition technologies and input to sign synthesis tools enabling signing by virtual signers (i.e. avatars).

In Efthimiou et al. (2016), a detailed account of the newly acquired lexicon resource provided information as regards the features associated with each lemma in the POLYTROPON lexicon database, as well as the way this information has been visualized (Fig. 1) to make the lexicon content accessible by end users outside the SL research community, mainly targeting: (i) the bilingual education of deaf children, and (ii) of the learning of GSL as a second language (L2).

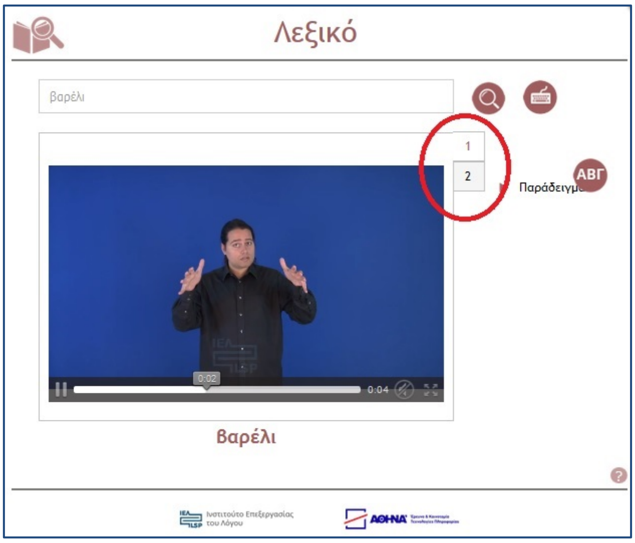
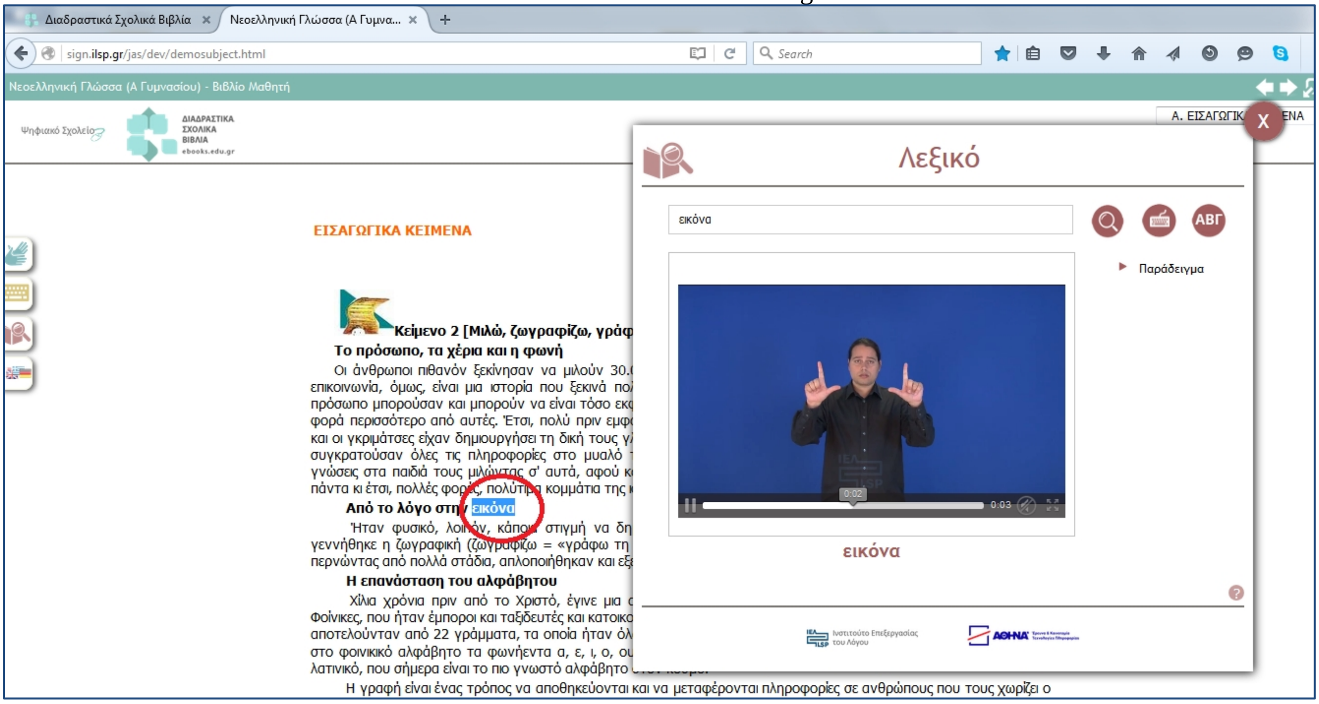


