Lost in Translation: the Problems of Using Mainstream MT Evaluation Metrics for Sign Language Translation

Sara Morrissey[1], **Andy Way**

National Centre for Language Technology
School of Computing,
Dublin City University,
Dublin 9, Ireland.

IBM CAS Dublin
Mullhuddart,
Dublin 15, Ireland.

{smorri,away}@computing.dcu.ie

Abstract

In this paper we consider the problems of applying corpus-based techniques to minority languages that are neither politically recognised nor have a formally accepted writing system, namely sign languages. We discuss the adoption of an annotated form of sign language data as a suitable corpus for the development of a data-driven machine translation (MT) system, and deal with issues that arise from its use. Useful software tools that facilitate easy annotation of video data are also discussed. Furthermore, we address the problems of using traditional MT evaluation metrics for sign language translation. Based on the candidate translations produced from our example-based machine translation system, we discuss why standard metrics fall short of providing an accurate evaluation and suggest more suitable evaluation methods.

1. Introduction

Large amounts of money and resources are invested in dominant languages both in terms of linguistic analysis and their machine translation (MT). However, such investment serves to increase the prominence and power of these languages and ignores the less dominant, minority languages (Ó'Baoill & Matthews, 2000). Sign languages are the first languages (L1s) of the Deaf community worldwide and, just like other minority languages, are poorly resourced and in many cases lack political and social recognition.

As with other speakers of minority languages, Deaf people are often required to access documentation or communicate in a language that is not natural to them. In an attempt to alleviate this problem, we propose the development of an example-based machine translation (EBMT) system to allow Deaf people to access information in the language of their choice. The language of choice for us is Irish Sign Language (ISL). While corpus creation for English—ISL is ongoing, and we hope to avail of this data in the near future, in order to seed the development of our EBMT system, we have instead used a corpus of Dutch Sign/Language Nederlandse Gebarentaal (NGT) data created by the ECHO project. The annotation scheme used for NGT is the same as for ISL, so we anticipate that migration to ISL will be reasonably seamless.

In this paper we begin by introducing sign languages (SLs) in section two, briefly discussing their current status. Section three provides an overview of related work in the area of sign language machine translation. Section four reviews writing systems available for SLs. This is followed by a discussion on the annotated format we chose in section five. Section six gives a brief overview of EBMT before describing our own approach. We describe in section seven the experiments carried out on our system, and discuss both their evaluation and the problems with traditional evaluation metrics in section eight. Finally we conclude the paper.

2. Sign Language

SLs are the primary means of communication of the Deaf community worldwide. In SLs, the hands are the main articulators and non-manual features (NMFs) such as eyebrow movement, head tilt, and blinks add vital morphological and grammatical detail. SLs are fully formed natural languages that have developed in such a way that full articulatory use is made of the signing space, i.e. the area in front of the signer from waist to head and the extension of the arms in which discourse is articulated (Ó'Baoill & Matthews, 2000).

Most countries have their own native SL, although some are dialects of more widespread languages. Despite the use of these manual languages as the L1s of Deaf communities, in most cases they lack political recognition and are often not recognised as languages at all. As a result, SLs remain less resourced than spoken languages. This is apparent in the areas of SL linguistics and machine translation of SLs. Both are relatively new areas in comparison with their spoken language counterparts. Significant SL linguistic research began about 45 years ago with the work of (Stokoe, 1960) and the earliest papers on research into the machine translation of SLs date back only approximately 15 years.

In Ireland, Irish Sign Language (ISL) is the dominant language of the Deaf Community. As with other SLs, it is grammatically distinct from spoken languages. Despite being in use in Ireland since the 1800s, its status has remained low and a standardized form of the language is not taught to children in Deaf schools in the same way that English is in spoken language schools. The development of the language is

[1] This work is generously sponsored by a joint IBM-IRCSET scholarship.

slow as a result of "its users' lack of access to technical, scientific and political information" (Ó'Baoill & Matthews, 2000).

NGT is the SL of the corpora we use for translation. NGT is the primary language of the Deaf community in the Netherlands with a population of approximately 15,000 deaf. Similar to ISL, NGT was originally derived from French Sign Language and, as is the case in Ireland and many other countries, it has not attained recognition as an official language (Gordon, 2005).

We hope that the development of our MT system with first an NGT, then ISL corpus will help to raise the status of SLs in these countries and facilitate communication of information to the Deaf community in their preferred language.

