Sign Language Recognition, Generation, and Modelling: A Research Effort with Applications in Deaf Communication

Eleni Efthimiou[1], Stavroula-Evita Fotinea[1], Christian Vogler[1], Thomas Hanke[2],
John Glauert[3], Richard Bowden[4], Annelies Braffort[5], Christophe Collet[6],
Petros Maragos[7], and Jérémie Segouat[8]

[1] Institute for Language and Speech Processing
{eleni_e,evita,cvogler}@ilsp.gr
[2] Universität Hamburg
thomas.hanke@sign-lang.uni-hamburg.de
[3] University of East Anglia
J.Glauert@uea.ac.uk
[4] University of Surrey
R.Bowden@surrey.ac.uk
[5]LIMSI/CNRS
annelies.braffort@limsi.fr
[6] Université Paul Sabatier
collet@irit.fr
[7] National Technical University of Athens
maragos@cs.ntua.gr
[8] WebSourd
jeremie.segouat@websourd.org

Abstract. Sign language and Web 2.0 applications are currently incompatible, because of the lack of anonymisation and easy editing of online sign language contributions. This paper describes Dicta-Sign, a project aimed at developing the technologies required for making sign language-based Web contributions possible, by providing an integrated framework for sign language recognition, animation, and language modelling. It targets four different European sign languages: Greek, British, German, and French. Expected outcomes are three showcase applications for a search-by-example sign language dictionary, a sign language-to-sign language translator, and a sign language-based Wiki.

Keywords: Sign Language, Deaf communication, HCI, Web accessibility.

1 Introduction

The development of Web 2.0 technologies has made the WWW a place where people constantly interact with another, by posting information (e.g. blogs, discussion forums), modifying and enhancing other people's contributions (e.g. Wikipedia), and sharing information (e.g., Facebook, social news sites). The choice of human-computer interface plays a critical role in these activities.

C. Stephanidis (Ed.): Universal Access in HCI, Part I, HCII 2009, LNCS 5614, pp. 21–30, 2009.

Today's predominant human-computer interface is relatively manageable for most Deaf people, despite lingering accessibility problems. The use of a language foreign to them is restricted to single words or short phrases. The graphical user interface, however, puts severe limitations on the complexity of the human-computer communication, and therefore it is expected that in many contexts the interface will shift to spoken human language interaction.

Obviously, with such a shift, a far better command of the interface language is required than with graphical environments. Most Deaf people would, therefore, be excluded from this future form of human-computer communication, unless the computer is also able to communicate in sign language. Moreover, they already are largely excluded from interpersonal communication among themselves on the Web, given the current lack of support for applications for sign language-to-sign language, but also spoken-to-sign language, and sign-to-spoken language.

Sign language videos, their current popularity notwithstanding, are not a viable alternative to text, for two reasons: First, they are not anonymous – individuals making contributions can be recognized from the video and therefore excludes those who wish their identity to remain secret. Second, people cannot easily edit and add to a video that someone else has produced, so a Wikipedia-like web site in sign language is currently not possible.

In order to make the Web 2.0 fully accessible to Deaf people, sign language contributions must be displayed by an animated avatar, which addresses both anonymisation and easy editing. The remainder of the paper describes the Dicta-Sign project, the overarching goal of which is to lay the groundwork for Web 2.0-style contributions in signed languages.

2 The Dicta-Sign Project

Dicta-Sign (http://www.dictasign.eu) is a three-year consortium research project that involves the Institute for Language and Speech Processing, the University of Hamburg, the University of East Anglia, the University of Surrey, LIMSI/CNRS, the Université Paul Sabatier, the National Technical University of Athens, and WebSourd. It aims to improve the state of web-based communication for Deaf people by allowing the use of sign language in various human-computer interaction scenarios. It will research and develop recognition and synthesis engines for signed languages at a level of detail necessary for recognizing and generating authentic signing.

In this context, Dicta-Sign aims at developing several technologies demonstrated via a sign language-aware Web 2.0, combining work from the fields of sign language recognition, sign language animation via avatars, sign language linguistics, and machine translation, with the goal of allowing Deaf users to make, edit, and review avatar-based sign language contributions online, similar to the way people nowadays make text-based contributions on the Web.

