Data-Driven Machine Translation for Sign Languages

Sara Morrissey

BSc.

A dissertation submitted in fulfilment of the requirements for the award of

Doctor of Philosophy (Ph.D.)

to the

Dublin City University
School of Computing

Supervisor: Prof. Andy Way

April 2008

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Ph.D. is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed:

(Candidate) ID No.:

Date:

Contents

Abstract **ix**

Acknowledgements **x**

1 Introduction **1**

2 Sign Languages **7**

2.1 Background 7

2.2 Sign Language Linguistics 10

2.2.1 Articulation and Structure 10

2.2.2 Non-Manual Features 11

2.2.3 Signing Space 11

2.2.4 Classifiers 13

2.2.5 Irish Sign Language 14

2.3 Transcription and Notation Systems 16

2.3.1 Stokoe Notation 17

2.3.2 HamNoSys 18

2.3.3 SignWriting 19

2.3.4 Manual Glossing and Annotation 20

2.4 Summary 24

3 Overview of Sign Language Machine Translation **25**

3.1 Background 26

3.2 Rule-based Approaches 27

3.2.1 The Zardoz System 29
3.2.2 Albuquerque Weather 31
3.2.3 The ViSiCAST Project 33
3.2.4 The TEAM Project 35
3.2.5 South African SL MT 36
3.2.6 The ASL Workbench 38
3.2.7 Spanish SL MT 40
3.2.8 A Multi-path Approach 43
3.3 Data-driven Approaches 45
3.3.1 Translating from German SL 48
3.3.2 RWTH Aachen University Group 50
3.3.3 Chinese SL MT 52
3.4 Summary 55

4 Problems for Sign Language Machine Translation 59

4.1 Data 59
4.1.1 Signs of Ireland Corpus 61
4.1.2 American Sign Language Linguistic Research Project Corpus . 62
4.1.3 European Cultural Heritage Online Project Corpus 63
4.1.4 Choosing Suitable Data for SL MT 65
4.2 Evaluation 67
4.2.1 Automatic Evaluation 68
4.2.2 Manual Evaluation 70
4.3 A Prototype EBMT System for Sign Language Machine Translation Illustrating Data and Evaluation Problems 72
4.3.1 Dataset Construction 72
4.3.2 System Description 72
4.3.3 Experiment One: Assessing the Suitability of a Small Dataset 77
4.3.4 Experiment Two: Assessing Evaluation 81

4.4 Summary 85

5 Data-Driven Sign Language Machine Translation: Experiments Using the MaTrEx MT System 88

5.1 Data 89

5.1.1 Data Translation 90

5.1.2 Data Annotation 91

5.1.3 Testing the New Data 95

5.2 The MaTrEx System 96

5.2.1 Word Alignment Module 97

5.2.2 Chunking Module 98

5.2.3 Chunk Alignment Module 98

5.2.4 Decoder 99

5.2.5 Non-SL Experiments using MaTrEx 100

5.3 Experiments 100

5.3.1 Translating Sign Language Gloss to Spoken Language Text . . 101

5.3.2 Translating Sign Language Video to Spoken Language Text . 113

5.3.3 Translating Sign Language to Synthetic Speech 116

5.3.4 Translating Spoken Language Text to Sign Language Gloss . . 117

5.3.5 Translating Spoken Language Text to Animated Sign Language118

5.4 Summary 118

6 Creating and Evaluating Sign Language 122

6.1 Generating Real Sign Language 123

6.1.1 Real Human Signing 123

6.1.2 Animated Avatar Signing 124

6.2 Previous Approaches to Sign Language Generation 125

6.2.1 The Zardoz System 125

6.2.2 The ViSiCAST Project 127

6.2.3 The TEAM Project 128

6.2.4 The South African SL Project 129
6.2.5 Spanish SL Project 130
6.2.6 The Multi-Path Approach 131
6.2.7 RWTH Aachen University Group 132
6.2.8 Comparing Previous Animation Approaches 134
6.3 Our Approach to Sign Language Animation 135
6.3.1 Animation Software 135
6.3.2 Creating the Animations 138
6.3.3 Comparing Robert Animations with Previous Approaches . . 140
6.4 Human Evaluation of Machine Translation 141
6.4.1 Judging Criteria 142
6.4.2 Scoring Scales 143
6.4.3 Number of Evaluators 143
6.5 Manual Evaluation of Sign Language Video Output from the MaTrEx System 144
6.5.1 Evaluation Methodology 144
6.5.2 Evaluation Interface and Format 146
6.6 Manual Evaluation Results 149
6.6.1 Intelligibility 150
6.6.2 Fidelity 152
6.6.3 Comparing Intelligibility and Fidelity Scores 154
6.6.4 Questionnaire 155
6.7 Summary 157

7 Conclusions 161

7.1 Future Work 165

Bibliography 167

List of Figures

2.1 Diagram of Signing Space . . . 12

2.2 Example of Classifier Usage in Chinese . . . 13

2.3 Example of Stokoe Notation: *don't know* . . . 17

2.4 Example of HamNoSys: *nineteen* . . . 19

2.5 Example of SignWriting: *deaf* . . . 20

2.6 An Example of the ELAN Interface Tool for Manual Glossing and Annotation of SL Video Data . . . 21

3.1 The Vauquois Pyramid . . . 28

3.2 The Zardoz System Architecture . . . 30

3.3 The Albuquerque Weather System Architecture . . . 31

3.4 The ViSiCAST System Architecture . . . 34

3.5 The TEAM Project Architecture . . . 35

3.6 The SASL System Architecture . . . 37

3.7 An Example of Feature Specification Fields for Movement-Hold Segments . . . 39

3.8 The ASL Workbench Architecture . . . 40

3.9 The Spanish SL MT System Architecture . . . 41

3.10 The Multi-path Approach Architecture . . . 43

3.11 An Example of the Partition/Constituent Methodology . . . 44

3.12 The DGS SL Recognition and MT System Architecture . . . 48

3.13 The DGS MT System Architecture . . . 49

3.14 The Chinese SL MT Architecture . . . 53

4.1 Screenshot of the SignStream Annotation Interface 63
4.2 Screenshot of the ELAN Annotator Interface 64
4.3 Outline of the Translation Process of Our Prototype EBMT system . 73

5.1 Screenshot of the ELAN Annotation Interface with ISL Video and Accompanying Annotations 92
5.2 The MaTrEx Architecture 97
5.3 MaTrEx Translation Directions 101
5.4 Graph Comparison of best RWTH results with MaTrEx results . . . 113

6.1 The Zardoz Animation Architecture 125
6.2 The ViSiCAST Avatar 127
6.3 The TEAM Project Avatar 128
6.4 The SASL Nancy Avatar 129
6.5 The Spanish Sign Language Project Avatar 130
6.6 The Multi-Path Approach Avatar 131
6.7 Average Scores of Multi-path Animations for Understandability, Naturalness and Grammaticality 132
6.8 The RWTH Aachen University Avatar 133
6.9 Manual Evaluation Interface Used in RWTH Experiments 133
6.10 The Poser 6 Animation Interface 137
6.11 Example of Interpolation Between Frames in Poser 137
6.12 Robert: The Poser 6 Avatar 139
6.13 ISL Evaluation Central Navigational Page 146
6.14 ISL Evaluation Sample Sentence Evaluation Page 147
6.15 Bar-Graph Showing Manual Evaluation Intelligibility Scores 150
6.16 Pie-Chart Showing Percentage of Scores for Intelligibility 151
6.17 Bar-Graph Showing Manual Evaluation Fidelity Scores 152
6.18 Pie-Chart Showing Percentage of Scores for Fidelity 153
6.19 Comparison of Intelligibility and Fidelity Scores 154

List of Tables

3.1 Comparison of Related Research 58

4.1 Manual Evaluation Scores for NGT Data 79

5.1 Comparison of Automatic Evaluation Scores for ATIS Data and ECHO Data 95

5.2 Evaluation Scores for MaTrEx Spoken Language Experiments . . . 100

5.3 Overview of ATIS Corpus Divisions: training, testing and development 102

5.4 Comparison of Evaluation Scores: MaTrEx baseline with ISL data, prototype with ISL data and prototype with NGT data 103

5.5 Comparison of Evaluation Scores: MaTrEx baseline alone compared with the addition of Type 1 chunks 105

5.6 Comparison of Evaluation Scores: MaTrEx baseline alone compared with the addition of Type 1 chunks and addition of Type 2 chunks . . 106

5.7 MaTrEx Evaluation Results for ISL–EN Gloss-To-Text Experiments 108

5.8 Evaluation Scores for MaTrEx Spoken Language Experiments compared with SL experiments 108

5.9 MaTrEx Comparative Evaluation Results for All Gloss-To-Text Experiments 110

5.10 Comparison of RWTH Evaluation Results with best of MaTrEx . . 112

5.11 Evaluation Scores for EN–ISL experiments: comparison with best ISL–EN scores shown 117

6.1 Comparison of MT Animation Approaches 135

6.2 Comparison MaTrEx Animation with Other MT Animation Approaches 141

Abstract

This thesis explores the application of data-driven machine translation (MT) to sign languages (SLs). The provision of an SL MT system can facilitate communication between Deaf and hearing people by translating information into the native and preferred language of the individual.

We begin with an introduction to SLs, focussing on Irish Sign Language - the native language of the Deaf in Ireland. We describe their linguistics and mechanics including similarities and differences with spoken languages. Given the lack of a formalised written form of these languages, an outline of annotation formats is discussed as well as the issue of data collection. We summarise previous approaches to SL MT, highlighting the pros and cons of each approach. Initial experiments in the novel area of example-based MT for SLs are discussed and an overview of the problems that arise when automatically translating these manual-visual languages is given.

Following this we detail our data-driven approach, examining the MT system used and modifications made for the treatment of SLs and their annotation. Through sets of automatically evaluated experiments in both language directions, we consider the merits of data-driven MT for SLs and outline the mainstream evaluation metrics used. To complete the translation into SLs, we discuss the addition and manual evaluation of a signing avatar for real SL output.

Acknowledgments

I would first like to thank Prof. Andy Way for his unrivalled supervision and support during my PhD studies. He has been an exemplary supervisor, always approachable and offered guidance and autonomy in equal measures. Thank you to IRCSET and IBM for their generous scholarship that has allowed me to undertake the research described in this thesis.

I am very grateful to the members of the Irish Deaf Community who have guided and assisted my work, particularly Kevin G. Mulqueen, Mary Duggan, Claire Power, Bernie Walsh and Phil Massey whose time and Irish Sign Language expertise greatly contributed to this thesis.

Thanks to Yanjun, John, Ventsi, Nicola, Ríona, Yvette, Karolina, Bart, Declan, Sywia and Mary, as past and present members of my research group for their support and interest in my work. Special thanks to Mary for her advice on thesis management, inspiring my career path and for her friendly, well-reasoned advice. Thanks also to my wider research group, Mark, Caroline, Gavin, Cara, Joachim, Gráinne, and James for providing welcome lunchtime distractions and some of the oddest conversation over the last few months! Thanks to Daniel Stein from RWTH Aachen University for being a kindred spirit and for not taking ourselves too seriously. Tell me again, Daniel, why there isn't just one sign language for everyone?