3. Related Work

Many different approaches have been applied to sign language machine translation (SLMT). As might be expected, most approaches have concentrated on translating from spoken languages to SLs.

3.1. Traditional 'Rule-Based' Approaches

Given that SLMT has been tackled only quite recently, most approaches to date are 'second generation', namely transfer- or interlingual-based.

Many transfer-based translation methodologies have been used. (Grieve-Smith, 1999) uses the domain of weather reporting and uses a literal orthography to represent American Sign Language (ASL) for translation into English by mapping the syntactic structure of one on to the other. No evaluation methods have been used in his work.

Other transfer approaches have been applied in (Marshall & Sáfár, 2002; Sáfár & Marshall, 2002). Their work employs discourse representation structures to represent the internal structure of linguistic objects then uses HPSG semantic feature structures for the generation of ASL. No automatic or manual evaluation is discussed

A more syntax-based transfer approach is described in (Van Zijl & Barker, 2003) in their translation work from English to South African Sign Language. Their focus is on producing a signing avatar for manual evaluation at a later stage.

Interlingual SLMT methodologies have also been employed that use language-independent intermediate representations as the basis of their translation. (Veale et al., 1998) developed the ZARDOZ system, a multilingual sign translation system for English to Irish, American and Japanese Sign Languages using this approach. (Zhao et al., 2000) used an interlingual approach for translating English to ASL and employed synchronized tree-adjoining grammars. Evaluation metrics are not mentioned for either interlingual approach.

(Huenerfauth, 2005) attempts to combine the two previous approaches, transfer and interlingual, with a more simplistic direct approach to create a hybrid "multi-path" approach. This system translates English to ASL using first an interlingual method, then failing that, a transfer then direct approach. His work concentrates on the translation of classifier predicates and will be manually evaluated by native signers.

3.2. Corpus-Based Approaches

The first statistical approach that we are aware of was that of (Bauer et al., 1999), but this is the only model we have come across where translation is not from spoken to sign language. Their approach consists of a video-based recognition tool for a lexicon of 100 signer-dependent German Sign Language signs and a translation tool composed of a translation and language model, which is standard in the statistical MT (SMT) paradigm.

An SMT model for spoken to sign language is described in (Bungeroth & Ney, 2004, 2006) to translate German weather reports into German Sign Language using HamNoSys (Prillwitz, 1989) notation. Their initial experiments are automatically and manually evaluated and show promising results for a data-driven approach.

The first Example-based approach is our own model in (Morrissey & Way, 2005). Using an NGT corpus, we developed a prototype system for translating English and Dutch into NGT. Although traditional evaluation metrics were not employed, through manual analysis of a set of experiments we show that encouragingly good translations were obtained.

4. Writing Systems

When applying EBMT techniques to SLs, the lack of recognition and under-resourcing of SLs, together with their having no formal or widely used writing system, make SL corpora difficult to find. Attempts have been made to develop notation systems for these visual languages, examples of which include Stokoe Notation, HamNoSys and SignWriting.

4.1. Stokoe Notation

Stokoe notation (Stokoe, 1960) was developed in the 1960s for ASL and initially described three factors to be taken into account for SL description, namely *tabulation,* referring to the location of a sign; *designator,* referring to the handshape; and *signation,* referring to the type of movement articulated. SL-specific additions by international linguists over the years including the addition of a fourth factor *orientation,* describing the orientation of the handshape, have resulted in no universally accepted version of the Stokoe notation system (Ó'Baoill & Matthews, 2000). While this approach describes a comprehensive analysis of an SL, the method is data-heavy and not practical for use as a writing system for Deaf people. Furthermore there are no large corpora available in this format for use in a data-driven MT system.

4.2. HamNoSys

Another explicit notation system for SLs is the Hamburg Notation System (HamNoSys) (Prillwitz, 1989) that uses a set of language-independent symbols to iconically represent the phonological features of SLs (Ó'Baoill & Matthews, 2000). This system allows even more detail than that of the Stokoe system to be described including NMFs and information about the signing space. For reasons similar to those above, this system is not suitable for adoption by the Deaf community as a writing system and again, no large SL corpora are available in this format.

4.3. SignWriting

An alternative method was developed in (Sutton, 1995) called SignWriting.[2] This approach also describes SLs phonologically but, unlike the others, was developed as a handwriting system. Symbols that visually depict the articulators and their movements are used in this system, where NMFs articulated by the face (pursed lips, for example) are shown using a linear drawing of a face. These simple line drawings make the system easier to learn as they are more intuitively and visually connected to the signs themselves. The SignWriting system is now being taught to Deaf children and adults worldwide as a handwriting version of SLs. The system is not yet widely used but its usability and the rate at which it is being adopted suggests that corpora may be available in the near future on suitable topics for MT.