Figure. 1: Indicative snapshot of the visualization environment of the POLYTROPON lexical database. Here, two GSL synonyms are linked to one sense and one corresponding lemma in Modern Greek.

Efthimiou et al. (2016) provided examples of use of the POLYTROPON lexical resource in two educational platforms, namely, the official educational content platform for secondary education in Greece, and an

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

e-class platform as adapted by the Technical Vocational Institute of Athens (TEI-A), demonstrating the usability of this resource in the context of SL technologies based on lemma matching, such as dynamic synthetic signing and written text accessibility (Fig. 2).


Διαδραστικά Σχολικά Βιβλία
Νεοελληνική Γλώσσα (Α Γυμνα...
sign.ilsp.gr/jas/dev/demosubject.html
Search
Νεοελληνική Γλώσσα (Α Γυμνασίου) - Βιβλίο Μαθητή
Ψηφιακό Σχολείο
ΔΙΑΔΡΑΣΤΙΚΑ ΣΧΟΛΙΚΑ ΒΙΒΛΙΑ ebooks.edu.gr
Α. ΕΙΣΑΓΩΓΙΚ ENA
X
Λεξικό
εικόνα
ΑΒΓ
Παράδειγμα
0:02
0:03
εικόνα
?
Ινστιτούτο Επεξεργασίας του Λόγου
ΑΘΗΝΑ

ΕΙΣΑΓΩΓΙΚΑ ΚΕΙΜΕΝΑ

Κείμενο 2 [Μιλώ, ζωγραφίζω, γράφ

Το πρόσωπο, τα χέρια και η φωνή

Οι άνθρωποι πιθανόν ξεκίνησαν να μιλούν 30.
επικοινωνία, όμως, είναι μια ιστορία που ξεκινά πο
πρόσωπο μπορούσαν και μπορούν να είναι τόσο εκ
φορά περισσότερο από αυτές. Έτσι, πολύ πριν εμφ
και οι γκριμάτσες είχαν δημιουργήσει τη δική τους γ
συγκρατούσαν όλες τις πληροφορίες στο μυαλό
γνώσεις στα παιδιά τους μιλώντας σ' αυτά, αφού κ
πάντα κι έτσι, πολλές φορές, πολύτιμα κομμάτια της

Από το λόγο στην εικόνα

Ήταν φυσικό, λοιπόν, κάποια στιγμή να δη
γεννήθηκε η ζωγραφική (ζωγραφίζω = «γράφω τη
περνώντας από πολλά στάδια, απλοποιήθηκαν και εξ

Η επανάσταση του αλφάβητου

Χίλια χρόνια πριν από το Χριστό, έγινε μια
Φοίνικες, που ήταν έμποροι και ταξιδευτές και κατοικ
αποτελούνταν από 22 γράμματα, τα οποία ήταν ό
στο φοινικικό αλφάβητο τα φωνήεντα α, ε, ι, ο, ο
λατινικό, που σήμερα είναι το πιο γνωστό αλφάβητο

Η γραφή είναι ένας τρόπος να αποθηκεύονται και να μεταφέρονται πληροφορίες σε ανθρώπους που τους χωρίζει ο

Figure 2: Web text accessibility tool exploiting GSL lexicon database content.