Dicta-Sign supports four European sign languages: Greek. British, German, and French Sign Language. Users make their contributions via webcams. These are recognized by the sign language recognition component (Section 3) and converted into a linguistically informed internal representation which is used to animate the contribution with an avatar (Section 4), and to translate it into the other respective three sign languages (Section 5).

Dicta-Sign differs from previous work in that it aims to integrate tightly recognition, animation, and machine translation. All these components are informed by appropriate linguistic models from the ground up, including phonology, grammar, and nonmanual features. A key aspect of the Dicta-Sign project is the creation of parallel corpora in the four above-mentioned different signed languages with detailed annotations. These not only greatly aid the development of language models for both recognition and animation, but also allow for the direct alignment of equivalent utterances across the four languages, which is useful for creating machine translation algorithms in a sign language-to-sign language translator (Sections 5 and 6).

The project will work closely with the Deaf communities in the countries of the project partners throughout its lifecycle to ensure that its goals are met, and to evaluate user acceptance. A major part of this evaluation consists of three showcase applications that highlight how the various aspects of the system work together (Section 6).

We now cover the three major components of the system —recognition, animation, and linguistic resources— in detail.

3 Sign Language Recognition

Despite intensive research efforts, the current state of the art in sign language recognition leaves much to be desired. Problems include a lack of robustness, particularly when low-resolution webcams are used, and difficulties with incorporating results from linguistic research into recognition systems. Moreover, because signed languages exhibit inherently parallel phenomena, the fusion of information from multiple modalities, such as the hands and the face, is of paramount importance. To date, however, relatively little research exists on this problem [1].

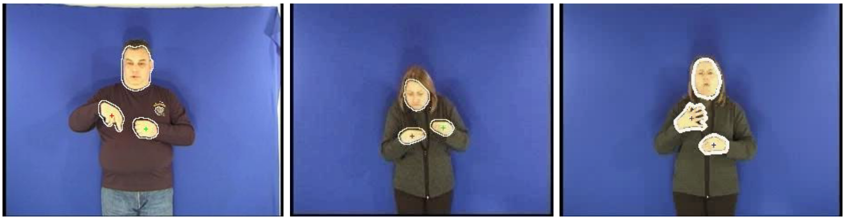


Fig. 1. Signer-independent visual tracking and feature extraction

3.1 Visual Tracking and Feature Extraction

The features that serve as input to the recognition system comprise a mix of measurements obtained by statistical methods, and geometrical characterisations of the signer's body parts, as shown in Figure 1. In the example shown in this figure, the face is roughly located via the Viola-Jones face detector [2], which then gives rise to a skin color model, which in turn is used to locate the signer's face and hands with a greater degree of precision. Based on these initial estimates, object-oriented morphological filtering extracts the silhouette of the face and the hands [3, 4].

In order to make the feature extraction process robust even when the image comes from commodity webcams, the computer vision algorithms need to operate on multiple scales. Moreover, the basic feature extraction processes need to be combined with statistical and learning-based methods, such as active appearance models for facial expression tracking [5, 6].

3.2 Continuous Sign Language Recognition

Hidden Markov model (HMM)-based approaches are the most popular approach to continuous sign language recognition, partly due to their great success in speech recognition [7, 8]. At the same time, there are important differences between speech and sign language recognition; foremost among them is the fact that sign language is inherently multimodal: both hands move in parallel, while the face and body exhibit grammatical and prosodic information [9]. Hence, sign language recognition must deal with the problem of fusing multiple channels of information.

Product and parallel HMMs have been suggested in the past as a possible solution to the problem [10, 8]; however, both approaches have the drawback that they require assigning weights that reflect the relative importance of each modality. Choosing these weights statically, as has been done in previous work, is ultimately unsatisfactory, because the reliability of the information in each channel can change dynamically, due to noise, the context in which the signs are executed, and the signing style of the particular person. A robust dynamic weighting scheme must, therefore, be chosen, so as to evaluate the amount of information that each modality carries, and to maximize their discriminative abilities.