4.4. Manual Annotation

One way around the problems with writing systems is to manually annotate SL video data. This approach involves transcribing information taken from a video of signed data. It is a subjective process where a transcriber decides the level of detail at which the SL in the video will be described. These categories can include a gloss term of the sign being articulated by the right and left hands (e.g. HARE if the current sign being articulated is the sign for the animal *hare*), information on the corresponding NMFs, if there is repetition of the sign and its location. The annotations are time-aligned according to their articulation in the video. As the process is subjective, the annotation may be as detailed or as simple as the transcriber or project requires. On the one hand, this makes annotations suitable for use with corpus-based MT approaches as they are not loaded with linguistic detail and can provide gloss terms for signs that facilitate translation from and into spoken language. On the other, however, the problem of inter-annotator agreement remains; discrepancies in the training data will hinder the capacity of the corpus-based MT system to make the correct inferences.

5. Annotated Corpora for EBMT

A prerequisite for any data-driven approach is a large bilingual corpus aligned at sentence-level from which to extract training and testing data. For translation between major spoken languages, such data is available in large amounts: in the recent OpenLab[3] evaluation, we used almost 1 million aligned Spanish—English sentence-pairs from the Proceedings of the European Parliament (Koehn, 2005) to seed our MaTreX system (Armstrong et al., 2006). While this is the largest EBMT system published to date, many SMT systems use much larger training sets than this, e.g. the Chinese-English SMT system of (Vogel et al., 2003) is trained on 150 million words.

5.1. Dutch Sign Language (NGT) Corpora

As discussed above, finding corpora suitable for the task we are confronted with in this paper can be difficult. However, a collection of annotated SL data—albeit on a much smaller scale than the training sets typical of data-driven approaches—has been made available through the ECHO project.[4] This EU-funded scheme, based in the Netherlands, has made fully annotated digitised corpora for Dutch Sign Language (NGT: Nederlandse Gebarentaal) available on the Internet. The corpora have been annotated using the ELAN annotation software.[5]

ELAN provides a graphical user interface in which corpora can be viewed in video format with their corresponding aligned annotations (cf. Figure 1). The name of the annotation category tiers may be seen in the column on the left and the time-aligned annotations for each tier are displayed horizontally in line with the tiers.

Annotation has been included that displays a time-aligned translation in the native spoken language and in English. Further annotation groupings include a gloss in both spoken languages of the signs of both hands and various NMF descriptions. An example of some annotations used in the NGT corpus can be found in (1) (where numbers indicate time frame of annotation):

(1)
```
3:09:500 3:10:380
(Gloss RH/LH English) TINY CURLS

3:09:500 3:10:380
(Gloss RH/LH) PIJPENKRULLEN

3:09:500 3:10:380
(Repetition) u

3:09:740 1461310
(Eye Gaze) l, d
```

Such suitably annotated corpora can be reasonably useful for an example-based approach to SLMT. Accompanying English and Dutch translation tiers and time-aligned annotations allow for easy alignment of corpora on a sentential level. The presence of time frames for each annotation also aids in the aligning of annotations from each annotation tier to form chunks that can then be aligned with chunks derived from the English/Dutch tier. As simultaneity (articulators signing two separate signs at the one time) and co-articulation (articulation of one sign being influenced by its neighbouring signs) are prevalent in natural signing, time-aligned annotations help tackle this issue by providing time boundaries to signs and NMFs so that each annotation remains complete and separate. As it is these annotations that are used in the translation output, once a satisfactory boundary width has been established, the issue of separating co-articulated words is removed automatically.