I have had a consortium of supporters outside of DCU to whom I am most grateful for reminding me of the outside world. Thank you to my parents, Aidan and Clare, and my sister Janis, your unconditional support and love on all my educational adventures has given me ambition and encouraged me to believe in myself. My friends Carol, Sorcha and Ríona for their understanding and willingness to listen, be it over a glass of wine or a cup of tea. Finally to Alex, for making me laugh, having faith in me, distracting me when I need distraction and pushing me to work when I need to work. Your patience, support and understanding have been greatly appreciated.

Namaste to you all.

ॐ

Chapter 1

Introduction

> *"Communication leads to community, that is, to understanding, intimacy and mutual valuing."*
>
> Rollo May
>
> American Existential Psychologist

Communication is the essence of human interaction. One of the most efficient means of communication for humans is through the use of languages, which themselves are born from human interaction. However, this natural means of communication can prove a barrier in cases where languages differ. This is nowhere more evident than in the communication barrier between languages of different modalities, namely spoken and sign languages (SLs).[1] Indeed, the contrary of the above quote is also true; that a lack of communication can lead to a breakdown in communities, and a *lack* of understanding, intimacy and mutual valuing. Commonly, users of SLs experience ambivalence at best toward their language from non-SL users, and in almost all cases it is less valued than the majority spoken language (Leeson, 2003).

While simple spoken language communication problems can often be partly resolved by a combination of basic foreign language knowledge, similar sounding words,

[1] The term 'spoken languages', in the context of this thesis, does not refer specifically to oral-only versions of these languages, but rather is used to distinguish them from visual–gestural sign languages.

and gesticulations, the cross-modal nature of SL–spoken language communication poses additional challenges. Not only are the words and structure different, but also the mode of communication. Instead of oral/aural–oral/aural, we are faced with oral/aural–visual/gestural.

To assume that this barrier can be overcome by lip-reading and speech on the part of the Deaf[2] person, and speaking and watching a signer 'mime' on the part of a hearing person, is to be ignorant of the rich, complex and fully expressive language that is an SL. While lip-reading is a common competency in Deaf communities, it relies on the speaker talking slowly with clear articulation of the lips at the very least. Furthermore, the reader must have a good grasp of the spoken language being used, and even then interpretation errors are easily made. Take, for example, the mouth patterns for the words 'mutter' and 'butter' in English. While the sounds of the first letters are aurally distinct, visually, they appear the same. For the case of speech articulated by a Deaf person, depending on the profoundness of their deafness, they may find it difficult to emulate and reproduce oral phonetic sounds. Taking a hearing person attempting to understand a person who is signing, we cannot assume that the information will be understood from iconic-like mime gestures. SLs are far more complex than this (cf. Chapter 2), employing their own grammatical rules and making full use of the signing space to articulate the concrete or abstract objects of the discourse. Furthermore, person-to-person communication within the confines of language barriers usually requires one or the other party to break from using their native language, something which may not be possible for either party in the context of Deaf–hearing communication.

While on the one hand, direct cross-modal person-to-person communication can

[2] It is generally accepted (Leeson, 2003) that 'Deaf' (with a capital D) is used to refer to people who are linguistically and culturally deaf, meaning they are active in the deaf community, have a strong sense of a Deaf identity and for whom SL is their preferred language. 'deaf' (with a small d) describes people who have less strong feelings of identity and ownership within the community, who may or may not prefer the local SL as their L1. Hard-of-hearing (HOH) is generally used to describe people who have lost their sense of hearing later in life and have little to no contact with the deaf community or SL usage for various social and cultural reasons. The boundaries of these categories are fuzzy and people may consider themselves on the border of one or another depending on their experiences and preferences.

cause problems, the Deaf community can also have issues with indirect communication, namely written language. Studies have shown that the average Deaf adult has the literacy competencies of 10-year-old (Traxler, 2000).[3] In addition, a study of the educational background and employment status of Deaf adults in Ireland shows that 38% of the study participants[4] did not feel fully confident to read a newspaper, and more than half were not fully confident writing a letter or filling out a form (Conroy, 2006):11. Regardless of the literacy competencies of a Deaf person, it should be their right to have information available to them in SL, their first and preferred language, and a language that is most natural to them. In Ireland alone, there is a total population of approximately 50,000 people who know and use ISL (Leeson, 2001), yet "there are no public services available in this language and provision of services in an accessible language — in all domains of life — is relatively *ad hoc*" (Leeson, 2003):150.

One way of breaking through the communication barrier is through SL interpreters (SLIs). An interpreter is a trained professional who is employed to facilitate communication between Deaf SL users and hearing people with no SL skills (Conroy, 2006). While SLIs play an important role here, there are some drawbacks. The first is the number of interpreters available. In Ireland, formal SLI training was only established in the early 90s, and according to Leeson (2003), the ratio of SLI:Deaf person is as large as 1:250. This means that frequently interpreter availability is low. The second drawback to using SLIs concerns confidentiality. The presence of an interpreter for legal or medical communication, for example, can be perceived as an intrusion in to the privacy of the Deaf person. Given the small number of SLIs, there is an increased likelihood of the Deaf person knowing or being related to the SLI. Furthermore, hiring an SLI can be expensive if he/she is only required for a small amount of interpreting, i.e. texts or short conversations that occur in

[3] This statistic is taken from the American Deaf community, but, conversations with the Director for the Centre for Deaf Studies, Dublin, indicate that this is also the case in Ireland.

[4] 354 Deaf people across Ireland participated which involved completing questionnaires or attending focus groups.

everyday life. Clearly, some means of interpretation or translation is required for these circumstances that protects the privacy of the individual and facilitates the exchange of information in context where the use of an SLI is impractical.

Advancements in technology have provided some answers to this problem. Teletype systems are available for telephone conversations, and closed-captioning and subtitles are available on all DVDs nowadays. While these tools do alleviate a Deaf person's problems with hearing, they assume good literacy skills and speed of reading and understanding (Huenerfauth, 2006). Often, for reasons of space and timing, subtitles are simplified meaning that some of the important information can be missing. In addition, this simplification can be seen as insulting to the Deaf communities and may be considered a 'dumbing down' of the information.

There is, however, one particular type of technology that can tackle the communication and comprehension problems of the Deaf community, namely machine translation (MT). Over the last 60 years or so, since the first proposal of using computational methods to automate translation (Weaver, 1949), advancements in MT research have shown to significantly bridge communication and understanding between different spoken languages.

The proliferation of this technology, coupled with the need for practical, objective tools for translation in the Deaf community, has motivated the research described in this thesis. We seek to explore and expand modern data-driven MT research to the field of SLs in order to produce assistive technology to allow Deaf people to receive and communicate information in their preferred language.

In recent years data-driven MT methodologies have taken centre stage as the most efficient and productive method of automatic translation (cf. Chapter 3). Given that a prerequisite of this approach is a bilingual corpus and that there is no formally recognised writing system for SLs, this leads us to our first research question:

(RQ1) *Is it possible to employ data-driven MT methodologies to SLs with no consistent written form?*

The sole concern of this thesis is not to investigate the applicability of MT methods to visual–gestural languages, but to also provide usable assistive technology to help alleviate the communication problems highlighted above. This leads us to our second research question:

(RQ2) *Can data-driven MT make information more accessible to the Deaf community by providing it in their first language?*

The remaining chapters of this thesis will seek to address these questions through the inclusion of background information, overviews of past approaches and a series of experiments.

Chapter 2 provides an introduction to languages on which this thesis focuses, namely SLs. We include descriptions of linguistic phenomena and transcription methods that both have an impact on MT.

Chapter 3 gives an overview of other SL MT approaches including rule-based and data-driven methodologies and includes an outline of SL MT-specific problems.

Chapter 4 discusses our initial sets of experiments using Dutch Sign Language data, highlighting the main problems for data-driven SL MT, namely data scarcity and format issues, as well as evaluation.

Chapter 5 describes our main experiments in SL MT. We first describe our purpose-built ISL data set, then the MaTrEx data-driven MT system and finally five sets of experiments detailing translations to and from SLs.

Chapter 6 completes our experimental work, detailing the creation and evaluation processes involved in the generation of an animated avatar for signing our translated output.

Chapter 7 concludes this thesis and outlines some future avenues of research.

The research presented in this dissertation was published in several peer-reviewed conference proceedings. (Morrissey and Way, 2005) presents our initial findings for English–SL translation using a prototype EBMT system. (Morrissey and Way, 2006) addresses the issue of evaluation for SL MT and includes our findings for the reverse translation direction. Our main experiments performed using the MaTrEx data-driven MT system and a purpose-built ISL data-set are presented in (Morrissey et al., 2007b) and (Stein et al., 2007) as well as (Morrissey et al., 2007a). The MaTrEx system itself is presented in joint work in (Armstrong et al., 2006). Finally, an outline of the human factors required for SL MT data collection and evaluation are presented in (Morrissey and Way, 2007).

Chapter 2

Sign Languages

Sign Languages (SLs) are the first languages of members of the Deaf communities worldwide (Ó'Baoill and Matthews, 2000). Naturally occurring and indigenous languages, they can be as eloquent and as powerful as any spoken language through their visual–spatial modality. It is this alternative communicative channel that poses interesting challenges for the area of MT that lies outside of the general spoken language MT field.

This chapter introduces SLs as fully expressive independent languages. Divided into two sections, we first give a general introduction to these visual-gestural languages, outlining their linguistic structure and highlighting linguistic phenomena prevalent in SLs that can have an impact on MT. Particular attention will be paid to Irish Sign Language (ISL), the SL native to Ireland and the primary SL addressed in this thesis. The second section addresses the lack of writing systems available for SLs and outlines notation systems that may be employed to circumvent this issue.

2.1 Background

SL research is still a relatively new area when compared to research into spoken languages. Significant study into the linguistic structure of SLs only began in the 1960s with the seminal work of Stokoe (1960), whose research on American Sign

Language paved the way for social recognition of SLs as real languages. More recent acknowledgment of this is included in the works of linguist such as Pinker (1994) and Chomsky (2000). Increased recognition of SLs as fully-formed, independent languages following political acknowledgment, such as the Resolution of the European Parliament in 1988,[1] has led to some level of research being carried out on SLs in most countries. Usually it is the national centre for Deaf Studies or other Deaf associations that investigate the sociological, educational, cultural and linguistic aspects of SLs. In Ireland, for example, the researchers at the Centre for Deaf Studies[2] engage in valuable ISL linguistic research (Ó'Baoill and Matthews, 2000; Leeson et al., 2006), and members of the Irish Deaf Society[3] have produced a book outlining the employment and poverty problems experienced by members of the Dublin Deaf Community (Conroy, 2006).

Despite common misconceptions, sign languages are not universal (Leeson, 2003).[4] If they were, there would be no barrier between Deaf communities. The Ethnologue of world languages[5] catalogues 121 sign languages for the Deaf worldwide. The majority of these languages are distinct languages in their own right that have evolved either naturally within Deaf communities themselves or have originally been borrowed from other SLs. Even within SLs in different countries there is some variation. German Sign Language/Deutsche Gebärdensprache (DGS), one of the SLs used in our experiments in Chapter 5, for example, has many dialects used in different areas across the country, and Ireland has the unusual division of male and female sign language (see p.14). As SLs differ from country to country, they are not mutually intelligible as is commonly thought (Ó'Baoill and Matthews, 2000). While there is a certain amount of iconicity displayed in SLs such that the meaning of some signs could be guessed by a non-native, generally even iconic signs differ in different languages. For example, *cat* in ISL is signed by placing the thumbs of each hand to the

[1] http://www.policy.hu/flora/ressign2.htm

[2] http://www.tcd.ie/slscs/cds/

[3] http://www.deaf.ie

[4] See 'Common Myths about Deafness' at http://www.deaf.ie/IDSinfo.htm

[5] http://www.ethnologue.com

cheeks, palms facing out and wiggling the fingers. In DGS, the same word is signed by linking the index finger and thumb, placing them either side of the nostrils, just above the upper lip and moving the hands away from the face two or three times. Both of these signs may be considered to be iconic in that they indicate that a cat is being signed from the whisker-like movements, but both are distinct signs in each language and are perhaps as mutually intelligible as say *cat* and *Katze* in English and German respectively.