In the rest of the paper, we will refer to the specific segment of the POLYTROPON resource, which is composed of GSL sentences added as examples of use for lexicon lemmas and their Greek translations. This set of data formed an independent parallel corpus resource, extensively annotated to serve development of SL technologies that crucially rely on availability of a "golden" corpus for machine learning purposes.

2. Corpus Content and Acquisition Methodology

As previously mentioned, the main objective that led to the creation of the POLYTROPON corpus was to build a bilingual parallel corpus for the language pair Greek Sign Language – Modern Greek that could serve as a "golden" corpus available to the community of SL technologies for experimentation with various approaches to SL processing, focusing on machine learning for SL recognition, MTand information retrieval (Efthimiou et al., 2015).

Corpus creation was based on the validation procedure of a set of 2,000 lemmas, originally derived from the GSL segment of the Dicta-Sign corpus (Matthes et al., 2012) in a three-step process:

- *In step 1*, all lemmas were spontaneously commented on by a working group of experts during unofficially recorded sessions.
- *In step 2*, selected sentences from these discussions were recorded in three repetitions each, in studio conditions at a later stage. Among the working group there was consensus that the selected sentences also form good examples of use of the discussed lemmas, so that they are eligible for inclusion in the lexicon database in the "example of use" information column.
- *In step 3*, one out of the three repetitions of each recorded sentence was annotated in iLex (Hanke & Storz, 2008; Efthimiou et al., 2016).

Annotation of the sentences on gloss level revealed the use of new lemmas within sentence content, not initially included in the lexicon. Thus, the above described procedure of lemma validation was repeated for the new lemmas we well. Discussion of new lemmas resulted in new sentences in an iterative process, enriching the GSL iLex lexicon DB with 1,600 new lemmas in total, while commentaries on new entries generated the new clauses which completed acquisition of the content of the POLYTROPON corpus following baseline elicitation principles as in Matthes et al. (2010).

In total, the POLYTROPON parallel corpus incorporates 3,600+ clauses in three repetitions each, captured in front view by means of one HD and one kinect camera.

In the next section, an account of the adapted annotation scheme and annotation findings is provided.

3. Corpus Annotation Scheme and Annotation Findings

3.1 Corpus Annotation Scheme

The corpus Annotation scheme (Fig. 3) entails the following tier set:

"*Clause*" defining clause boundaries of the signed utterances,
"*Gloss*" assigned to each identified token and provided in Greek,
"*Greek equivalent clause*", which provides the Greek translation of each signed unit within the "Clause" time frame,
"*Classifier*", which provides one tier for each classifier handshape and allows feature values assignment with respect to one- or two-hands and same or different activity and the classifier's semantic content (Fig. 4), using the labels [entity], [shape], [handling] and [predicative],
"*S-type*" to mark main vs. subordinate constructions,
"*S-category*" for the marking of sentence categories according to syntactic classification that receives feature values based on classical descriptive grammar classification.

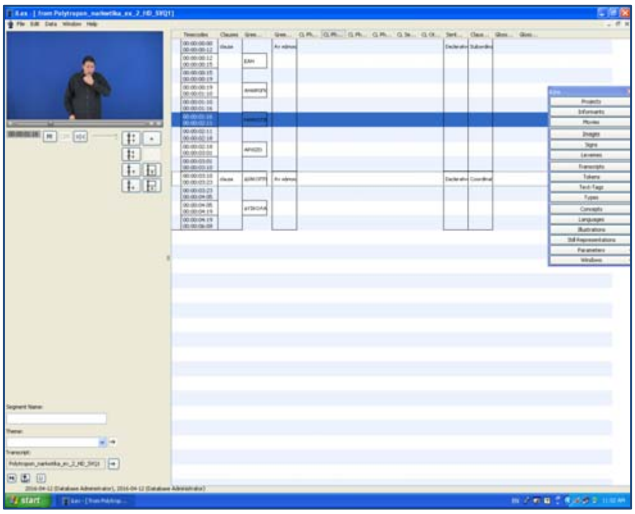


Figure 3: The overall POLYTROPON corpus annotation scheme in iLex.

