Cross Modal Comprehension in ZARDOZ

An English to Sign-Language Translation System

Tony Veale and Alan Conway

Hitachi Dublin Laboratory, O'Reilly Institute, Trinity College, Dublin 2, Ireland.
Phone: +353-1-6798911, Fax: +353-1-6798926
E-mail: Tony.Veale@hdl.ie, Alan.Conway@hdl.ie

The sign languages used by deaf communities around the world represent a linguistic challenge that natural language researchers have only recently begun to take up. Zardoz is a system which tackles the cross-modal machine-translation problem, translating speech and text into animated sign language. Native sign languages, such as ISL (Ireland), BSL (Britain) and ASL (U.S.A.) have evolved in deaf communities as natural methods of gestural communication. These languages differ from English, not only in modality, but in grammatical structure, exploiting the dimensions of space as well as time. This paper presents an architectural overview of Zardoz, and describes the methods employed to analyse the verbal input and generate the corresponding signed output.

1. Introduction

The fluid articulation of animated sign language from English input represents a unique linguistic challenge of cross-modal translation. There is a sizeable body of sign language users world-wide, for whom such technology can provide valuable tools for education and information access. From a linguistic perspective, the pursuit of cross-modal translation poses new problems in translation and generation, and forces us to question our conceptions of language universals.

This paper describes the architecture and methodology of the ZARDOZ multilingual sign translation system, which is designed to translate spoken language (specifically English text) into a number of different sign-languages, in particular ISL (Irish) , ASL (American) and JSL (Japanese).

The paper has the following structure: Section 2 presents a brief introduction to sign languages which places the problem in context. Section 3 describes the system architecture of ZARDOZ. Section 4 discusses the conceptual Interlingua representation used in translation. Section 5 discusses syntactic generation, while Section 6 addresses articulation and animation issues. Finally Section 7 summarises the present status of the project and our future research goals.

2. Sign Language as a Communication Medium

There is a strong tendency among the speaking community to trivialise the capacity of sign as a full communication medium. It is a common assumption that sign language, being iconic in nature, is a universal language shared by the deaf communities of the world. In fact countries which share the same spoken language (e.g. English in the cases of Britain, Ireland and America) do not necessarily employ the same form of sign (BSL, ISL and ASL respectively). Certainly iconicity plays a stronger role in sign language than sound symbolism does in spoken language but, as with any language, there is a strong tendency to move from iconicity to arbitrariness (see Klima & Bellugi 1979).

A second common misconception is that sign language is a gestural coding of spoken language. While sign languages can be employed for this type of coding (e.g. SEE: Signed Exact English), *native* sign languages possesses a syntax which is independent of any spoken language, and is considerably better adapted to manual communication. Thus there is a genuine translation problem in generating native sign language from English, as well as the obvious articulation problem of generating animated signs.

3. An Overview of the ZARDOZ system

In this section we present an overview of the system architecture of ZARDOZ, a modular system organised around a blackboard control structure (Cunningham & Veale 1991, Veale & Cunningham 1992). This blackboard is built upon the frame-based KR-language KRELL (Veale & Smyth 1992).

A process-oriented view of the system is illustrated in Figure 1, which presents the blackboard compartmentalised into distinct *panels*. Task-specific knowledge agencies (composed of autonomous, write-activated demons) communicate by reading from and writing to these panels.

Taking a clockwise tour around Figure 1, system operation proceeds as follows: (i) incoming text is processed by a swarm of *Lexperts* - lexical expert demons which implement morphological rules and heuristics for recognising compound word constructs. The digested text then undergoes (ii) *idiomatic reduction,* before it is (iii) *parsed* , using a unification grammar, to produce a deep syntactic/semantic representation. From the unification structure a first-cut Interlingua representation is (iv) *composed;* but before this representation can be considered language-independent, metaphoric and metonymic structures specific to the source language are removed by (v) *schematization* (see Section 4). The refined interlingua provides grist for the (vi) *discourse tracking* agency, which does anaphoric resolution, before being passed to the generation panels of the system: (vii) the *sign syntax* agency, which employs a robust scheme of spatial dependency graphs (see Section 5), and (viii) the *sign*