While we were grateful to avail of the ECHO data, there are two main problems with it: firstly, the data consists of annotated videos of two versions of Aesop's Fables and an NGT poetry file—this is hardly the most suitable genre for *any* MT system. Secondly, despite combining all NGT data files available, the corpus amounted to a mere 40 minutes of data, or just 561 sentences. This small corpus size obviously results in data sparseness; for any data-driven approach, the larger

[2] http://www.signwriting.org
[3] http://www.tc-star.org/openlab2006/
[4] http://www.let.kun.nl/sign-lang/echo/data/html
[5] http://www.mpi.nl/tools/elan.html

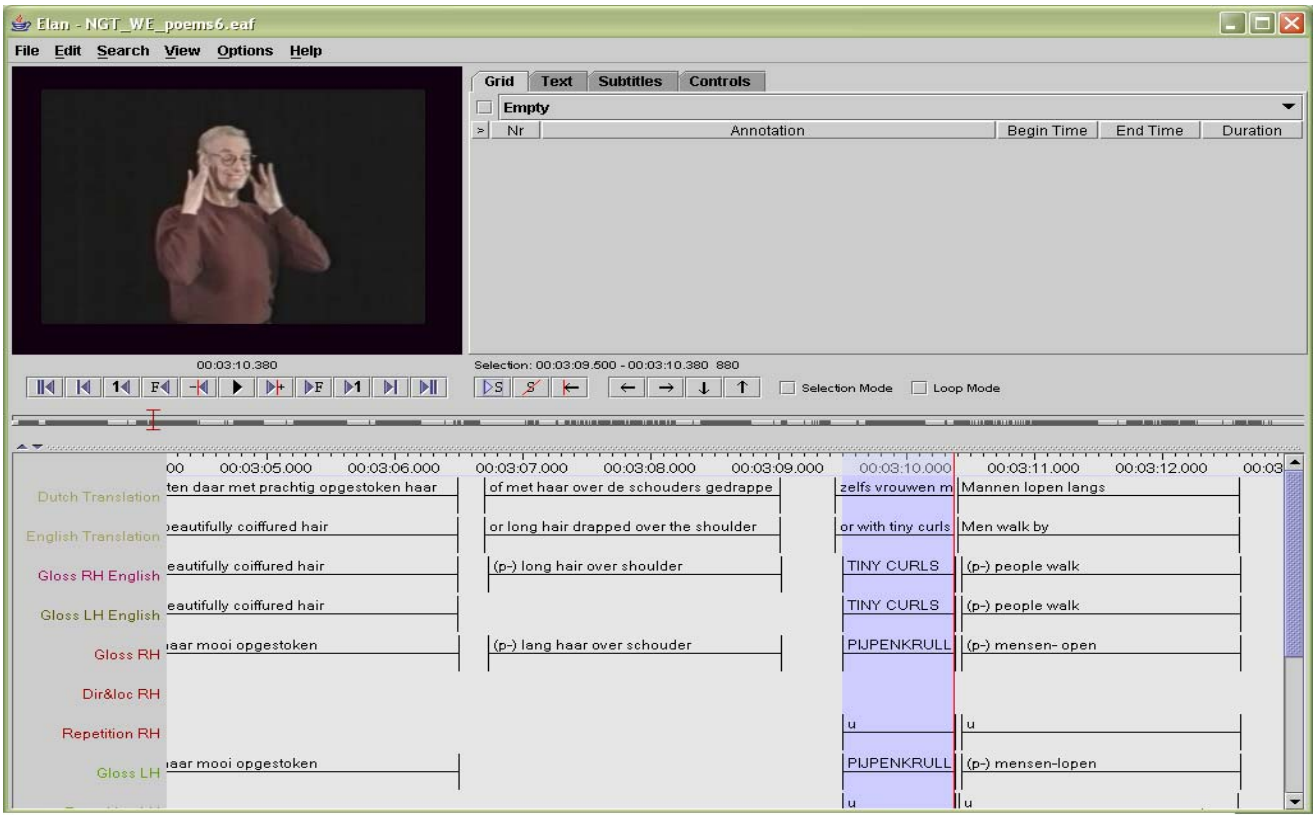

Figure 1 ELAN user interface

the amount of training data available, the greater the set of sub-sentential alignments that can be created. This provides a larger scope for finding translation matches for input string, which correspondingly increases the chances of improving system output. The ECHO project team have funding to increase their corpus creation by 2008, so we hope to increase our training data when this becomes available.

5.2. Irish Sign Language Corpora

Currently, a large annotated corpus of ISL data is under construction. The Centre for Deaf Studies[6] in Dublin is in the process of annotating a corpus of approximately 40 hours of ISL data. However, the subject matter of the ISL data is similar to that of the NGT data, namely stories and conversation. While the larger amount of training data will help increase the translation quality of our system, we will still be confronted with 'unsuitable' data.