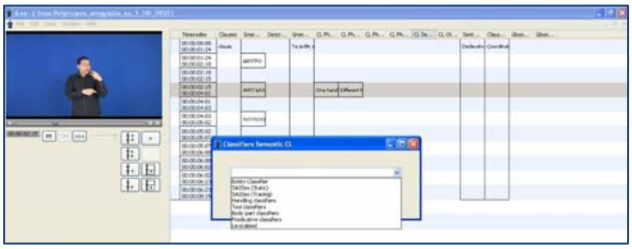


Figure 4: Classifier annotation according to semantic function.

Annotation was performed by a coda GSL expert in the iLex annotation environment and was cross-checked by two more linguists with expertise in annotation and analysis of GSL data.
The translation of the annotated sentences in Greek was performed in two phases:

a) a strongly GSL-influenced initial translation version provided directly by the annotator, followed by
b) a "corrected" translation version, which provided fully acceptable Greek sentences as regards naturalness and grammaticality, performed by an expert in Greek language.

The overall annotation scheme was designed to provide a range of information expressed in terms of lexical and sentential feature bundles, aiming to allow for search options targeting morphology, semantics and syntax relevant events (Liddell and Johnson, 1986; Pfau and Quer, 2010; Quer et al., 2017).
Regarding representation-level related information, it must be noticed that although no articulation (phonological-phonetic) information is visible in the corpus as regards sign tokens, this is directly available in the POLYTROPON lexical database (Efthimiou et al., 2016), where lemmas receive full phonetic articulation descriptions according to the HamNoSys notation system (Prillwitz et al., 1989; Hanke, 2004) and the SiS-Builder embedded non-manuals notation tool (Goulas et al., 2010; Efthimiou et al., 2014).
By the completion of the annotation process, the POLYTROPON corpus proved to be a rich source of GSL grammar information, equally useful to SL technologies oriented research and SL theoretical linguistic analysis.

3.2 Annotation Findings

In this section, the major linguistic findings of the annotation process are listed, since they bring insight as to a series of phenomena SL technologies need to tackle.

Gloss-level findings
Gloss-level annotation has made visible three types of lemmas not previously contained either in the iLex lexicon or in the GSL-Greek bilingual lexicon database in the SiS-Builder environment. These involve:

a) A set of new lemmas which were not previously included in any of our two databases. These were directly added in the iLex lexicon during annotation process, and also created a set of new reference type entries which enriched the SiS-Builder lexicon database. This procedure enriched both databases with 1600 entries in total.

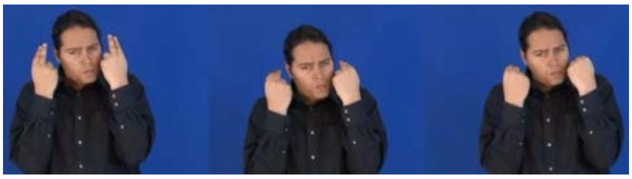


Figure 5: Sequence in manual and non manual activity when articulating the GSL expression EARS-DOWN.

b) A set of GSL-specific expressions with no direct translation equivalent in Greek or English, such as EMPTY-POCKETS to imply the meaning of "I am broke" or EARS-DOWN to express the meaning of "obey" (Fig. 5). These were classified under "GSL special expressions" in both lexicon databases.

c) A set of exclamation gestures with semantic value in direct equivalence to embodied signals of oral expression as, for instance, the embodied exclamatory expression adding affect-related extra-linguistic information to strengthen utterances such as “what can I say!”, “I don’t know!” etc. (Fig. 6).