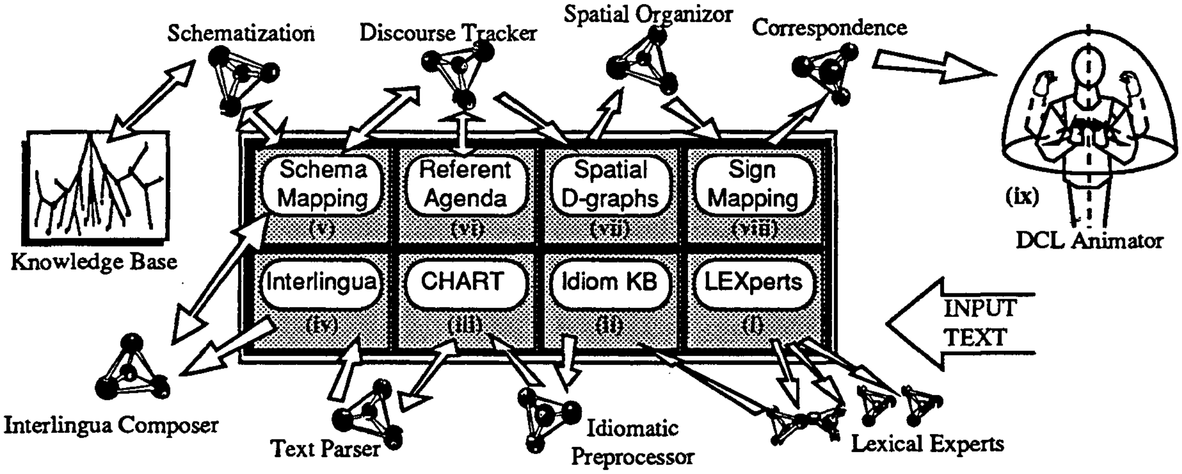


Figure 1: The ZARDOZ Blackboard Architecture.

mapping agency, which assigns concept-to-sign correspondences to the tokens in the interlingua structure. The syntax and mapping agencies transduce the interlingua structure into a flat output stream of sign tokens, which are compiled into a Doll Control Language (DCL) program by (ix) the *DCL animator* . The DCL program controls an on-screen animated doll, causing the correct gesture sequence to be articulated to the end-user.(Conway & Veale 1994).

4. Interlingua and Schematization

To decouple the input and output languages, ZARDOZ adopts an *Interlingua* approach (e.g. Mitamura et al 1991), which places a language-independent interface between source and target. Rather than attempting to construct a *universal grammar* generalising over the syntactic forms of many languages, we take the knowledge based path of modelling sentence meaning in the interlingua. This reflects the origins of ZARDOZ in the TWIG knowledge-acquisition system (Veale & Cunningham 1992).

The first-cut interlingua representation of an utterance is derived compositionally from lexeme-to-concept correspondences. Next *schematization* removes conventional metonymies and metaphors as illustrated in Figure 2, which demonstrates the use of the core English metaphor POSSESSION-AS-ABSTRACT-STATE (see Veale & Keane 1992 for a discussion of the computational treatment of metaphor).

The first-cut representation is the interlingua frame HAVE-0, with the concepts *SPEAKER* and HEADACHE-0 in the slots POSSESSOR and POSSESSION. Next the system looks for the most suitable schema for this frame, using spreading activation from the nodes HAVE, *SPEAKER* and HEADACHE. On finding a suitable schema, SUFFER-FROM-AILMENT, the concepts *SPEAKER* and HEADACHE-0 are re-mapped into the slots SUFFERER and AILMENT.

The importance of the schematization phase can be seen when one considers that ASL has a sign for HAVE (possession), but does not use the metaphor of possession for ailments. Thus a translation from the first-cut representation meaning "I posses a headache" is possible, but incorrect in ASL.

I have a terrible headache

InterLingua

HAVE-0	
Possessor:	*SPEAKER*
Posession:	HEADACHE-0
Tense:	PRESENT
Surface-Form:	ACTIVE-VOICE

Schematization

Suffer-From-Ailment-0	
Sufferer:	*SPEAKER*
Ailment:	HEADACHE-0
Tense:	PRESENT
Surface-Form:	ACTIVE-VOICE

Generation

ASL-ME ASL-INTENSE Forehead::ASL-HURT

Figure 2: Example of Interlingual representation, Schematization, and ASL output

5. Sign Generation: Syntactic Issues

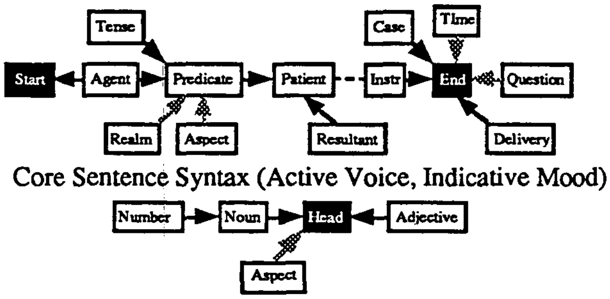
In parsing, structure is imposed upon a flat input stream. Conversely, generation removes structure from a meaning representation to produce a flat output stream. The heart of a generation system is a *linearizer* which selects and orders elements of the meaning representation.

5.1 Spatial Dependency Graphs

In this section we introduce the syntactic framework of *Spatial Dependency Graphs.* An SD-graph is a partial ordering of case types from the syntactic/semantic case ontology, which indicates which elements are to be selected from an interlingua structure, and their relative order in the output stream.