Although there is no universal SL *per se*, an International SL often termed *Gestuno* does exist.[6][7] Rather than being a fully-formed language, it is a vocabulary of signs, more iconic in nature than SLs in general. It is primarily used to facilitate communication in international contexts when there is no common SL. It is not standardised nor does it have its own grammar. Given this flexibility and lack of native users, International SL can only go so far to bridge the communication gap between different SL communities. Rather than standardising an international form of the language which would require tens of thousands of people to learn a new language, never mind the language development needed, this is another area where SL MT could significantly facilitate and improve international communication on an SL level.

Another common misconception is that SLs are grammatically dependent on spoken languages.[8] This is not the case. If it were, then the role of MT would be minimal and only a direct dictionary look-up system for finding the gesture sequence corresponding to each spoken language word would be needed. Having said that, SLs do exist where the grammar and often the morphology of the spoken language is followed. An example of this is Signed Exact English (SEE). These versions of sign languages adhere to spoken language grammar (in this case, English), where there is a sign for each word of the spoken language sentence and often for each morpheme too. This format is sometimes used in the media as a means to interpret

[6] http://listserv.linguistlist.org/cgi-bin/wa?A2=ind0202&L=slling-l&P=82

[7] http://www.deaflibrary.org/asl.html

[8] See ‘Common Myths about Deafness’ at http://www.deaf.ie/IDSinfo.htm

spoken news to a signed version when there is little time for a proper translation and change of grammar, for example the News for the Deaf shown on the national television station in Ireland. SEE is often used in schools for the Deaf where there is an oralist[9] teaching tradition or by people who have become deaf later in life and learned manual communication as a second language. While it does speed up the translation process, as there is no need to alter the grammar structure, the manual signing process itself is laborious and slows down the communication of information, because full and best use of the signing space and simultaneous spatial linguistic phenomena of SLs are ignored. For most concerned, SLs are independent of spoken languages, make the best use of the modality of communication and thus bring to light interesting issues for MT to tackle in cross-modal translation.

2.2 Sign Language Linguistics

Now that we have shown that there are multiple SLs and that they are distinct from spoken languages, we will discuss the linguistics of SLs in general and point out the relevance of these issues for MT.

2.2.1 Articulation and Structure

Compared to spoken languages where the primary articulators are the throat, nose and mouth, the main articulators in SLs are the fingers, hands and arms. The signs themselves are analogous to morphemes in spoken languages and the articulations of the hands and body can be categorised as phonemes like those in spoken languages, (Stokoe, 1972). However, unlike speech, these phonemes are not linear and sequential, but rather occur simultaneously. There are five categories of phoneme in an articulated sign: the shape, orientation, location and movement of the hand as well as the non-manual features. The simultaneous production of these phonemes

[9] Oralism is an educational method for deaf children that eschews the use of sign language in favour of spoken language through the use of text, residual hearing, lip-reading and technological aids. It was introduced into Irish Deaf schools in the 1930s and remains the predominant teaching methodology (Conroy, 2006).

allows a signer to communicate different ideas or aspects of a sentence while at the same time exploiting the signing space of their articulation to the full. In this way, complete ideas can be represented in the signing space at any one time. Typical linguistic phenomena used by signers to communicate thoughts and ideas in a way that best uses the visual medium are non-manual features, classifiers, and spatial deictics, for example. These are discussed in more detail in the following sections.

2.2.2 Non-Manual Features

Integral to the transmission of information through a manual modality is the inclusion of *non-manual features* (NMFs). Predominantly concomitant with manual signs, they consist of movements or expressions of parts of the body other than the hands that can express emotion, intensity or act as morphological and syntactic markers. These consist of "eyebrow movement, movement of the eyes/cheeks, mouth patterns, tilting of the head, movement of the upper body and shoulder movements" (Ó'Baoill and Matthews, 2000):45. The inclusion of an NMF with a manual sign in discourse can alter the meaning of a sign and its absence can render a sign meaningless. For example, the sign for *cat*, as described previously, can be augmented into a question by simply adding the NMF of raising the eyebrows. Given that a person watching another person signing only looks at the hands directly 30% of the time with the remaining 70% spent looking in the person's eyes or the rest of the body[10], the importance of including NMFs in the development of an SL MT system is important.

2.2.3 Signing Space

Sign languages are gestural languages that are articulated in such a way to make the best use of the space in which articulation can take place, in a similar way that oral speech makes best use of the vocal tract, nose and mouth. This articulation area in SLs is called the *signing space*. This space extends from just above the head

[10] In conversation with ISL tutors and our ISL animation evaluators discussed in Chapter 6.

down to the waist and outwards about as far as the arms can extend. A diagram of this space taken from Ó'Baoill and Matthews (2000):40 is shown in Figure 2.1.

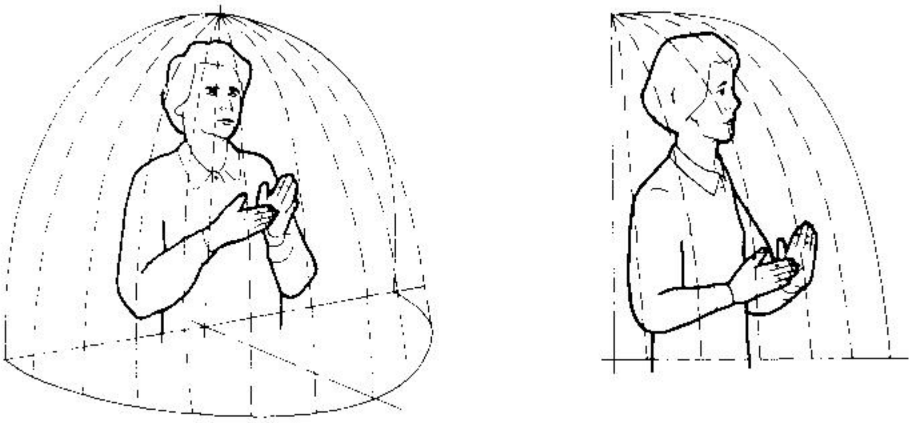
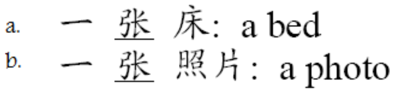


Figure 2.1: Diagram of Signing Space

All manual and non-manual articulations take place within this space. SLs have developed to make the best use of the visual communication channel. The signing space in front of a signer can be subdivided during communication into neutral signing space as well as deictic reference points, and showing directional verbs and classifier predicates (discussed in Section 2.2.4). The neutral signing space is the space directly in front of the signer that normally does not contain reference points, and is where most non-body contact, non-referential signing takes place. Locational points in the signing space are chosen to 'locate' nominal items for subsequent reference. Once the nominal has been signed it is 'placed' with a pointing gesture at a specific point in the signing space that remains exclusive to that nominal for the conversation unless moved or changed. A signer may then use that point in the signing space to refer back to the nominal in an anaphoric reference or may use a directional verb moving from one referenced object to another. For example, if a signer articulates the sign for 'house' and points to a location on the left of the signing space, this indicates the assignation of a new reference point for that house in the signing space. If the signer then articulates a man walking, he can show that the man is walking to or from the house by using the house's 'location' in the signing space as a start- or end-point of the movement. The use of a semantic 3–D space is not something that occurs in spoken languages, and given that locational information is important to the meaning within a sentence, it is, therefore, important that this is taken into account both in the notation of the SLs and in their translation.

2.2.4 Classifiers

A linguistic phenomenon that is prevalent in SLs and makes excellent use of the gestural modality of these languages is that of *classifiers.* Classifiers generally group vocabulary together according to "certain formal or semantic properties" (Ó'Baoill and Matthews, 2000):63. Classifier languages exist in the spoken language domain that require the affixation of certain morphemes to a word if it belongs to a particular semantic class. Classifiers are frequently found in Eastern languages. Figure 2.2 shows the use of classifiers in Chinese.



a. 一 张 床: a bed
b. 一 张 照片 : a photo

Figure 2.2: Example of Classifier Usage in Chinese

In both examples, the first character is just the number *'1'*, indicating the quantity of the subsequent objects discussed. The second, underlined, character is the classifier used to denote a flat object. The subsequent characters indicate exactly what this object is, in this instance a 'bed' and a 'photo'. Some classifiers do exist in English, such as the term *head* in 'fifty head of cattle' where it is a mass noun.

In SLs, classifiers are purportedly used in the majority of SL sentences[11] although their usage differs slightly from those of spoken languages. Classifier signs are used to denote the shape or arrangement or consistency of objects, for instance. Generally the classifier is preceded by a citation form of the lexical item which is followed by the relevant classifier which can demonstrate the action or location of the cited object. To put this in some context, we visit the example from Ó'Baoill and Matthews

[11] From conversations with Deaf colleagues.

(2000). Here the sign for *car* is signed (two fists, palms facing the signer at chest level that move alternately up and down), followed by a classifier sign for all vehicles (a flat palm of the dominant hand with fingers extended and together and the thumb facing up). This vehicle classifier sign then takes on the meaning *car* given the context and the previously signed citation form. The classifier hand shape can then be moved to indicate the action of the car.

The prevalent use of similar classifiers throughout SLs in this *citation–classifier* format leads us to acknowledge the importance of including these features in the processing of SLs for MT.

2.2.5 Irish Sign Language

In Ireland, Irish Sign Language (ISL) is the dominant and preferred language of the Irish Deaf community. There are approximately 5,000 Deaf people in Ireland (Matthews, 1996) that have ISL as their L1 with an additional 45,000 deaf and hearing people using ISL in addition to their first language (Leeson, 2003). These include family members and friends of Deaf people, and people in educational, research or social roles. This means that approximately 1.2% of the Irish population use ISL, with 0.12% using it as their first language. Despite this, ISL remains a poorly resourced minority language that still lacks social and political recognition. ISL is not yet recognised as an official language despite European Parliament legislation calling for "Member States to abolish any remaining obstacles to the use of sign language".[12]

It is perhaps due to this lack of recognition and related under-resourcing that linguistic research into ISL began as late as the 1980s. Today, this is a developing area with The Centre for Deaf Studies, Dublin, as the primary research unit involved in ISL linguistic research.

ISL is a derivative of French Sign Language/La Langue des Signes Français (LSF). The history of ISL is somewhat difficult to trace but it is thought that there

[12] Resolution for Sign Languages for the Deaf, Irish Deaf Society 1988.

was some usage of British Sign Language (BSL) in Dublin in the early 1800s before Dominican nuns set up a school for deaf girls using LSF in the mid-1800s. This developed into a language in its own right and was later adopted by the Christian Brothers who had set up a school for deaf boys. Interestingly, this gender-based educational segregation led to the development of different varieties of ISL and localised differences in signing vocabulary. 'Female' ISL in female-to-female conversations is generally within a reduced signing space, with shorter movements and different lexical choices to those used in female-to-male conversations (Ó'Baoill and Matthews, 2000). While the male form is generally considered the standard variation, it is noted in Le Master (1990) that the female version has a longer historical provenance than the male.