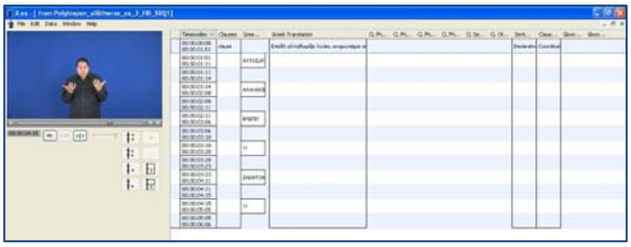


Figure 6: Embodied extra-linguistic expression commonly met in GSL and Greek.

Compounds

Regarding compound formations, the POLYTROPON corpus annotation allowed for identification of the following compounding options in line with formal descriptions as in Sandler & Martin (2006):

- *classifier+classifier*: as in formation of the GSL sign “lighthouse” incorporating the scheme: [CL-5C+CL-5]
- *sign+sign*: as in formation of the GSL sign “air hostess”: involving the sign combination [AIRPLANE]+[ACCOMPANY], or the sign “cow” exploiting [ANIMAL]+[MILK].
- *sign+ classifier*: as in formation of the GSL sign “pilot”, exploiting the combination: [AIRPLANE]+[CL-S]
- *classifier+sign*: as in formation of the GSL sign “letter”, by means of the combination [CL-C1]+[SEAL]
- *ad-hoc formations*: such formations involve concatenation of signs, as for example in [PAPER-NOTE-REMEMBER] to express the meaning of “take notes”, and are characterized by their unique appearance in the entire corpus. This set of lemmas has not been incorporated in the lexicon databases yet, since they need to be further cross-checked with more native signers in order for their compound status to be validated.

Classifiers

As regards the classifier content of POLYTROPON, annotation was performed (i) on morphophonemic level involving markings related to formation and including information as to handshape, two- or one-hand activity, and same or different activity performed by the hands, and (ii) on semantic-content level, assigning the labels: [entity] [shape] [handling] and [predicative] from a drop-down menu, following an analysis as in Efthimiou et al. (2010).

Sentence-level findings

On sentence level, two major clause categories were annotated, that is, *main* and *subordinate*.

Under main clauses, there are also coordination constructions which in many cases replace oral language subordination as when coordination makes use of the INDEX/TOPIC mechanism, which replaces Relative Clause subordination, met in a number of oral languages. Subordinate clause formation involves constructions which present clear subordination markers extensively indicated via the combination of manual and non-manual signals as in the case of the causative marker [BECAUSE] (Fig. 7), or in conditional constructions where the semantics of “if” are expressed both via manual and non-manual elements and where presence of the non manuals is obligatory.

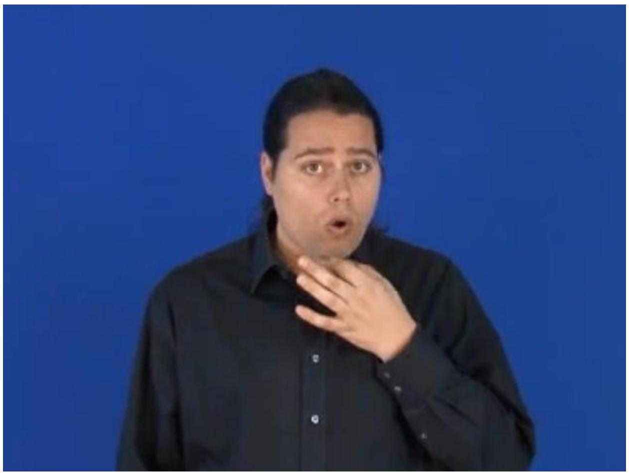


Figure 7: The causative marker [BECAUSE] that introduces subordinate Causative Clause in GSL.