A more suitable corpus which would have a practical use for the Deaf is, for example, the area of travel information. In airports and train stations, announcements of changes in travel information are usually announced over a PA system; often such information does not appear on the information screens until a later stage if at all. For this reason many Deaf people find themselves uninformed about changes to schedules etc. through no fault of their own. In many airports and train stations worldwide travel information is entered into a system that announces the changes in an electronic voice. It is quite possible that this system could be extended to accommodate SLs. The limited range of statements and information used in these circumstances could be compiled into a corpus and the information that is announced could be translated into sign language and displayed on video screens for the Deaf to view. We are negotiating with our local airport authority to obtain such data in order to compile a corpus of commonly used announcements suitable for the seeding of our EBMT system's memories.

6. Marker-Based EBMT

An example-based approach necessitates a set of sentences aligned in the source and target languages. Three processes are used to derive translation for an input string:

1. Searching the source side of the bitext for 'close' matches and their translations;
2. Determining the sub-sentential translation links in those retrieved examples;
3. Recombining relevant parts of the target translation links to derive the translation.

The methodology employed in our system is to make use of the 'Marker Hypothesis' (Green, 1979). Here closed class words are used to segment aligned source and target sentences and to derive an additional set of lexical and phrasal resources. In a pre-processing stage, (Gough & Way, 2004b) use 7 sets of marker (or closed class) words for English and French (e.g. determiners, quantifiers, conjunctions etc.) to segment the text into chunks which together with cognate matches and mutual information scores are used to derive three new data sources: sets of marker chunks, generalised templates and a lexicon.

[6] http://www.tcd.ie/Deaf_Studies/

Within our system, sentential alignments are extracted using the time-aligned borders of the English/Dutch translation tiers and grouping all annotations within those time frames together to form the corresponding SL sentence. The English/Dutch sentences are segmented on the basis of closed class words. As an example, consider the sentence in (2), from (Gough & Way, 2004a):

(2) `The first part of the book describes the components of the desktop.`

This string is automatically tagged with marker words, as in (3):

(3) `<DET> The first part <PREP> of the book describes <DET> the components <PREP> of the desktop.`

Given the tagged strings in (3), the marker chunks in (4) are automatically generated:

(4) (a) `<DET> The first part`
(b) `<PREP> of the book describes`
(c) `<DET> the components`
(d) `<PREP> of the desktop`

By generalising over the marker chunks we produce a set of marker templates. This is achieved by replacing the marker word by its relevant tag. From the examples in (4), we can produce the generalized templates in (5):

(5) (a) `<DET> first part`
(b) `<PREP> the book describes`
(c) `<DET> components`
(d) `<PREP> the desktop`

These templates increase the robustness of the system and make the matching process more flexible.

A different approach is used on the sign language side of the corpus. The annotations are segmented according to the NGT gloss time divisions and other corresponding annotations within the same time frame are grouped with that gloss to form a chunk. We therefore segment the sign language corpus into concept chunks to match the content of the English chunks. The example below demonstrates segments from both data sets (English (6) and NGT (7)) and their usability for chunk alignment:

(6) `<CONJ> or with tiny curls`

(7)
```
<CHUNK>
(Gloss RH English) TINY CURLS
(Gloss LH English) TINY CURLS
(Repetition LH) u
(Repetition RH) u
(Eye gaze) l,d
```

Despite the different methods used, they are successful in forming potentially alignable chunks. Both chunks indicate the possession of "tiny curls", articulated by the words in the English chunk and the right and hand left hand in the sign chunk. Extra information is added in by the NMFs *repetition* and *eye gaze*. Repetition shows that the left and right hands signs are articulated a number of times, the 'u' indicates uncountable repetitions in a wiggling manner showing the plurality of the sign, i.e. many curls. The eye gaze 'l,d' indicates that the gaze of the signer goes from the left of the signing space (where the curls start at the signer's head) downwards, following the movements of the hands. This is an important feature in SLs. The gaze of the signer usually follows the movement of the main articulators. Eye gaze is also used to indicate distance of an object in relation to the signer. If eye gaze was not taken into account vital information on the location of objects in the signing space or the distance of objects from the signer would be lost.

7. Experiments and Results

We extracted 561 English— and Dutch—NGT sentence pairs. In order to provide an indication of data complexity, the English translations had an average sentence length of 7.89 words (min. 1 word per sentence, max. 53).

We began testing the system for translation of English and Dutch into NGT. The data was divided into an approximate 90:10 training-testing splits, with 55 randomly selected sentences withheld for testing purposes. Each test sentence is entered into the system and a translation produced based on best matches found at a sentential, sub-sentential (chunk) or word level.