An SD-graph represents a syntactic context, or general *state of affairs*, rather than a rule of grammar; in effect, an SD-graph is a collection of weak rules (or preferences) folded together. Figure 3 depicts the SD-graph representation of some basic ASL syntax. These graphs represent *stand-alone*, syntactic contexts, inasmuch as they are capable of transforming (i.e.

linearizing) an interlingual frame without recourse to additional syntactic information.



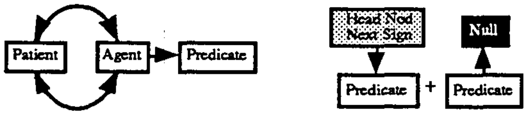
(a) Core Sentence Syntax (Active Voice, Indicative Mood)

(b) Basic Noun Phrase Syntax

*Figure 3: SD-graph of core ASL Sign Syntax.. Key: left to right arrows indicate **Before**; right to left arrows indicate **After**; vertical arrows indicate **Same Position As**; Grey arrows indicate **Closer Proximity**; Grey nodes indicate **Sign Literals** as opposed to constituent types, while black nodes represent the fixed points of the graph)*

An SD-graph is a collection of constraints for ordering the elements of an interlingua frame structure. Following the constraints of Figure 3, the linearizer will place the occupants of the AGENT and ASPECT cases *before* the predicate in the output, but will also ensure that the ASPECT *follows* the AGENT and directly precedes the verb.

As well as stand-alone syntactic contexts there are *augmentative* graphs for syntactic flourishes of surface form in the target language - e.g. passive voice, and verb gapping in ASL (Figure 4). These augmentations are triggered by *style markers* in the SURFACE-FORM slot of each interlingual frame.



(a) Passive Voice (b) Verb Gapping

Figure 4: SD-graphs for augmentations to the core syntax of Figure 3.

When linearizing, the augmentation graph is combined with the core syntax by pooling constraints, giving precedence to the constraints in the augmentation graph. The constraints of the combined graph are instantiated with the contents of the current frame, and resolved relative to the fixed nodes START and END to produce the final linearized ordering.

5.3 Content-Dependent Syntactic Contexts

The graphs of Figures 3 and 4 are *content-independent* , i.e. are applicable to an interlingua frame regardless of its conceptual content. It is often necessary to employ *content-dependent* contexts which are triggered by particular elements of the interlingua structure. Such a context is depicted in Figure 5.

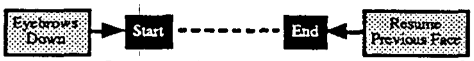


Figure 5: SD-graph for WH-questions in ASL

WH-question words in ASL require an eyebrows downward facial pose for the duration of the interrogative context (see Liddell 1980). Thus the content-dependent context of Figure 5 is associated with ASL-WH-QUESTION in the sign hierarchy, where it is inherited by all ASL wh-question signs. When the target translation contains any member of this class, the wh-question context is invoked to add the appropriate facial features.

5.4 Anaphora and Spatial Designation

Anaphoric resolution is required in translation whenever source and target languages use different anaphoric discrimination systems. For instance, the English "They", neutral in gender, can map onto either of "Elles" or "Ils" in French. It is thus necessary to resolve the reference of a pronoun before translation, so that the correct referring term can be generated in the target language. ZARDOZ employs the basic Hobbs algorithm for this task (see Hobbs 1978), augmented with discourse *registers* which track the movement of referents between peripheral and central focus.

Sign language makes use of *index locations* in space to refer to entities in a conversation. Thus, locations should be allocated sign space in such a way that possible referential conflicts are minimised. ZARDOZ strives to allocate a different locations to the major cases of an utterance (e.g., agent, patient), and to maintain those assignments throughout a narrative as far as is possible.

5.2 Word Order in Sign

Word order, the dominant syntactic constraint in English, has a reduced role in ASL which can also employ the dimensions of space to indicate case roles (see Liddell 1980). The referents of a verb may be established at index locations in signing space, and the direction of movement of the verb between locations then indicates which is the agent and which the patient.

For example, if BILL is signed on the left and MARY on the right, then a left to right motion while signing the verb ASL-CHASE, indicates BILL is the pursuer and MARY the pursued. Thus using the passive voice in ASL is simply a matter of reversing the order of agent and patient. Of course, the verb/predicate will now have to be signed *after* both agent and patient have been articulated. The SD-graph representing this transformation is presented in Figure 4(a).

6. Mode-Interleaved Sign Generation

We have already mentioned the distinction between native sign languages (e.g. ASL) and borrowed sign system (like SEE). Native sign language is the dominant means of communication among deaf signers, however, borrowed sign language is often used for educational purposes (where hearing signers are involved), and for such ends as signed news summaries on television. As a result, most native signers are comfortable with both types of sign, and encounter little difficulty in segueing between the two forms.