Sentence type

For both main and subordinate clauses, further classification assigns sentence category values from the following list:

- Declarative-Affirmative
- Declarative-Negative
- Interrogative (Yes/No, Wh)
- Rhetoric Q&A
- Imperative
- Exclamation

Rhetoric Q&As are declarative constructions, which incorporate a WH-like utterance without exhibiting the full non-manual activity usually present in WH-questions, aiming to enforce the focus catching effect of the signed message. Examples of rhetoric Q&A are utterances like [TODAY-EAT-WHAT-RICE-WITH-CHICKEN], which expresses the message “Guess what I will eat today! Rice with chicken!”.

3.3 Corpus Exploitation

It has already been mentioned that creation of the POLYTROPON parallel corpus has been directed towards its application in the SL technologies domain, mainly targeting the need for annotated data. Given work with the language pair GSL-Greek, the main aim has been to provide a significant amount of GSL data annotated, which may allow reaching a similar level of abstraction for both language representations. This abstraction should be succeeded by making exploitable the inherent characteristics of both languages, thus, reaching a state where we can apply deep learning experiments on GSL data, where representation of both words and signs takes the form of a vector of characteristics as in (Fig. 8).

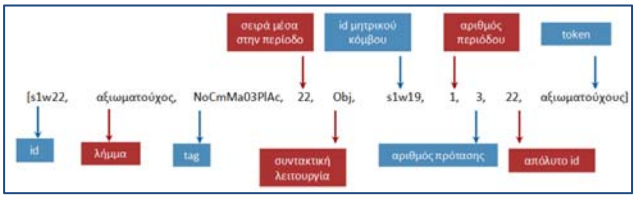


Figure 8: Feature vector representation of lexical items.

Furthermore, since the goal of the acquisition team has been to provide the research community with a golden corpus for machine learning in the areas of SL corpus mining and MT, the current release of the POLYTROPON parallel corpus is available via the *clarin:el* repository, which is the Greek sector of CLARIN[2], the European infrastructure for language resources and technology. The corpus is available free of charge but subject to Creative Commons (*CC - BY*) licensing[3]. For its identification in *clarin:el*, the POLYTROPON parallel corpus has been assigned the persistent identifier (PID):

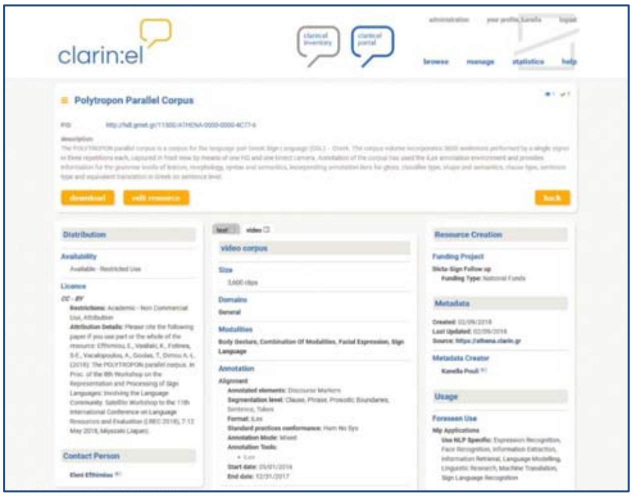


Figure 9: Word/sign representation as a vector of characteristics.

http://hdl.grnet.gr/11500/ATHENA-0000-0000-4C77-6 (Fig. 9), while academic users may directly reach the resource in the *clarin:el* platform by following the link: https://athena.clarin.gr/resources/browse/polytropon-parallel-corpus/197061c20d9711e89c26aa3fc8d33ad8b716f4f795884a8792b708207d02bd84/.

Regarding sentence-level representation, experimentation is currently oriented towards exploiting Dependency Tree Structure representations of input text and signed clauses using Tree Editor TrEd 2.0.

4. Conclusion and Future Plans

The POLYTROPON parallel corpus was created to mainly address SL processing needs in the framework of human language technologies applications seeking mainly ways to extend our current knowledge with respect to corpus-based and statistical approaches to MT, but also in service of SL technologies with focus on sign recognition, information extraction and information retrieval from video sources. The resource aims to trigger new challenges both on technological and SL linguistic grounds. In this context and in order to better serve the goal of exploiting the corpus in the context of machine learning by providing a multi-signer approach to the already acquired data, a crowd-sourcing activity is currently planned, which will invite native GSL signers to repeat a selected segment of the corpus content with the aim to enrich the variability of signers and signing space conditions towards serving machine-learning purposes more effectively.