To ensure user acceptance, the recognition system must be able to work in a signer-independent way. To this end, it employs well-known HMM adaptation methods from the speech recognition. Even so, given the current state of the art in sign language recognition, one cannot expect the system to recognize the full range of expressiveness in signed languages. We deal with this limitation in two ways: First, the prototype application is domain-specific, with a restricted vocabulary of no more than 1500 signs. Second, the system employs a dictation-style interface (hence the name "Dicta-Sign"), where the user is presented with the closest-matching alternatives if a sign is not recognized reliably.

The output of the recognition component is converted into a linguistically informed representation that is used by the synthesis and language modelling components, respectively.

4 Synthesis and Animation

Speech technology has exploited properties of phonological composition of words with respect to spoken languages, so as to develop speech synthesis tools for unrestricted text input. In the case of sign languages, a similar approach is being experimented with, with the goal of generating signs (word level linguistic units of sign languages) with an avatar not by mere video recording, but rather by the composition of sign phonology components (Figure 2) [11, 12].

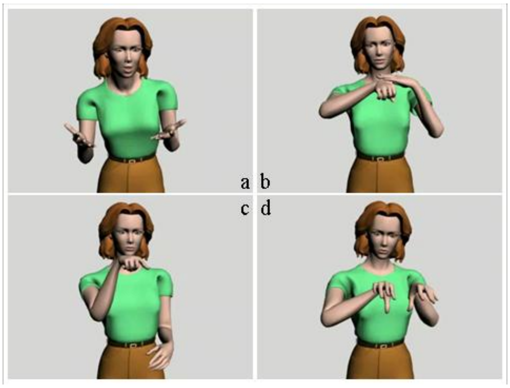


Fig. 2. The signing avatar

Sign language synthesis is heavily dependent on the natural language knowledge that is coded in a lexicon of annotated signs, and a set of rules that allows structuring of core grammar phenomena, making extensive use of feature properties and structuring options. This is necessary in order to guarantee the linguistic adequacy of the signing performed. In the Dicta-Sign project, the annotated parallel corpora provide the basis for these rules (see also Section 5.3), which encompass manual and non-manual features, as well as the role of placement of signs in space [13].

The internal representation of sign language phrases is realized via SiGML [14], a Signing Gesture Markup Language to support sign language-based HCI, as well as sign generation. The SiGML notation allows sign language sequences to be defined in a form suitable for execution by a virtual human, or avatar, on a computer screen. The most important technical influence on the SiGML definition is HamNoSys, the Hamburg Notation System [15], a well-established transcription system for sign languages. The SiGML notation incorporates the HamNoSys phonetic model, and hence SiGML can represent signing expressed in any sign language.

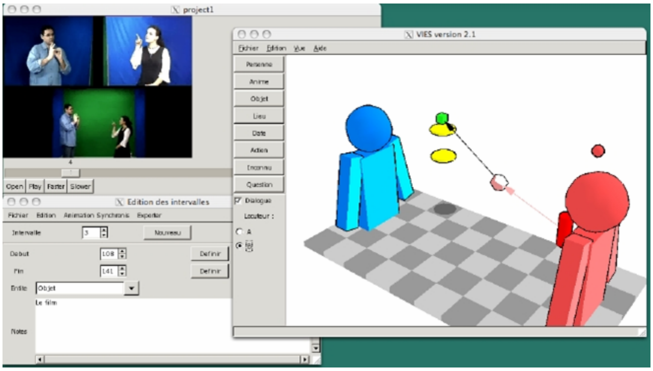
One of the most difficult problems in sign synthesis is converting a linguistic description of the signed utterance into a smooth animation via inverse kinematics, with proper positioning of the hands in contact with the body, and generating realistic prosodic features, such as appropriate visual stress. To this end, the sign language corpus, as described in the next section, does not only encompass phonetic and grammatical information, but also prosodic information. Together with the features derived from the visual tracking and recognition component, this allows for greatly increased realism in the animations.

5 Sign Language Linguistic Resources

In the following, we describe the linguistic resources that contribute to all the other components of the Dicta-Sign project. They can broadly be divided into language modelling, support for annotation tools, and the collection of parallel sign language corpora.

5.1 Linguistic Modelling

Linguistic modelling will develop a coherent model from the phonetic up to the semantic level of language representation, envisaged to be language-independent in most aspects. This modelling will cover a broad range of phenomena, including the use of the signing space (Figure 3), and the coordination of manual with nonmanual features, such as facial expressions and eye gaze. The input data for the development of the linguistic model will be provided by the lemmatized project corpora (see also Section 5.3).