Another feature of ISL is the use of initialised signs. The alphabet of ISL is a one-handed alphabet similar to LSF and also American Sign Language (ASL).[13] Initialised signs are thought to be influenced by the introduction of oralism into deaf education. The hand shape used in these signs is the first, or initial, letter of the word in the spoken language. For example, the hand shape for the sign *nice* is an '*n*' and the hand shape for *mother* is an '*m*'. The use of such signs is prevalent in ISL. These signs may be differentiated from *finger signed* signs that are abbreviations of concepts that are usually finger spelled, such as place names. Finger signs also use alphabetical signs taken from the spoken language spelling, but generally merge multiple alphabetical hand shapes together sequentially in a short space of time rather than maintaining the one hand shape of the initial letter. An example, taken from the corpus used in our experiments in Chapter 5, is the sign for the city *Cork*, where the '*c*' and '*k*' may be clear but the letters in between may be blurred somewhat by the speed of movement.

For reasons of education and cultural involvement, the language level of members of the Deaf community in Ireland varies. The majority are bilingual in ISL and English and have varying competencies in each language (Ó'Baoill and Matthews,

[13] ASL is also a derivative of LSF, coming from the same family of languages.

2000; Conroy, 2006). Members of the Deaf community for the main part use ISL as their first language, but other versions of the language exist such as SEE as described in Section 2.1. To communicate with non–ISL users, Deaf people usually require the assistance of an interpreter. This can cause problems in a number of respects. As noted in Leeson (2003) there is a distinct shortage of interpreters in Ireland with an interpreter:Deaf person ratio of 1:250. This low number means that it can be difficult to avail of an interpreter when needed, as frequently resources are stretched. The low number of interpreters also means that the chances of knowing your interpreter personally are increased which can impinge on privacy, particularly in the case of legal or medical contexts. Furthermore, if interpretation is required for only a small amount of information, hiring an interpreter could be a prohibitively expensive process. While interpreters will always be necessary and valued in the Deaf community, there are many instances when interpreting or translation is required but may not be available or practical. In these instances, the provision of an automatic SL MT system that could translate from spoken language into SLs and vice versa could greatly facilitate inter-community communication and would provide an objective and more private means of communicating.

2.3 Transcription and Notation Systems

So far we have addressed SLs from a person-to-person communication point of view. This could be compared to a discussion of speech for spoken languages. But we are primarily concerned about SLs for MT and for this a text or symbolic representation is required (cf. representations used for the MT systems discussed in Chapter 3 and our data-driven experiments in Chapters 4 and 5). One of the striking differences between sign and spoken languages is the distinct lack of a formally adopted, or even recognised, writing system for SLs. There are many possible reasons for this, one being that as SLs are minority languages, they are poorly resourced as much investment in languages serves to increase the prominence and power of dominant languages and ignore the less significant, minority ones (Ó'Baoill and Matthews, 2000). On a more linguistic level, simultaneous articulation of the phonemic structure of SLs, as described previously in this section, does not lend itself to a linear writing system.

The lack of a formalised or widely used writing system for SLs means that SLs remain as visual–spatial languages that cannot be 'read' as spoken languages can. There have been many attempts at creating writing systems for SLs, but most are not usable by the general public as they consist of numeric codes or symbols to encapsulate the phonetics or phonology of signs and are not easily learned, written nor is there a standardised accepted form. The models discussed in the following sections outline the most commonly used systems and those that have been adopted by the SL MT systems described in Chapter 3.

2.3.1 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 hand shape; 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 hand shape, have resulted in no universally accepted version of the Stokoe notation system (Ó'Baoill and Matthews, 2000). An example of the Stokoe notation for the sign *don't know* in ASL can be seen in Figure 2.3.

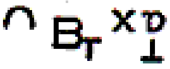


Figure 2.3: Example of Stokoe Notation: *don't know*

This example here clearly shows the use of tabulation, signation, designator and orientation symbols for a single sign. The first symbol, like an upside-down *u*, is a tabulation symbol referring to the forehead, brow or upper face. The letter *B* that follows it is a designator symbol indicating a *B* hand shape where all fingers are extended and side-by-side but with the thumb folded in. The subscripted capital *T* symbol beside it refers to the orientation of the designator, in this case facing the signer. The superscripted *x* symbol indicates a signation where the designator hand shape *B* moves to touch the tabulation location of the upper head. The next superscript symbol, similar to an inverted, reversed lower case *a*, is also a signation symbol indicating the palm is turned down. The final symbol, a subscripted, upside-down *T*, indicates that the hand now faces away from the signer.

While this approach describes a comprehensive analysis of signs articulated using the hands, the method is data-heavy and not practical for use as a writing system for Deaf people and lacks any means for describing NMFs. Furthermore, Stokoe-based systems are primarily useful for notating single, decontextualised signs, for example citation forms in SL dictionaries, and not for signs used within a context. In addition, there are no large corpora available in this format for use in a data-driven MT system.

2.3.2 HamNoSys

Another explicit notation system for SLs is the Hamburg Notation System (HamNoSys) (Hanke, 2004) that uses a set of language-independent symbols to iconically represent the phonological features of SLs (Ó'Baoill and Matthews, 2000). This system, rooted in the Stokoe system described in the previous section, allows even more detail to be described but in most cases it is only a description of the hand shape. In Figure 2.4 the symbols that denote the sign *nineteen* are shown.

This example illustrates most of the description categories used in HamNoSys notation. The first colon-like symbol indicates that the hand arrangement is a two-handed sign where the hands are parallel but not mirroring each other. The square

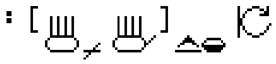


Figure 2.4: Example of HamNoSys: *nineteen*

brackets denote a grouping of simultaneously occurring movements. Within these braces, the oval symbol with four vertical lines indicates one hand has four fingers extended and aligned together, the same symbol repeated with a fifth diagonal line denotes the same hand shape but with the thumb extended this time. The horizontal line with the dash through it in between these two hand shapes shows the movement is repeated from the starting point of the original movement. Outside of the brackets, the next symbol, a horizontal line with an upward facing arrow above it, indicates that the hand orientation is away from the body, while the next symbol, an oval with the lower half shaded, signifies that the palm of the hand is facing down. The final symbol, a vertical line with a clockwise circular arrow to its right, shows that the movement of the hands is circular, moving vertically and away from the body.

As this system can be laborious to transcribe, it is not suitable for adoption by the Deaf community as a writing system and again, no large SL corpora are available in this format. Furthermore, the symbols of both this and the Stokoe Notation are not easily machine readable for MT, requiring specific technology (such as that created for avatar development in the ViSiCAST system described in Chapter 6).

2.3.3 SignWriting

An alternative method was developed in (Sutton, 1995) called SignWriting. 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, such as that shown in Figure 2.5, make the system easier to learn as they are more intuitively and visually connected to the signs themselves.

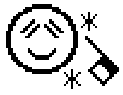


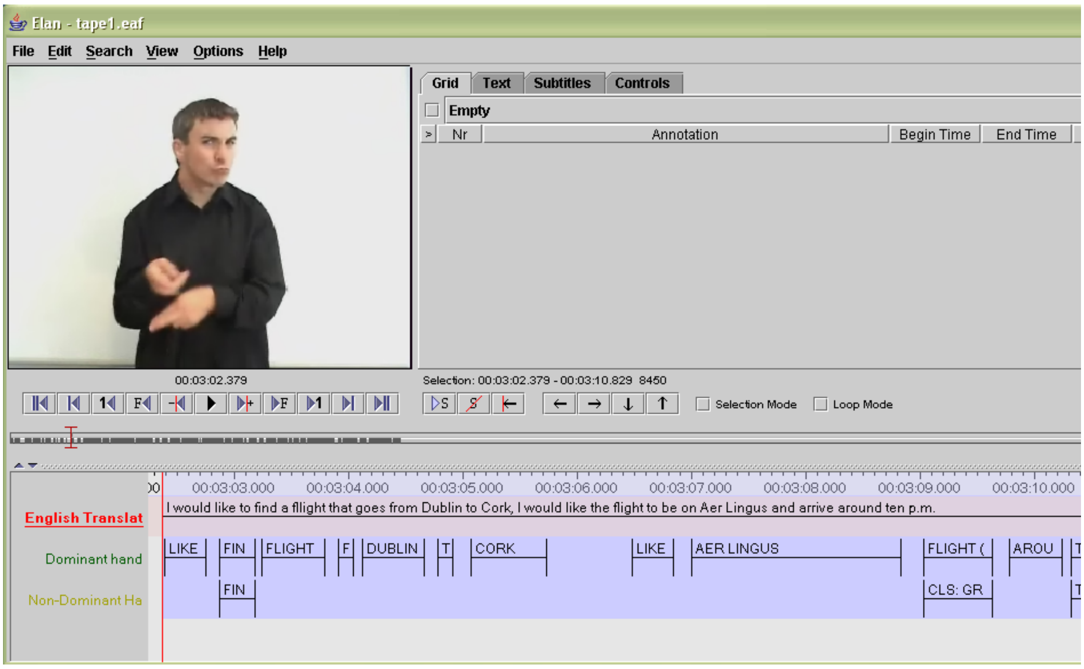
Figure 2.5: Example of SignWriting: *deaf*

This example of SignWriting shows a circle that denotes the head of the signer with two sets of two parallel opposing diagonally slanting lines indicating eyes and a three further lines joined to form a smile. The square to the right of the head with a line extending from it indicates a hand with the index finger pointing out, and the semi-shaded square denotes that the back of the hand is facing outward. The asterisk symbols above and below the hand symbol and against the top right and lower right of the face indicate that the index finger touches these points.

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 has resulted in the publishing of SignWriting books for learning the system such as *'Goldilocks & The Three Bears'* (Clark Gunsauls, 1998). The linguistic detail of SLs are encoded in the line drawings of SignWriting. Essentially, a secondary 'language' such as ASCII code representation of the SignWriting pictures or some form of recogniser would be needed to transfer the necessary detail from SignWriting text to a representation suitable for use in SL MT.

2.3.4 Manual Glossing and Annotation

As previously mentioned, the 3–D nature of SLs does not lend itself to a linear format but rather a multi-linear and time-aligned format with multiple descriptive tiers (Leeson et al., 2006). One way around such problems with writing systems is to manually annotate SL video data. This moves away from the task of solving the writing problem, but rather addresses the issue of SL representation. This approach



English Translat: I would like to find a fllight that goes from Dublin to Cork, I would like the flight to be on Aer Lingus and arrive around ten p.m.

Dominant hand: LIKE FIN FLIGHT F DUBLIN T CORK LIKE AER LINGUS FLIGHT (AROU T

Non-Dominant Ha: FIN CLS: GR T

Figure 2.6: An Example of the ELAN Interface Tool for Manual Glossing and Annotation of SL Video Data

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, whether there is repetition of the sign and its location or any other relevant linguistic information (cf. see Section 2.2). The annotations are time-aligned according to their articulation in the corresponding SL video. An example of an annotation interface (ELAN) is shown in Figure 2.6. Both this process and the interface is described in more detail in Chapter 4.

The video of signed data is displayed on the upper left hand side of the screen. Below, in horizontal lines are the annotations, where each line represents a different annotation feature called an annotation tier, and all are on a time-line relative to the signing in the video. Users can add any number of annotations and annotation tiers to suit their needs.