Manual examination of the output showed that the system performed reasonably well and appeared to have correctly translated most of the central concepts in the sentences (Morrissey & Way, 2005). However, annotations can be complex and it is difficult for an untrained eye to discern the correctness of the output. Furthermore, due to the subjectivity and varying format of the annotations, there lacks a 'gold standard' against which they may be formally evaluated using traditional MT evaluation metrics.

In light of this issue, we chose to reverse the translation process taking in annotations as input and producing either English or Dutch output. Output into spoken language takes the form of written text and output in sign language takes the form of grouped annotations.

While reversing the directionality of translation enables automatic evaluation metrics to be used, the exercise is quite artificial in that there is little or no demand for translation from SL to spoken language. Of course, situations can be envisaged where this might be useful, e.g. in a post office, a Deaf customer could ask for stamps by signing into a camera and having it translated into text/speech for the hearing sales assistant, while the process could be reversed for communicating the information from the sales assistant to the Deaf person via a signing avatar.

From the change in direction we were able to obtain evaluation scores for the output as we had reference translations against which to measure the output. However, as SLs by their very nature do not contain closed class lexical items, the output was sparse in terms of lexical data and rich only with respect to content words. This resulted in decidedly low evaluation scores. In an attempt to improve these scores we experimented with inserting the most common

marker word (*the* in English and *de* in Dutch) into the candidate translations in what we determined to be the most appropriate location, i.e. whenever an INDEX was found in the NGT annotations indicating a pointing sign to a specific location in the signing space that usually refers back to an object previously placed there. This was an attempt to make our translations resemble more closely the gold standard.

7.1. Automatic Evaluation Metrics

The system was evaluated for the language pair NGT—English using the traditional MT evaluation metrics BLEU (Papineni et al., 2002), SER, WER and PER. BLEU score is a precision-based metric that compares a system's translation output against reference translations by summing over the 4-grams, trigrams, bigrams and unigram matches found divided by the sum of those found in the reference translation set. It produces a score for the output translation of between 0 and 1. Sentence Error Rate (SER) computes the percentage of incorrect full sentence matches. Word Error Rate (WER) computes the distance between the reference and candidate translations based on the number insertions substitutions and deletions in the words of the candidate translations divided by the number of correct reference words. The Position-independent word Error Rate (PER) computes the same distance as the WER without taking word order into account. With all error rates, a lower percentage score indicates better candidate translations.

7.2. NGT—English Results

For the 55 test sentences, the system obtained a SER of 96%, a PER of 78% and a WER of 119%.[7] Due to the lack of closed class words produced in the output, no 4-gram matches were found, so the system obtained a BLEU score of zero. Ongoing experiments using the 'Add-One' ploy of (Lin & Och, 2004) will circumvent the 'Zero-BLEU' problem described here. An example of the candidate translation capturing the central content words of the sentence may be seen in (8) compared with its reference translation in (9).

(8) `mouse promised help`

(9) `'You see,' said the mouse, 'I promised to help you'.`

Here it can be seen that our EBMT system includes the correct basic concepts in the target language translation, but for anyone with experience of using automatic evaluation metrics, the 'distance' between the output in (8) and the 'gold standard' in (9) will render the quality to be scored very poorly.

In the next section, we hypothesize whether a different evaluation metric might be more useful, both for SLMT, but also for MT as a whole.

8. Discussion of the Evaluation Process

As shown in the previous section, we struggled to use mainstream MT evaluation metrics such as BLEU, WER and PER, albeit in a rather artificial exercise. Of the related work mentioned in section 3, only the translations produced by (Bungeroth & Ney, 2005) have been evaluated using these metrics. The standard evaluation technique applied to SLMT seems to be a manual assessment by native and non-native signers.

We contend that in general, the traditional string-based metrics are inappropriate for the evaluation of SLMT systems, where the primary goal is translation from an oral to a non-oral language, as there is no 'gold standard' underlying sign language annotation available.

A typical annotation taken from our corpus was shown in (7). For our purposes, we concentrate mostly on the 'GLOSS' field, but other relevant information appears in other fields too, such as lip rounding, puffing of the cheeks etc. The absence of the semantic information provided by these NMFs affects the translation and thus the evaluation scores, so we intend to incorporate this information into the system in the next phase of development.