Intervalle 3

Debut 108

Fin 141

Entite Objet

Notes Le film

Fig. 3. Editor used to model the signing space

Whereas the first sign language grammar models developed in previous projects (ViSiCAST [12] and SYNENNOESE [16]) were mainly dedicated to generation purposes, Dicta-Sign aims to extend modelling capabilities toward a common representation of sign language grammar and the lexicon —or alternatively two coherent representations— to accommodate both sign language recognition and synthesis. Overall, this represents a major advance over previous work, since language modelling has been largely neglected particularly in the recognition field.

5.2 Annotation Tools

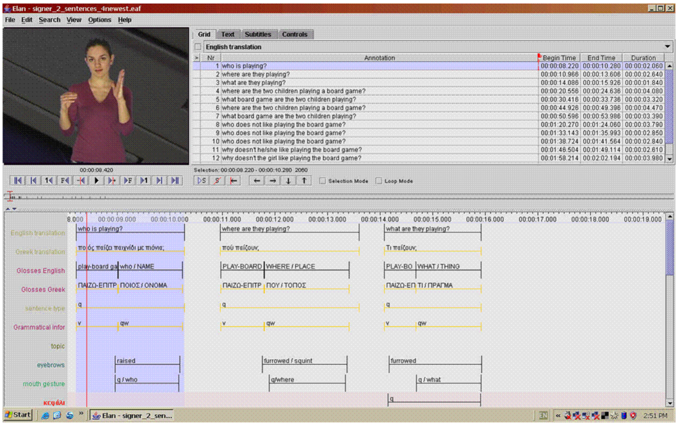
Most mainstream annotation tools, such as ELAN and Anvil, are geared toward the processing of spoken languages. As such, they lack some features that would facilitate the processing of signed languages. These include a graphical representation of sign language utterances, and special input methods for sign language notation systems (e.g., HamNoSys [15]). Although some tools exist for specifically processing signed languages, such as iLex, none of these tools currently provide any kind of automated tagging, so the annotation process is completely manual.

An experimental version of the AnCoLin annotation system allows some image processing tasks to be initiated from within the annotation environment and to compare the results with the original video [17,18]**.** It also connects to a 3D model of the signing space, but still lacks a coherent integration into the annotation workflow.

It is expected that one of the major outcomes of the Dicta-Sign project will be greatly improved annotation tools, with image processing and recognition integrated into the annotation workflow. Their long term utility can be judged by the uptake by other sign language researchers.

5.3 Sign Language Corpora and Translation

An electronic corpus is of the utmost importance for the creation of electronic resources (grammars and dictionaries) for any natural language. For multi-lingual research and applications, parallel corpora are basic elements, as in the case of translation-memory applications and pattern-matching approaches to machine translation. Furthermore, a substantial corpus is needed to drive automatic recognition and generation, so as to obtain sufficient data for training and language representation.



Nr	Annotation	Begin Time	End Time	Duration
1	who is playing?	00:00:08.220	00:00:10.280	00:00:02.060
2	where are they playing?	00:00:10.966	00:00:13.606	00:00:02.640
3	what are they playing?	00:00:14.086	00:00:15.926	00:00:01.840
4	where are the two children playing a board game?	00:00:20.556	00:00:24.636	00:00:04.080
5	what board game are the two children playing?	00:00:30.416	00:00:33.736	00:00:03.320
6	where are the two children playing a board game?	00:00:44.926	00:00:49.396	00:00:04.470
7	what board game are the two children playing?	00:00:50.596	00:00:53.986	00:00:03.390
8	who does not like playing the board game?	00:01:20.270	00:01:24.060	00:00:03.790
9	who does not like playing the board game?	00:01:33.143	00:01:35.993	00:00:02.850
10	who does not like playing the board game?	00:01:38.724	00:01:41.564	00:00:02.840
11	why doesn't he/she like playing the board game?	00:01:46.504	00:01:49.114	00:00:02.610
12	why doesn't the girl like playing the board game?	00:01:58.214	00:02:02.194	00:00:03.980