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 do not have to be loaded with linguistic detail and can provide gloss terms for signs that facilitate translation from and into spoken language. Furthermore, it is a useful feature as it means that the phonological structure of SLs can be transcribed simultaneously and in context. Such a provision can assist in real-time avatar production, as discussed in Chapter 6. On the other hand, however, the problem of inter-annotator agreement remains. For data-driven MT, the approach taken for this thesis, this could mean discrepancies in the training data that could hinder the capacity of the corpus-based MT system to make the correct inferences. For our ISL data in Chapter 5, we show that one way around this issue is to confine annotation to just one transcriber.

The provision of an English translation for each signed sentence, along with other time-aligned annotation tiers allows for easy alignment of corpora on a sentential level, as information within the time limits of the English translation can then be aligned with that annotation. For example, all NMFs or phonetic features signed with the word 'hare' can be grouped with that word. The presence of time spans for each annotation also aids in the aligning of information from each annotation tier to form chunks that can then be aligned with chunks derived from the English tier. We demonstrate this in our experiments described in Chapters 4 and 5.

Time-aligned annotations are also useful for tackling the issue of coarticulation of signs. The phenomenon of co-articulation in sign languages is analogous to co-articulation in spoken languages where the articulation of a phoneme may be altered relative to its neighbouring phonemes (Jerde et al., 2003). Co-articulation can occur in fluent signing when the shape of the hand for one sign is altered relative to the hand shape for the subsequent sign. Annotation can overcome this as, even if signs are co-articulated in the videos, the annotations for the signs will be separate. Either way they are easy to separate using the time span figures. As it is these annotations that can be used in the translation output, the issue of separating co-articulated words is removed automatically.

These annotated corpora also provide a solution to the SL translation issue of NMFs. In sign language, meaning is conveyed using several parts of the body in parallel (Huenerfauth, 2005), not solely the hands (cf. Section 2.2.2). NMFs are used to express emotion or intensity, but also can be used morphologically and syntactically as markers (Ó'Baoill and Matthews, 2000). The annotations of the ECHO corpora, described in Chapter 3, contain explicit NMF detail in varying tiers such as eye aperture and mouth that combine with other tiers to form complete signs and, therefore, more linguistically complete translations.

The example in (1) shows the effect NMFs have on a sign. The `Gloss RH/LH English` is the manual hand sign articulated by the right and left hand, in this case showing that of a hare running. The annotation `n` on the head tier indicates a nod of the head. This combined with the furrowing marked in the brows tier (signified by `f`), the squinting marked in the eye aperture tier (signified by `s`) and the puffing of the cheeks marked in the cheeks tier (signified by `p`) shows the intensity of the running that the hare is doing. Without these NMFs the hare would be understood to be running at a normal pace.

(1)
```
(Gloss RH English) running hare
(Gloss LH English) running hare
(Head) n
(Brows) f
(Eye Aperture) s
(Cheeks) p
```

In many cases NMFs are essential for providing the full sense of the sign. The more detail that is contained in the annotation tiers, the better the translation and the more suitable the translations will be for use with a signing avatar.

Given that annotated video allows for a transcription of SL video data at a level of granularity to suit any user, as well as having the means to facilitate production of an SL using an avatar, we consider annotation a suitable transcription method for data-driven MT, as we show from our experiments in Chapter 5.

2.4 Summary

In this chapter we showed that SLs are still only in the early stages of linguistic analysis in comparison with spoken languages. Furthermore they are not universal nor linguistically dependent on spoken languages. Despite the existence of International SL and SL variants such as SEE, there is still a communication barrier between different SLs, as well as between SLs and spoken languages. Through an outline of some of the more common linguistic feature of SLs, we demonstrated the spatial and visual nature of SLs and predicted that these will be interesting challenges for the field of MT. We have shown the need for SL MT technology in Ireland by outlining ISL as a poorly resourced language with communication problems brought on by the lack of interpreter availability, a problem that could be partially alleviated by SL MT.

In the second section, we discussed the requirement of MT for a notation system to represent SLs resulting from the lack of a formalised writing system. By outlining systems used in the past, we demonstrated that the process of manual annotation with glossing, given its flexible architecture, adaptability and provision to include important SL linguistic phenomena, is most suited to SL MT.

Chapter 3

Overview of Sign Language Machine Translation

MT for spoken languages is a thriving area of research and development. A clear indication of this is the proliferation of MT products for sale, such as Systran[1] and Language Weaver,[2] as well as freely available on-line MT tools such as AltaVista's Babel Fish[3] and Google Translate.[4] The funding of large research projects such as the Global Autonomous Language Exploitation (GALE) project,[5] the TC-STAR project[6] and the most recent Centre for Science, Engineering and Technology (CSET) project on Next Generation Localisation[7] further demonstrate the importance given to such areas of research in the EU, the US and in Ireland. The same level of activity cannot be said for SL MT, with little more than a dozen systems having tackled this area of translation. Most papers describe prototype systems that often focus primarily on SL generation rather than applying MT techniques to these visual-gestural languages.

This chapter introduces past and current systems that have employed rule-based

[1] http://www.systran.co.uk

[2] http://www.languageweaver.com/home.asp

[3] http://babelfish.altavista.com

[4] http://www.google.com/language_tools

[5] http://projects.ldc.upenn.edu/gale/

[6] http://www.tc-star.org

[7] http://www.dcu.ie/engineering_and_computing/research/Overview/NGL.shtml

and data-driven approaches to SL MT. We show that despite the mainstream movement in MT communities toward data-driven methodologies, SL MT research has predominantly been rule-based. In addition, we demonstrate how, over time, systems within this paradigm have progressed, as well as the more recent move toward more empirical methodologies. Summarising these attempts, we show how they have fallen short of addressing the issue of SL MT.

3.1 Background

Early interest into MT of spoken languages can be traced back to the late 1940s, with a significant expansion of interest in the late 70s and 80s (Trujillo, 1999). A similar level of development cannot be said for SL MT. Widespread documented research in this area did not emerge until the early 1990s. This is a somewhat understandable delay given the comparatively late linguistic analysis of SLs (cf. Chapter 2). Despite this late venture into SL MT, and even within the short time-frame of research, the development of systems has roughly followed that of spoken language MT from 'second generation' rule-based approaches toward data-driven approaches.

This chapter outlines past and current attempts at creating SL MT systems and discusses the pros and cons of each individual system. Section 3.2 outlines rule-based approaches and Section 3.3 then explores the recent movement toward data-driven techniques. For the purposes of brevity and comparison with the work in this thesis, we have chosen to restrict the analysis of systems in this chapter to those that have shown either a reasonable demonstration of the development of MT paradigms for SLs, those that have addressed paradigms not previously nor subsequently researched in-depth, or those that have made some significant impact on this relatively novel area. Where appropriate and depending on the approach, each system will be analysed with respect to:

- the methodology,
- languages used,

- translation direction,
- domain or context of translation,
- format and amount of SL data,
- system description,
- experiments run,
- evaluations performed and their results,
- details of avatar production.

3.2 Rule-based Approaches

The ‘second generation’, or rule-based, approaches to MT emerged in the 1970s/1980s with the development of systems such as Météo (Chandioux, 1976, 1989) and Systran[8]. These systems are examples of the first commercially adopted MT systems to successfully translate spoken languages.

Rule-based approaches may be sub-classified into *transfer*– and *interlingua*–based methodologies. The Vauquois Pyramid, shown in Figure 3.1, is a diagram taken from (Hutchins and Somers, 1992) and widely used in MT circles to demonstrate the relative effort involved in translation processes. The length of the arrows signifies the amount of ‘effort’ required for translation. A direct approach is shown in green, transfer in red and interlingua in blue.

In this chapter we focus on transfer and interlingual approaches. Despite the different paths the transfer and interlingual methods take when translating a source string into a target string, they are inherently similar in their processing as both involve an analysis and generation procedure.

In a transfer approach, *analysis* of the source language input sentence is usually shallow (when compared with interlingual approaches rather than a *direct* method-

[8] http://www.systransoft.com

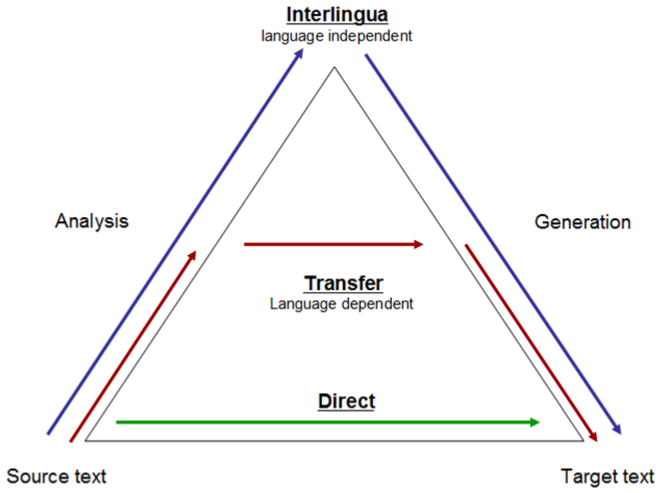


Figure 3.1: The Vauquois Pyramid

ology) and on a syntactic level, often producing constituent structure-based parse trees. Interlingual approaches tend to enact a deeper analysis of the source language sentence that creates structures of a more semantic nature.

A parse generated in an interlingual approach can be transferred directly into a language-independent semantic representation of the sentence and produce target language translations directly from this intermediate representation. A transfer approach must apply another step, the *transfer* module, before *generation*. This step exploits the linguistic knowledge in a language pair-specific rule database to map the syntactic structure of the source onto the target language. Following this, target language translations can be generated from the results of the rule transfer.

Both methods have their advantages and disadvantages. Transfer approaches, being language-dependent, need to know the source and target languages. This hampers their extensibility as each new language pair requires separate language pair-specific transfer modules. These can be expensive and time-consuming to create. In addition, monolingual grammars for both languages are required. On a more positive note, the inclusion of a grammar for the target language does increase the likelihood of grammatical output and the better the analysis and generation modules are, the simpler the transfer modules become.

Interlingual approaches have the advantage of being language-independent making them more extensible to new language pairs as only the analysis and generation modules need to be added. One of the main arguments against interlingual approaches is based on the Sapir-Whorf hypothesis that there is no real 'universal' language and issues such as cultural differences and how people and languages carve up the world differently can hamper the creation of such an intermediate 'language'.

The following sections discuss systems using the paradigms described above in an SL context. Unless further detail is provided, it may be assumed that the structure of the systems as labelled 'transfer' or 'interlingua' are in line with the above description.

3.2.1 The Zardoz System

One of the first significant attempts at using MT approaches for the translation of SLs was undertaken by Veale et al. (1998) under the title of the Zardoz system. This interlingual approach, in the domain of health, describes a multilingual translation system using a blackboard control structure (Cunningham and Veale, 1991; Veale and Cunningham, 1992) for translating English text into Japanese SL (JSL), ASL and ISL. The development of this system is based on extending Artificial Intelligence (AI) research to cross-modal translation.

In the system, signs are notated using a glossing process in a linear string format consisting of English stems preceded by the SL in capitals (e.g. ASL–) with features such as locational information and NMFs included. NMFs are considered important in the implementation of this cross-modal system.

The system is composed of various 'panels' on the blackboard structure that are linked by different processors that feed from and into each panel producing a flow of information through the architecture. An illustration of this is shown in Figure 3.2 (Veale et al., 1998).