Our experiments were further hampered by the fact that we were generating root forms from the underlying GLOSS, so that a lexeme-based analysis of the gold standard and output translations via a morphological analysis tool might have had some positive impact. This remains an avenue for future research.

Subsequent experiments to (i) insert the most common marker word corresponding to the appropriate marker tag (to make our translations resemble more closely the 'gold standard', and (ii) delete marker words from the reference translations (to make them closer to the translations output by our system) had little effect on overall BLEU score.

In fact, we have come to the conclusion that rather than continue to attempt these 'transformations' in order to try to reconcile the differences between the reference and candidate translations, we would in fact be better off developing automatic MT evaluation metrics that were more suitable to sign language data

One measure that might have some promise is evaluation on the level of syntactic (or, even better, semantic) relations. For example, compare the (invented) reference and candidate translations in (10):

(10) *Reference*: `I went to the shops yesterday.`
Candidate: `Yesterday I went to the shops.`

Despite being a 'perfect' translation in many ways, the candidate translation in (10) obtains a BLEU score of just 0.669. However, at the level of syntactic relations, the sentences in (10) are identical.

One method of evaluating the 'goodness' of translations would be at the level of syntactic dependencies, rather than by measuring the distance between two strings. Dependency parsers for many languages exist already; two examples that one of the authors has been involved in are the LFG parsers of (Cahill et al., 2004) for English, and (Cahill et al., 2005) for German. New strings are parsed using a variety of PCFG-based LFG parsers, and LFG trees and f-structures are produced. Gold standard sets of f-structures exist (e.g. the PARC-700), and reference and

[7] It is possible to obtain a WER of more than 100% if there are fewer words in the reference translations than in the candidate translations.

system-generated f-structures can be compared and evaluated using F-score. An alternative means of evaluation would be to read the semantic forms (subcategorisation frames) off the f-structures generated using the method of (Ó'Donovan et al., 2005) and compare the 'predicate(filler, arg)' triples. Given examples such as (8), for example, it might be sufficient for our purposes to evaluate only the 'PRED' (or headword) triples.

All this remains for further research, and is outside the scope of this paper, but we are quite confident that such an evaluation would be more useful not only for sign language MT, but for all models of translation. Clearly, in addition, human evaluation remains crucial for all such approaches.

9. Conclusions

We have described ongoing work on our EBMT system to translate between oral and sign languages. Like other minority languages, SLMT suffers from the lack of suitable corpora for the training of corpus-based models of translation. In order to bootstrap the system, we have used 561 English— and Dutch—NGT sentence pairs from the ECHO corpus.

Despite the subjective nature of the corpus and its size, the availability and ease of use of the annotations facilitates speed of development of such an SLMT system. Were a larger corpus to be made available in another SL, the approach described above could easily be applied.

One disadvantage of a corpus-based approach, as discussed in this paper, is its evaluation. No 'gold standard' is available for evaluating candidate translations in SL and the metrics used for evaluating the English/Dutch output fall short of recognising that the candidate translations capture the essence of the sentence. We are confident that a syntactic- or semantic- based evaluation metric would better reflect the performance of an SLMT system while at he same time providing an improved evaluation approach for written languages.

References

Armstrong, S., D. Groves, M. Flanagan, Y. Graham, B. Mellebeek, S. Morrissey, N. Stroppa, and A. Way. (2006). The MaTreX System: Machine Translation Using Examples. Available at: `http://www.tc-star.org/openlab2006/day1/Groves_openlab.pdf`

Bauer, B., S. Nießen and H. Heinz. (1999). Towards an Automatic Sign Language Translation System. In *Proceedings of the International Workshop on Physicality and Tangibility in Interaction: Towards New Paradigms for Interaction Beyond the Desktop*, Siena, Italy.

Bungeroth, J. and H. Ney (2004). Statistical Sign Language Translation. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages (LREC 04)*, Lisbon, Portugal.

Cahill, A., M. Forst, M. Burke, M. McCarthy, R. O'Donovan, C. Rohrer, J. van Genabith and A. Way. (2005). Treebank-Based Acquisition of Multilingual Unification Grammar Resources. *Journal of Language and Computation: Special Issue on Shared Representations in Multilingual Grammar Engineering*, pp.247—279.

Cahill, A., M. Burke, R. O'Donovan, J. Van Genabith and A. Way. (2004). Long-Distance Dependency Resolution in Automatically Acquired Wide-Coverage PCFG-Based LFG Approximations. In *ACL-04: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain, pp.319—326.