English translation	Greek translation	Glosses English	Glosses Greek	sentence type	Grammatical infor	eyebrows	mouth gesture	κεφάλι
who is playing?	ποιός παίζει παιχνίδι με πιόνια;	play-board ga who / NAME	ΠΑΙΖΩ-ΕΠΙΤΡ ΠΟΙΟΣ / ΟΝΟΜΑ	q	v qw	raised	q / who	
where are they playing?	πού παίζουν;	PLAY-BOARD WHERE / PLACE	ΠΑΙΖΩ-ΕΠΙΤΡ ΠΟΥ / ΤΟΠΟΣ	q	v qw	furrowed / squint	q/where	
what are they playing?	Τι παίζουν;	PLAY-BO WHAT / THING	ΠΑΙΖΩ-ΕΠ ΤΙ / ΠΡΑΓΜΑ	q	v qw	furrowed	q / what	q

Fig. 4. Annotation of existing Greek Sign Language corpus with ELAN

The quality and availability of sign language corpora has improved greatly in the past few years [19, 20], where, among others, high-quality corpora exist for Greek, American, and German sign language (Figure 4). Yet, to date, multi-lingual sign language research has been hampered by the lack of sufficiently large parallel sign language corpora. One of the most important goals of Dicta-Sign is to collect the world's first large parallel corpus of domain-specific utterances across four signed languages (Greek, British, German, and French), with a minimum of three hours of signing in each language, and a minimum vocabulary of 1500 signs.

This corpus will be fully annotated, showcase best practices for sign language annotations, and be made available to the public. It is expected that the availability of this corpus will significantly boost the productivity of sign language researchers, especially those who are interested in comparing and contrasting multiple languages. In addition, the utterances in the respective four languages can be aligned automatically, thus opening the door for implementing shallow machine translation techniques [21,22], similar to state-of-the-art techniques for spoken languages (see also the showcase application in Section 6).

6 Application Domains

Dicta-Sign is an ambitious project that aims to integrate recognition, synthesis and linguistic modelling on a hitherto unseen scale. One of its key metrics of success is acceptance by the respective Deaf communities in the participating countries. To this end, three proof-of-concept prototypes will be implemented and evaluated within Dicta-Sign.

First, a search-by-example system will integrate sign recognition for isolated signs with interfaces for searching an existing lexical database. Aside from the obvious utility to sign language learners, this prototype will also showcase the technology behind the dictation characteristics of the user interface, where multiple alternatives are presented if a sign cannot be recognized reliably.

Second, a sign language-to-sign language translation prototype will pioneer a controlled-vocabulary sign language-to-sign language translation on the basis of the parallel language resources developed within the project. It will be the first project of its kind to make use of shallow translation technologies. This prototype will also serve as the project demonstrator.

Third, a sign language-based Wiki will be developed, providing the same service as a traditional Wiki but using sign language. This prototype will specifically showcase the integration of all major components of the project. At the same time, it will also demonstrate a Web 2.0 application that is accessible to the Deaf from the beginning to end.

7 Conclusions

Today, just a few months after the “European Year of Equal Opportunities for All,” it is important that drastic measures are taken to prevent new barriers from arising, as new forms of communication establish their role in the society at large. Dicta-Sign will be a key technology to promote sign language communication, and to provide Web 2.0 services and other HCI technologies to Deaf sign language users, an important linguistic minority in Europe so far excluded from these new developments.

As the field of sign language technology is still very young, it is beyond the scope of a three-year project to catch up completely with mainstream language technology, and to deliver end-user products. Nevertheless, Dicta-Sign is poised to advance significantly the enabling technologies by a multidisciplinary approach, and to come close enough to let designers of future natural language systems fully take sign languages into account.

Acknowledgements. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 231135.