The analysis phase is composed of three stages (Roman numerals in brackets refer to the corresponding section of the architecture): the input data is lexically

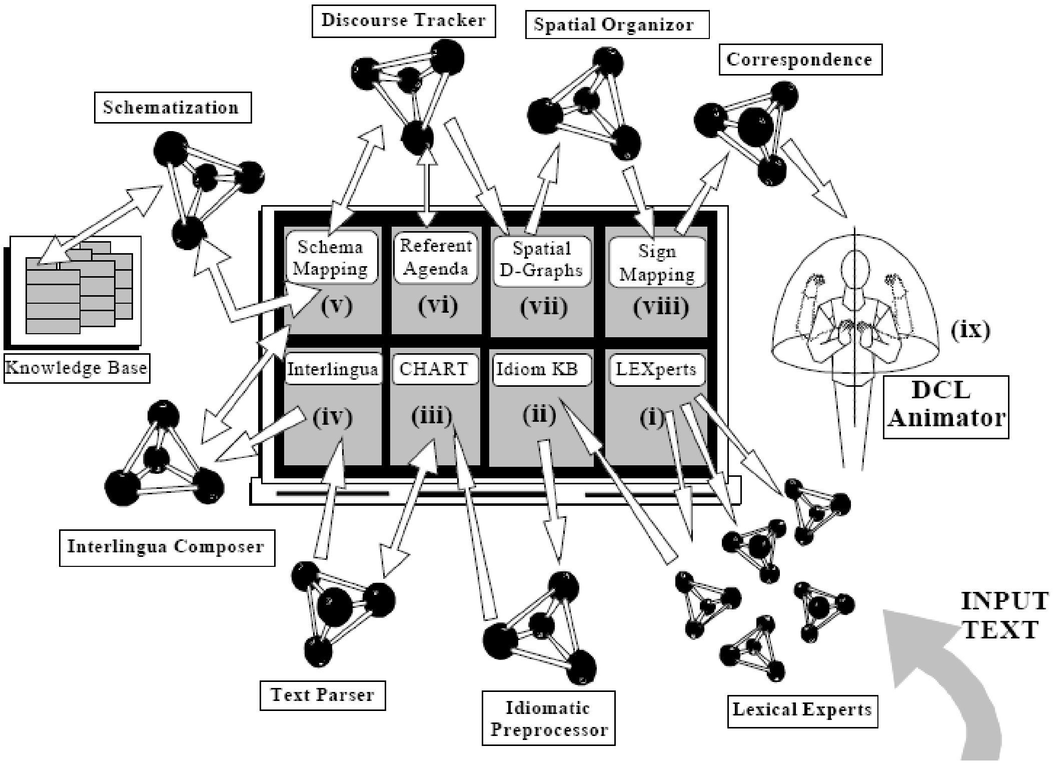


Figure 3.2: The Zardoz System Architecture

analysed for compound word structures (i), idiomatically reduced (ii) then finally parsed using a PATR-based unification grammar (iii). This parse creates the initial interlingua (iv) consisting of a deep syntactic/semantic representation of the input.

This initial interlingua is merged into an interlingual frame (v). Chunking is performed to remove metaphoric and metonymic information to make the representation more language-independent. Next, discourse entities are tracked and anaphora resolutions are determined (vi) before spatial dependency graphs (vii) are used to construct the correct sign syntax (viii) in a linear order for gesture production.

For generation, the tokens in the revised interlingual structure are mapped in concept-to-sign correspondences using a look-up table and heuristic measures. This linear output stream of sign tokens is formed into a 'Doll Control Language' that animates an avatar to produce the output (ix). This animation process is discussed in more detail in Chapter 6.

The primary focus of this work is the extension of AI research to SLs for a complete text to avatar system including NMFs to address the challenges of cross-modal translation, rather than producing a system for evaluation and use. The language-independent processing in this interlingual system allows for the extension of the system to other sign languages although this extension is hampered by the requirement of additional lexicons. Even within a restricted domain the construction of sign dictionaries remains a problem, with only 100 signs in the ISL and ASL dictionaries used in this system.

3.2.2 Albuquerque Weather

Grieve-Smith (1999) describes a transfer-based system for translating English into ASL within the domain of weather forecasting. The restricted domain choice affords little variation in sentence structure and vocabulary, thereby reducing the number of rules and the likelihood of unseen input. The system architecture is outlined in Figure 3.3.

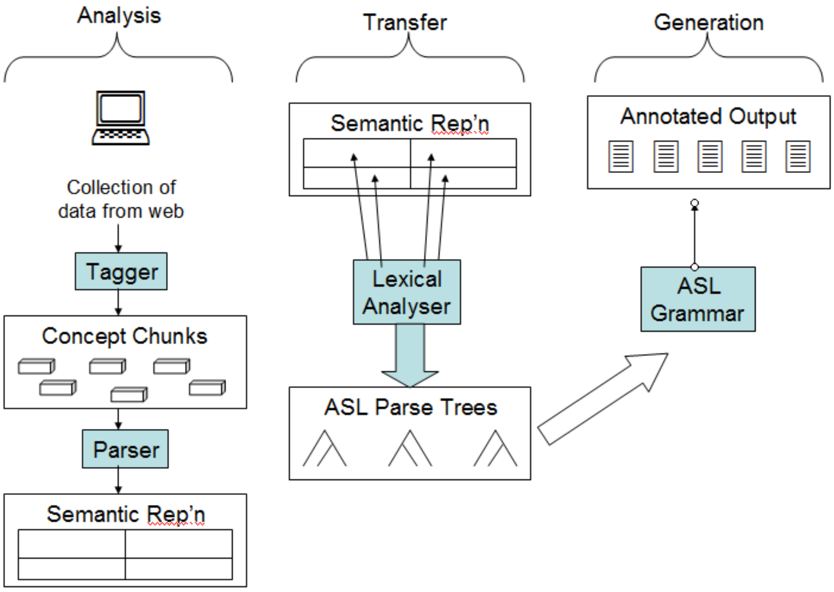


Figure 3.3: The Albuquerque Weather System Architecture

The analysis module creates semantic representations rather than syntactic structures. In terms of SL data format, Grieve-Smith makes use of Literal Orthography (Newkirk, 1986), a coded glossing technique which uses the letters of the alphabet to represent phonetic features such as handshape, movement and location. For example the notation *'so-bles w'* represents the sign *breezy*. 'so' indicates the hands are making a contrary movement, e.g. the right hand is moving toward the right and the left hand is moving toward the right. 'bl' denotes the handshape for the number 5 in ASL (all fingers spread out). 'e' means the hands move toward the non-dominant side of the body and 's' means the movement is repeated. The 'w' is a novelty added by Grieve-Smith to support the lack of NMF capabilities of the notation system. It indicates pursing of the lips. This method displays close to the full range of lexical and classifier signs of ASL although it does not cater for NMFs so these were added by Grieve-Smith along with finger spelling features.

The system follows the usual methodology of a transfer-based system. Weather data are collected from the web and tagged with lexical categories such as 'determiner' and 'preposition', but the input data is also categorised in terms of 'precipitation', 'sky', and 'time', for example. This results in the formation of concept chunks where, in some cases, words or phrases are tagged as one chunk, e.g. 'rain showers <Precip>'.[9] A recursive descent parser[10] then produces a semantic representation based on four primary domain-specific semantic components: *sky*, *precipitation*, *wind* and *heat*.

The transfer module takes the form of a simple look-up table where the semantic representations are mapped onto an ASL parse tree. The look-up table consists of a lexical analyser where each item has an idiomatic translation in ASL Newkirk notation. A formulaic ASL phrase is then created in the generation phase where an ASL grammar is employed to produce grammatically correct ASL output in annotated format.

Other than the creation of the system, no experiments are documented and Grieve-Smith states that the purpose of the work was to show that such an approach was feasible. Despite this, the work was not evaluated although manual assessment is suggested for future work. While the system described in (Grieve-Smith, 1999)

[9] This chunking process can be compared with chunking methodologies explored in Chapter 4 and 5.

[10] http://search.cpan.org/dist/Parse-RecDescent/

does not include a description of an avatar, (Grieve-Smith, 2001) details further work using Stokoe Notation (cf. Section 2.3.1) as a base. Previous attempts at MT are not discussed in detail, but merely as an extension to sign synthesis research.

While Grieve-Smith's approach does seem to be promising, it lacks significant development including any evaluation to determine this accurately. The choice of weather as a domain does facilitate a transfer-based approach meaning that only a limited number of rules need be created for the simple structure and short concise sentence structures in the weather reporting domain (cf. the Météo) system (Chandioux, 1976, 1989). However, such a domain is not ideal for SL MT as SLs in general lack the colloquialisms and register of that used in spoken language weather reporting.

3.2.3 The ViSiCAST Project

Another system employing a transfer approach has been developed within the ViSiCAST[11] project (Marshall and Sáfár, 2002; Sáfár and Marshall, 2002; Marshall and Sáfár, 2003). Working with the language pair English–BSL, they collected sign narratives and information presentations as their choice of domain. SL data for the output takes the form of HamNoSys notation (cf. Chapter 2). An outline of the system architecture is shown in Figure 3.4.

Although Marshall & Sáfár employ a transfer approach that is language pair-specific, in the initial analysis phase they do parse to a semantic as well as syntactic level drawing similarities with interlingua systems. They work with the Carnegie Mellon University (CMU) parser (Sleator and Temperley, 1991) which creates linkages that characterise syntactic dependency parses of the English sentences with a user intervention option if multiple parses are offered. Discourse representation structures (DRSs) (Kamp, 1981) are then generated from the linkages which capture

[11] *ViSiCAST* was a project funded by the EU Framework Programme 5 that researched virtual signing technology to improve Deaf accessibility with the support of the UK Independent Television Commission and the British Post Office. The avatar software was subsequently expanded under the eSIGN Project. Information on both are available at: http://www.visicast.cmp.uea.ac.uk/

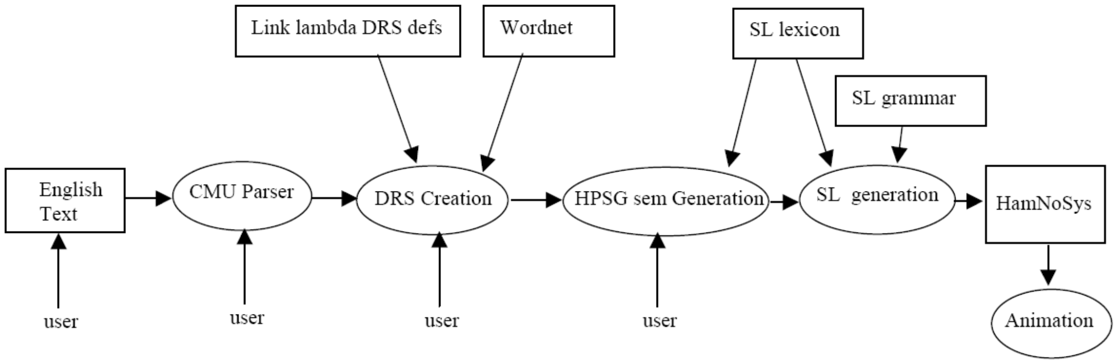


Figure 3.4: The ViSiCAST System Architecture

semantic content in the sentences as well as resolving anaphora.