Gordon, R. G., Jr. (ed.), (2005). *Ethnologue: Languages of the World, Fifteenth edition*. Dallas, Tex.: SIL International.

Gough, N. and A. Way. (2004a). Example-Based Controlled Translation. In *Proceedings of 9th EAMT Workshop*, Valetta, Malta, pp.73—81.

Gough, N. and A. Way. (2004b). Robust Large-Scale EBMT with Marker-Based Segmentation. In *Proceedings of the Tenth Conference on Theoretical and Methodological Issues in Machine Translation (TMI-04),* Baltimore, MD., pp.95—104.

Green, T. (1979). The Necessity of Syntax Markers. Two experiments with artificial languages. *Journal of Verbal Learning and Behavior* **18**:481—496.

Grieve-Smith, A.B. (1999). English to American Sign Language Machine Translation of Weather Reports. In D. Nordquist (ed.) *Proceedings of the Second High Desert Student Conference in Linguistics (HDSL2),* Albuquerque, NM., pp.23—30.

Huenerfauth, M. (2005). American Sign Language Generation: Multimodal NLG with Multiple Linguistic Channels. In *Proceedings of the ACL Student Research Workshop (ACL 2005)* Ann Arbor, MI., pp.37—42.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *MT Summit X*, Phuket, Thailand, pp.79—86.

Lin, C-Y. and F.J. Och. (2004). ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, pp.501—507.

Marshall, I. and É. Sáfár. (2002). Sign Language Generation using HPSG. In *Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-02)*, Keihanna, Japan, pp.105—114.

Morrissey, S. and A. Way. (2005). An Example-based Approach to Translating Sign Language. In *Proceedings of the Workshop in Example-Based Machine Translation (MT Summit X)* Phuket, Thailand, pp. 109—116.

Ó'Baoill, D. and P.A. Matthews. (2000). *The Irish Deaf Community (Volume 2): The Structure of Irish Sign Language*. The Linguistics Institute of Ireland, Dublin, Ireland.

O'Donovan, R., M. Burke, A. Cahill, J. van Genabith and A. Way. (2005). Large-Scale Induction and Evaluation of Lexical Resources from the Penn-II and Penn-III Treebanks. *Computational Linguistics* **31**(3):329—365.

Papineni, K., S. Roukos, T. Ward and W. Zhu. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, Philadelphia, PA., pp.311—318.

Prillwitz, S. (1989) *HamNoSys Version 2.0; Hamburg Notation System for Sign Language. An Introductory Guide.* Signum Verlag.

Sáfár, É. and I. Marshall. (2002). The Architecture of an English-Text-to-Sign-Languages Translation System. In *Proceedings of Recent Advances in Natural Language Processing (RANLP-01)*, Tzigov Chark, Bulgaria, pp.223—228.

Stein, D., J. Bungeroth and H. Ney (2006). Morpho-Syntax Based Statistical Methods for Automatic Sign Language Translation. In *Proceedings of 11th EAMT Annual Conference,* Oslo, Norway.

Stokoe, W.C. (1960). An Ouline of the Visual Communication Systems of the American Deaf. In *Studies in Linguistics: Occasional papers, No. 8*, Department of Anthropology and Linguistics, University of Buffalo, Buffalo, NY., [Revised 1978 Lincoln Press].

Sutton, V. (1995). Lessons in Sign Writing, Textbook and Workbook (Second Edition). The Center for Sutton Movement Writing, Inc.

Van Zijl. L. and D. Barker. (2003). South African Sign Language Machine Translation System. In *Proceedings of the Second International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction in Africa (ACM SIGGRAPH)*, Cape Town, South Africa, pp.49—52.

Veale, T., A. Conway and B. Collins. (2000). The Challenges of Cross-Modal Translation: English to Sign Language Translation in the Zardoz System. *Machine Translation* **13**(1):81—106.

Vogel, S., Y. Zhang, F. Huang, A. Tribble, A. Venugopal, B. Zhao and A. Waibel. (2003). The CMU Statistical Machine Translation System. *MT Summit IX*, New Orleans, LA., pp.402—409.

Zhao, L., K. Kipper, W. Schuler, C. Vogler, N. Badler, and M. Palmer. (2000). A Machine Translation System from English to American Sign Language. In *Envisioning Machine Translation in the Information Future: Proceedings of the Fourth Conference of the Association for Machine Translation (AMTA-00)*, Cuernavaca, Mexico, pp.293—300.