References

1. Ong, A.C.W., Ranganath, S.: Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning. IEEE Trans. PAMI 27(6), 873–891 (2005)
2. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vision 57(2), 137–154 (2004)

3. Maragos, P.: Morphological Filtering for Image Enhancement and Detection. In: Bovik, A.C. (ed.) The Image and Video Processing Handbook, 2nd edn., pp. 135–156. Elsevier Acad. Press, Amsterdam (2005)
4. Sofou, A., Maragos, P.: Generalized Flooding and Multicue PDE-based Image Segmentation. IEEE Transactions on Image Processing 17(3), 364–376 (2008)
5. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active Appearance Models. IEEE Trans. PAMI 23(6), 681–685 (2001)
6. Papandreou, G., Maragos, P.: Multigrid Geometric Active Contour Models. IEEE Trans. Image Processing 16(1), 229–240 (2007)
7. Starner, T., Weaver, J., Pentland, A.: Real-Time American Sign Language Recognition Using Desk and Wearable Computer-Based Video. IEEE Trans. Pattern Analysis and Machine Intelligence 20(12), 1371–1375 (1998)
8. Vogler, C., Metaxas, D.: A Framework for Recognizing the Simultaneous Aspect of ASL. CVIU 81, 358–384 (2001)
9. Neidle, C., Kegl, J., MacLaughlin, D., Bahan, B., Lee, R.G.: The Syntax of American Sign Language: Functional Categories and Hierarchical Structure. MIT Press, Cambridge (2000)
10. Gravier, G., Potamianos, G., Neti, C.: Asynchrony modeling for audiovisual speech recognition. In: Proc. Human Language Technology Conference, San Diego, California (March 2002)
11. Fotinea, S.-E., Efthimiou, E., Karpouzis, K., Caridakis, G., Glauert, J. (eds.): A Knowledge-based Sign Synthesis Architecture. Emerging Technologies for Deaf Accessibility in the Information Society: Editorial. Journal of Universal Access in the Information Society 6(4), 405–418 (special issue, 2008)
12. Marshall, I., Sáfár, E.: Grammar Development for Sign Language Avatar-Based Synthesis. In: Proceedings HCII 2005, 11th International Conference on Human Computer Interaction (CD-ROM), Las Vegas, USA (July 2005)
13. Braffort, A., Bossard, B., Segouat, J., et al.: Modélisation des relations spatiales en langue des signes française. In: TALS 2005, atelier de TALN 2005 (2005)
14. Elliott, R., Glauert, J.R.W., Kennaway, J.R., Marshall, I.: Development of Language Processing Support for the Visicast Project. In: ASSETS 2000 4th International ACM SIGCAPH Conference on Assistive Technologies, Washington, DC, USA (2000)
15. Hanke, T.: HamNoSys - representing sign language data in language resources and language processing contexts. In: Streiter, O., Vettori, C. (eds.) LREC 2004, Workshop proceedings: Representation and processing of sign languages, pp. 1–6. ELRA, Paris (2004)
16. Efthimiou, E., Sapountzaki, G., Karpouzis, K., Fotinea, S.-E.: Developing an e-Learning Platform for the Greek Sign Language. In: Miesenberger, K., Klaus, J., Zagler, W.L., Burger, D. (eds.) ICCHP 2004. LNCS, vol. 3118, pp. 1107–1113. Springer, Heidelberg (2004)
17. Braffort, A., Choisier, A., Collet, C., et al.: Toward an annotation software for video of Sign Language, including image processing tools and signing space modelling. In: LREC 2004 (2004)
18. Gianni, F., Collet, C., Dalle, P.: Robust tracking for processing of videos of communication's gestures. In: International Workshop on Gesture in Human-Computer Interaction and Simulation (GW 2007), Lisbon, Portugal (May 2007)
19. Efthimiou, E., Fotinea, S.-E.: GSLC: Creation and Annotation of a Greek Sign Language Corpus for HCI. In: Stephanidis, C. (ed.) HCI 2007. LNCS, vol. 4554, pp. 657–666. Springer, Heidelberg (2007)

20. Neidle, C., Sclaroff, S.: Data collected at the National Center for Sign Language and Gesture Resources, Boston University (2002), `http://www.bu.edu/asllrp/ncslgr.html`
21. Koehn, P., Och, F.J., Marcu, D.: Statistical Phrase-Based Translation. In: Proceedings of the Human Language Technology Conference 2003 (HLT-NAACL 2003), Edmonton, Canada (May 2003)
22. Diab, M., Finch, S.: A Statistical Word-Level Translation Model for Comparable Corpora. In: Proceedings of the Conference on Content-Based Multimedia Information Access, RIAO 2000, Paris, France, April 12-14 (2000)