The transfer module is divided into three phases. First semantic transfer takes place with the English DRS transformed into the BSL DRS. Next, the BSL DRS is converted into a head phrase-structure grammar (HPSG) (Pollard and Sag, 1994) representation. Finally, generation of HamNoSys phonetic descriptions are propagated from the HPSG representations from a bank of 250 lexical items in BSL. There are grammar constraints here that the signs selected must satisfy in order to be designed a valid sequence. These include appropriate structuring such as topic-comment structures and agreement of signing space location for nominals and directional verbs. This agreement is achieved by a model of the signing space in the HPSG structure. The HamNoSys output generated is then animated using an interface developed at the University of East Anglia and an avatar illustrated and developed by Televirtual, Norwich UK. The animation generation phase is discussed further in Chapter 6.

Other than the basic system described above no specific experiments and results are detailed in this work further than "the functionality of the system is demonstrable on a laptop computer" (Marshall and Sáfár, 2003):4.

While the system is successful in its development of a complete text-to-sign MT system, it is hampered by the need for human intervention in many modules. The system does cater for a wide array of SL linguistic phenomenon but it does lack functionality for NMFs, a future direction of work according to the authors.

Furthermore, as we show in Chapter 4 of this thesis, the domain of sign stories is not the most suitable domain for SL MT.

3.2.4 The TEAM Project

The TEAM project (Zhao et al., 2000) was developed using an interlingua-based approach for translating English into ASL. The architecture of the system is shown in Figure 3.5.

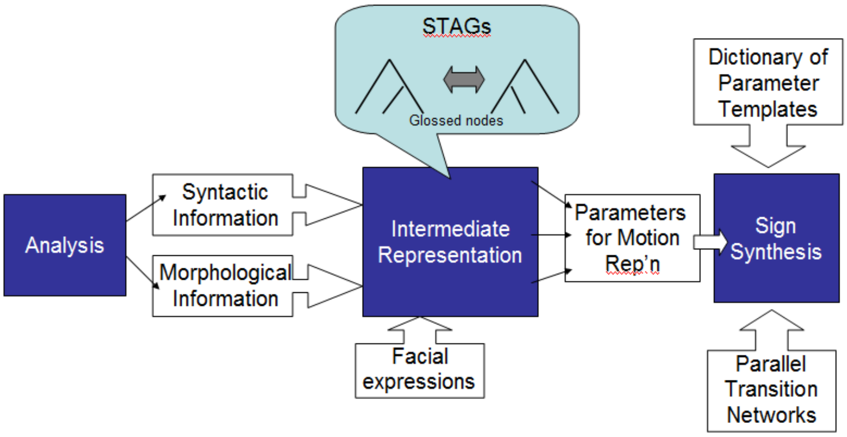

STAGs
Glossed nodes
Dictionary of Parameter Templates
Analysis
Syntactic Information
Morphological Information
Intermediate Representation
Parameters for Motion Rep'n
Sign Synthesis
Facial expressions
Parallel Transition Networks

Figure 3.5: The TEAM Project Architecture

Input data is analysed for syntactic and morphological information and from this an intermediate representation (IR) is generated in the form of gloss notation with some embedded parameters such as facial expressions. During the generation of the IR and the analysis of the English word order, the sign order is also decided. In the next stage, synchronous tree-adjoining grammar (STAGs) (Shieber, 1994) is used to map information from English to ASL. Simultaneous to the parsing of the source language tree, the target language tree can be created using STAG where a complete parse on the English source side means the creation of a complete parse on the ASL target side. Provision is made for grammatical differences between the languages such as negation and topicalisation.

In the second phase of the translation process the information from the IR is fed into the sign synthesiser. A small set of qualitative parameters are interpreted as a motion representation from the IR that can later be converted to a large set of low-level qualitative parameters that control the avatar. A sign dictionary of parameterised motion templates for ASL signs is then used, and default parameters for the avatar model are appropriated to each sign. Parallel transition networks are then used to smooth between signs for generating fluid signed output. Further description of the animation process is given in Chapter 6.

As with many of the other systems described in this chapter, this project does not include information on any evaluations carried out on its performance. However, it must be noted that these systems pre-date automatic evaluation metrics, and that most non-SL MT systems of this time did not include formal evaluations. In any case, as is shown in Section 4.2, the use of automatic evaluation metrics can cause problems for SL MT.

The flexibility of an interlingual approach in terms of extensibility to new language pairs is highlighted here, but these are still within the constraints of general interlingual approaches as outlined previously in this chapter. The description of the system does demonstrate that it can handle SL linguistic phenomena such as topicalisation structures as well as taking visual and spatial information into account, but it fails to completely address NMFs. As described by the authors, this work was a prototype system and within those limitations no experiments were performed so the accuracy and potential usability of the system remains open to question.

3.2.5 South African SL MT

The TEAM project described above is the basis of part of the translation work carried out on South African SL (SASL) in Stellenbosch University (van Zijl and Combrink, 2006; van Zijl and Olivrin, 2008). This system variant differs from the TEAM project's interlingual approach by using a more language-dependent transfer methodology and augments it with the use of STAG. SASL grammar trees and transfer rules were manually constructed from a prototype set of sentences. The database consists of word and phrase lists of hand-annotated SASL videos. An outline of the SASL system architecture (van Zijl and Olivrin, 2008) is shown in

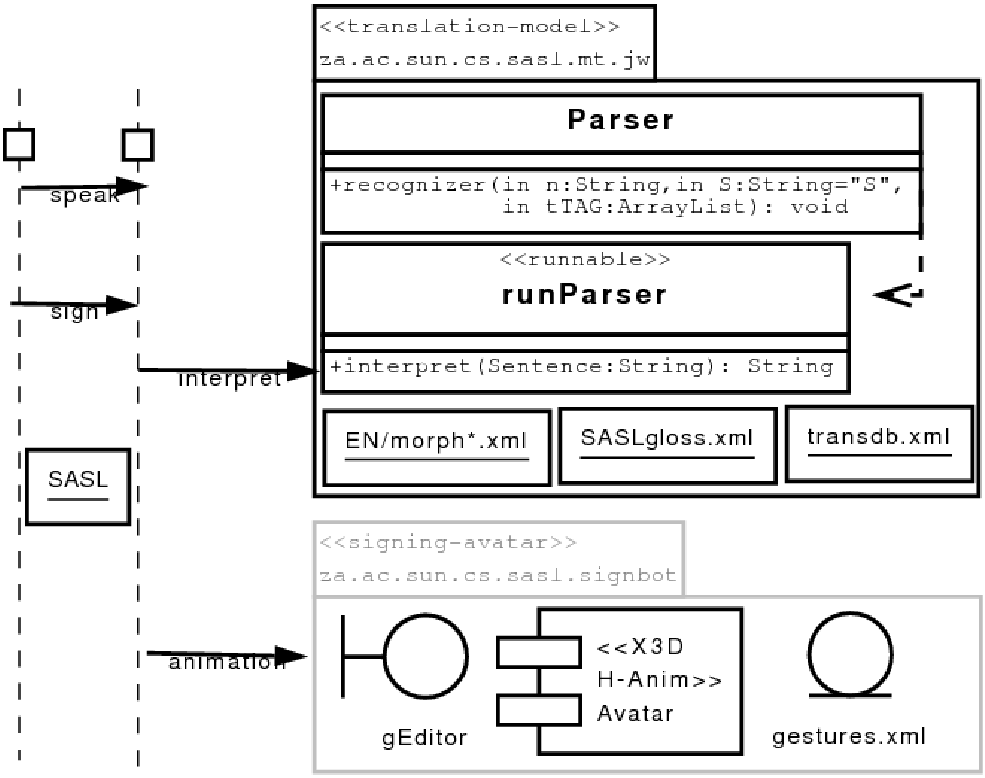


Figure 3.6: The SASL System Architecture

Figure 3.6.

In the initial analysis phase, a GNApp framework (a pre-processing addition in (van Zijl and Olivrin, 2008)) is used for word-sense disambiguation. This takes the form of an interface where English sentences can be entered annotated with POS by the user with the help of SignWriting icons. This annotation is fed into a sentence parser that creates a syntactic parse tree from the tagged English input. At this stage the input TAG syntax parse tree is examined for objects (in this case, the 'objects' are restricted to people) that in SASL may need to be allocated a location in the signing space. Each new object is then assigned a location in the signing space and flagged either as an original location or as a reference point and this assignation continues within each paragraph for agreement with relevant verbs and pronouns. Further analysis is carried out to incorporate NMFs into the system. The input STAG trees are tagged for emotional or expressive content using WordNet (Fellbaum, 1998). These tags take the form of word-level glosses. A further stage of analysis identifies the sentence type, e.g. a question or assertion, to further interpret the possible NMFs appropriate for translation into SASL.

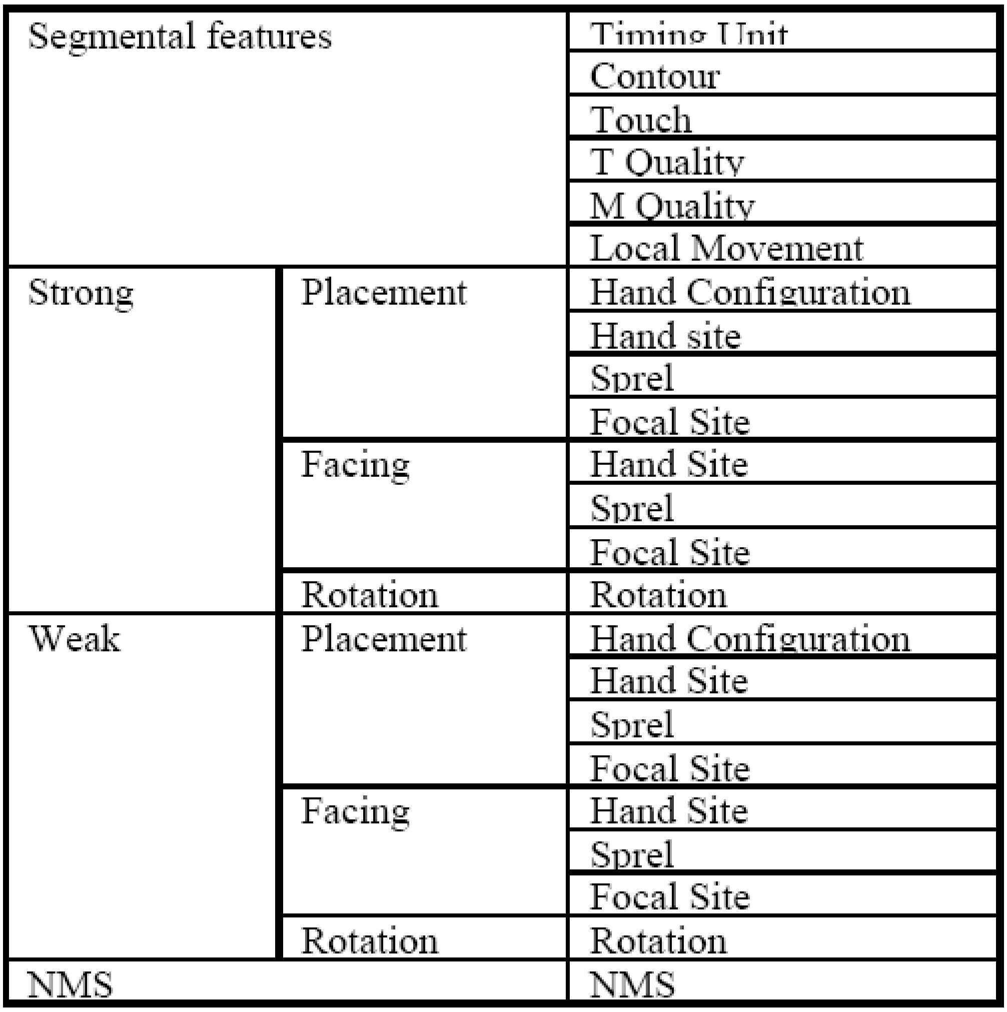
Following this in-depth analysis, transfer from the metadata-annotated English TAG parse tree is mapped onto SASL trees which contain glossed annotation leaves to represent the SL. The generation phase is performed by a mapping of the notation of the SASL trees to a graphical description which is then plugged into a generic signing avatar (Fourie, 2006). This process is described in more detail with an accompanying figure of the avatar in Chapter 6.