This ability affords a system such as ZARDOZ with a base-level performance that can be guaranteed by the system. Should the source parser be unable to generate a syntactic structure that spans the entire input, the system can still produce a full output representation by interleaving native and coded sign. Native sign is used for those input fragments which produce case-frame interlingual representations, while the coded sign is used for troublesome link words which cause the parser to fail.

6. Sign Articulation and Animation

Having generated a *syntactic* form of a sign utterance it remains to create a fluid *articulation* of the utterance to display as an animated sequence. In the case of coded sign systems such as SEE this is a simple matter of mapping sign tokens to pre-stored animation sequences and smoothing over inter-sign gaps. However, native sign languages employ a much richer sign structure, which requires a correspondingly richer representation for the output lexicon.

Native sign languages are heavily inflected, with a much of syntactic information encoded in parallel on a single lexeme. One example already mentioned is the use of verb movement to indicate agent and patient. Another example in ASL is the rich aspectual inflection system employed. For example the sign ASL-SICK with a circular motion added means 'sickly' or 'sickness prone'. A repeated, tense motion indicates the meaning 'extremely sick'. These and many more inflections apply in a regular manner to ASL signs, and may be compounded. For example, the verb ASL-LOOK-AT can be inflected to mean 'he watches it regularly' or 'I look at each of them in turn'.

Because of the richness of the inflection system, it is impractical to store every inflected form directly as an animation sequence. We adopt the approach of storing signs in their *citation* or root form only, and storing inflection rules separately. Inflected signs are generated as needed by applying the appropriate rules to the root sign forms. Signs are stored using a *phonological* model of sign structure, based on Sandler's Hand-Tier model (Sandler, 1989). The phonological representations are not mapped to concrete animation values in a DCL program until after inflection rules have been applied.

7. Summary and Future Research

To date, we have implemented the infrastructure of the Zardoz system, including parsing, interlingua, generation and animation components, but have yet to implement a comprehensive sign grammar or lexicon. The phonological model of sign structure and inflection rules, mentioned in Section 6, is also in an early stage of development.

Our current research efforts are concentrated on developing more a comprehensive computational grammar, morphology and lexicon for ISL, the native sign language of Ireland where our research is based. The examples in this paper are taken from ASL, as linguistic information on ASL is more readily available, but in future work will focus on ISL, as we feel that the evaluation and advice of native signers will be crucial to the success of our research.

Though our work is still in an early stage, we are confident that the framework outlined here will provide a sound basis for tackling the challenges of cross-model translation. The issues of translation between different language media holds considerable theoretical interest, but we also believe that the A.I./linguistic technology is mature enough to build systems of value to sign users in the near future. We hope to contribute to the development of such systems.

References

Conway, A. & T. Veale. (1994). A Linguistic Approach to Sign Language Synthesis, *to be presented at the Human Computer Interaction conference, HCI'94, Glasgow.*

Cunningham, P. & T. Veale. (1991). Organizational issues arising from the integration of the Concept Network & Lexicon in a text understanding System, in *the proceedings of the 12th International Joint Conference on Artificial Intelligence.* Morgan Kaufmann.

Hobbs, J. (1978). Resolving Pronoun References, *Lingua* **44**, p 311-338.

Klima, E. & U. Bellugi. (1979). *The Signs of Language.* Harvard University Press.

Lee, J. & T. L. Kunii. (1992). Visual Translation: from Native Language to Sign Language, in *the proceedings of the IEEE workshop on Visual Languages*, Seattle Washington.

Liddell, S. K. (1980). *American Sign Language Syntax.* Mouton.

Mitamura, T., E. H. Nyberg and J. G. Carbonell. (1991). An Efficient Interlingua Translation System for Multi-lingual Document Production, in *the proceedings of Machine Translation summit III*, Washington D.C., July 2-4, 1991.

Patten, T. & Hartigan, J. (1993). Automatic Translation of English to American Sign Language, presented at *the 1993 National Conference on Deafness*, Columbus Ohio.

Sandler, W. (1989). *Phonological representation of the sign: Linearity and non linearity in American Sign Language.* Foris Publications.

Veale, T. & B. Smyth. (1992). KRELL: Knowledge Representation Entry-Level Language, the User Guide Version 1.0. *Hitachi Dublin Laboratory Technical Report*, HDL-TR-92-051.

Veale, T. & M. T. Keane. (1992). Conceptual Scaffolding: A spatially founded meaning representation for metaphor comprehension, *Computational Intelligence* 8(3), p 494-519.

Veale, T. & P. Cunningham. (1992). Competitive Hypothesis Resolution in TWIG: A Blackboard-Driven Text-Understanding System, in *the proceedings of the 10th European Conference on Artificial Intelligence*, Chichester: John Wiley.