5. Acknowledgements

We acknowledge support of this work by the project “Computational Science and Technologies: Data, Content and Interaction” (MIS 5002437) which is implemented under the Action “Reinforcement of the Research and Innovation Infrastructure”, funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

6. Bibliographical References

Dimou, A-L., Goulas, T., Efthimiou, E., Fotinea, S-E., Karioris, P., Pissaris, M., Korakakis, D., and Vasilaki, K. (2014). Creation of a multipurpose sign language lexical resource: The GSL lexicon database. Proceedings of 6th Workshop on the Representation and Processing of Sign Languages, LREC’14, Reykjavik, Iceland, pp. 37–42.

E. Efthimiou, S-E. Fotinea, A-L. Dimou, T. Goulas, P. Karioris, K. Vasilaki, A. Vacalopoulou, M. Pissaris, and D. Korakakis. (2016). From a sign lexical database to an SL golden corpus – the POLYTROPON SL resource. Proceedings of the 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining. Satellite Workshop to the 10th International Conference on Language Resources and Evaluation (LREC-’16), 28 May 2016, pp. 63–68, Portoroz, Slovenia.

Efthimiou, E., Fotinea, S-E., Goulas, T., and Kakoulidis, P. (2015). User friendly Interfaces for Sign Retrieval and Sign Synthesis. In M. Antona, M. & C. Stephanidis (Eds.). Proceedings 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., and Pissaris, M. (2014). SiS-builder: A Tool to Support Sign

[2] CLARIN: Common Language Resources and Technology Infrastructure (www.clarin.eu).

[3] https://creativecommons.org.

Synthesis. Proceedings of 2nd Int'l Conference on the Use of New Technologies for Inclusive Learning, pp. 26–36. York, UK.

Efthimiou, E., Fotinea, S-E., and Dimou, A-L. (2010). Towards decoding Classifier function in GSL. In Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010), 2010, Satellite workshop of the LREC-'10 Conference, Valetta, Malta, pp.76–79.

Goulas, T., Fotinea, S-E., Efthimiou, E., and Pissaris, M. (2010). SiS-Builder: A Sign Synthesis Support Tool. In Dreuw, P. et al. (eds.), Proceedings of 4th Workshop on Representation and Processing of Sign Languages, LREC-'10, pp. 102–105.

Hanke,T. (2004). HamNoSys - representing sign language data in language resources and language processing contexts. Proceedings of 1st Workshop on Representing and Processing of Sign Languages, LREC-'04, 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.* Proceedings 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,* 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. Proceedings of 5th Workshop on the Representation and Processing of Sign Languages, LREC-'12, Istanbul, Turkey.

Matthes S., Hanke T., Storz J., Efthimiou E., Dimou A-L, Karioris P., Braffort A., Choisier A., Pelhate J., and Safar E. (2010). Elicitation tasks and materials designed for Dicta-Sign's multi-lingual corpus, LREC '10, 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.

Quer, J., Cecchetto, C., Donati, C., Geraci, C., Kelepir, M., Pfau, R., and Steinbach, M. (eds): (2017). *SignGram Blueprint: A Guide to Sign Language Grammar Writing*. DE GRUYTER MOUTON. ISBN 978-1-5015-1180-6.

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

7. Language Resource References

POLYTROPON Bilingual Dictionary. (2015). Clarin:el repository persistent identifier (PID): http://hdl.grnet.gr/11500/ATHENA-0000-0000-42D5-5.

POLYTROPON Parallel Corpus. (2017). Clarin:el repository persistent identifier (PID): http://hdl.grnet.gr/11500/ATHENA-0000-0000-4C77-6