Little detail is given of any particular experiments performed other than the details of linguistic analysis of the input in order to address spatial and NMF features of SLs. Despite the heavy analysis (which remains unfinished), this system does make ample acknowledgment of important SL linguistic features and makes a good attempt at incorporating such features to address these issues. At the same time, the portability and extensibility of the system comes into question with the required level of manually constructed rules and trees. This system shows potential but there is a lot of unfinished work that will require extensive effort before a fully functioning system is developed.

3.2.6 The ASL Workbench

A transfer approach using Lexical Functional Grammar (LFG)[12] is described by Speers (2001) in his dissertation outlining the ASL Workbench. The focus of this work is on the representation of ASL rather than the development of a complete system. Speers adopts the Movement-Hold model of sign notation (Liddell and Johnson, 1989). This method divides and documents signs as a sequence of segments according to their phonological representation where each phoneme comprises a set of features specifying its articulation. Variability within the features accounts for

[12] LFG (Kaplan and Bresnan, 1982) assumes a two-tier syntactic representation for sentences: a constituent structure (c-structure) comprising a phrase structure tree that encodes the order of words and phrases, and a functional structure (f-structure) that encodes the grammatical relations in a sentence.



Segmental features		Timing Unit
		Contour
		Touch
		T Quality
		M Quality
		Local Movement
Strong	Placement	Hand Configuration
		Hand site
		Sprel
		Focal Site
	Facing	Hand Site
		Sprel
		Focal Site
	Rotation	Rotation
Weak	Placement	Hand Configuration
		Hand Site
		Sprel
		Focal Site
	Facing	Hand Site
		Sprel
		Focal Site
	Rotation	Rotation
NMS		NMS

Figure 3.7: An Example of Feature Specification Fields for Movement-Hold Segments

grammatical features and changes in the sentence. Spatial data are also taken into account. An example of the feature specifications for a segment are shown in Figure 3.7. This structure outlines features such as timing, spatial relationship (sprel), hand placement and orientation.

The architecture of the ASL Workbench is shown in Figure 3.8 (Speers, 2001).

In the analysis phase the English input sentence is parsed into LFG structures. LFG correspondence architectures are used in the transfer module to convert the English f-structure in to a proposed ASL f-structure. Then an ASL c-structure is derived in the generation phase with a corresponding p-structure, or phonetic structure, representing spatial and NMF detail using the Movement-Hold notation mentioned above. The system automatically chooses the simplest form of output

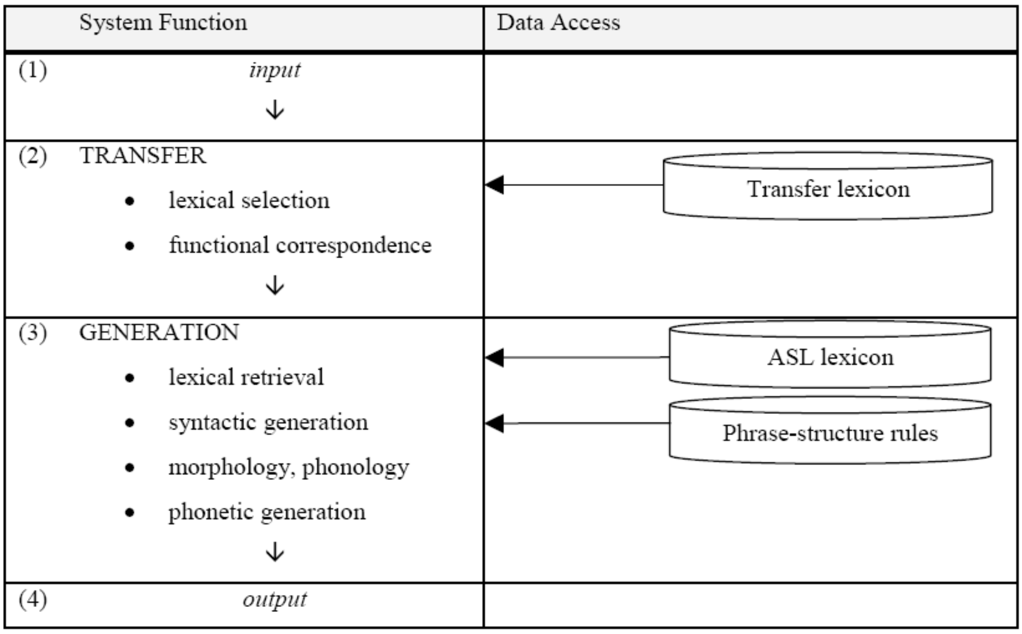


Figure 3.8: The ASL Workbench Architecture

but a human can interact at this point as choose the utterance to be produced. The translated output takes the form of Movement-Hold notation. Speers notes that the output 'document' would ideally be produced in various forms including a gloss-based output, a linguistic notation and animated signing but these have not yet been designed. No specific MT experiments are documented in the dissertation nor are any evaluation methods described which take either automatic or manual assessment into account.

While the notation of this system does manage to grasp important features of SLs such as spatial information and NMFs, they note that it is a complex system that is often difficult to use. Also, retaining this notation for output restricts the usability of the system. As the author states, the output data is only useful if the user understands the notation system and has a grasp of ASL grammar. Further work is envisioned including Stokoe notation output as well as animated sign.

3.2.7 Spanish SL MT

The traditional transfer MT paradigm of text-to-text/-avatar has been extended in (San Segundo et al., 2007) to include a speech recogniser component. In the

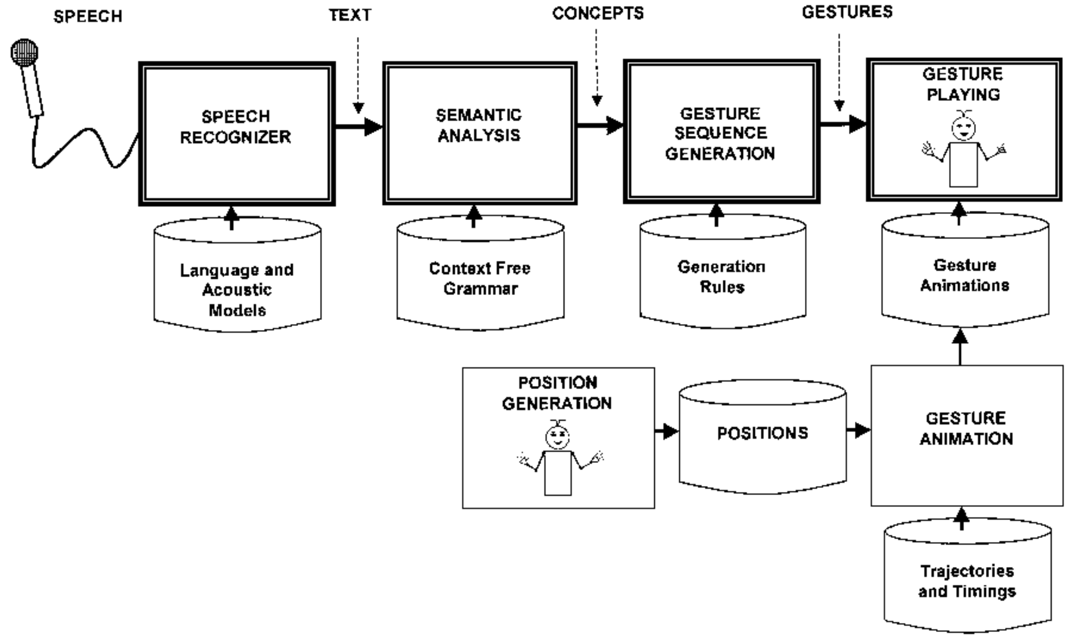


Figure 3.9: The Spanish SL MT System Architecture

restricted domains of railway stations, flights and weather information, this group translate from Spanish speech to a Spanish SL (SSL) avatar.

Outside of the MT system, a speech recogniser, IBM ViaVoice Standard Edition,[13] is used to convert Spanish speech into text. The acoustic and language models of the user are tuned to ensure the most accurate text output. Within the transfer-based methodology, the Phoenix v3.0 parser (Ward, 1991) developed at the Centre for Spoken Language Research, University of Colorado, uses a context-free grammar to parse the input into semantic frames in the analysis phase. Grammar rules and a set of frames must be pre-created and loaded by the system developer. The architecture for the system is shown in Figure 3.9.

In the transfer phase, grammar rules are compiled into Recursive Transition Networks (RTNs) and a top-down chart-parsing algorithm searches for the best set of semantic frames for the input. The more input words accounted for, the higher the score, and thus the most complete, least fragmented parse is derived. The result of the transfer phase is a sequence of parse slots representing the SSL where each slot is a semantic concept.

[13] http://www-306.ibm.com/software/pervasive/embedded_viavoice/

The generation phase is a gesture sequencing phase where SSL gestures are assigned in n:m alignments to the semantic concepts. The rules here are restricted to the sentence types and context of the chosen domain and extension for different domains is handled manually.

In the final avatar production phase, a simplified human figure composed of geometrical shapes has been developed to minimise the effort in gesture design. This figure is 2–D in design with functionality in the fingers, arms and facial features as well as two hair strands to reinforce facial expression. A system developer defines a few basic positions for the signing agent as well as several interpolation strategies which are later used to create continuous signing from the avatar. The animations of the gestures are based on the semantic concepts. This is further explored in Chapter 6.

While no specific MT experiments are noted, this system is one of the few described in the literature that does perform some amount of evaluation. Preliminary human evaluations were carried out where Deaf people were employed to assess the systems output. In preliminary experiments, the avatar produced only the letters of the alphabet. The evaluators found that less than 30% of letters were difficult to understand, i.e. of the 26 letters in the alphabet, approximately 7 were indecipherable on first look. This does not test the functionality of the MT system itself.

The level of human interaction required between the creation of language pair-specific rules and pre-created gesture animations makes the system labour-intensive. The practical usability of a simplified signing avatar, while circumnavigating the complexity of developing a sophisticated human model, does lead to the question: does such a simplified avatar have the functionality to sign accurately and would a Deaf person understand and use it? The addition of the extra speech recognition component, while being an interesting addition that improves the completeness and functionality of the system, multiplies the amount of 'noise' and potential errors between modules and requires user-specific training.

3.2.8 A Multi-path Approach

All the above described systems have employed either a transfer- or interlingua-based approach. One system that stands apart from the others is the multi-path approach of Huenerfauth (2006). In his PhD dissertation, Huenerfauth describes an approach that combines interlingual, transfer and direct methodologies in what he terms a 'multi-path' approach. Although his work primarily focuses on the generation of classifier predicates (CPs) (cf. Chapter 2 Sections 2.2.3 and 2.2.4), he describes the architecture of the English to ASL MT system despite the non-CP elements not being implemented. The architecture for this approach is shown in Figure 3.10

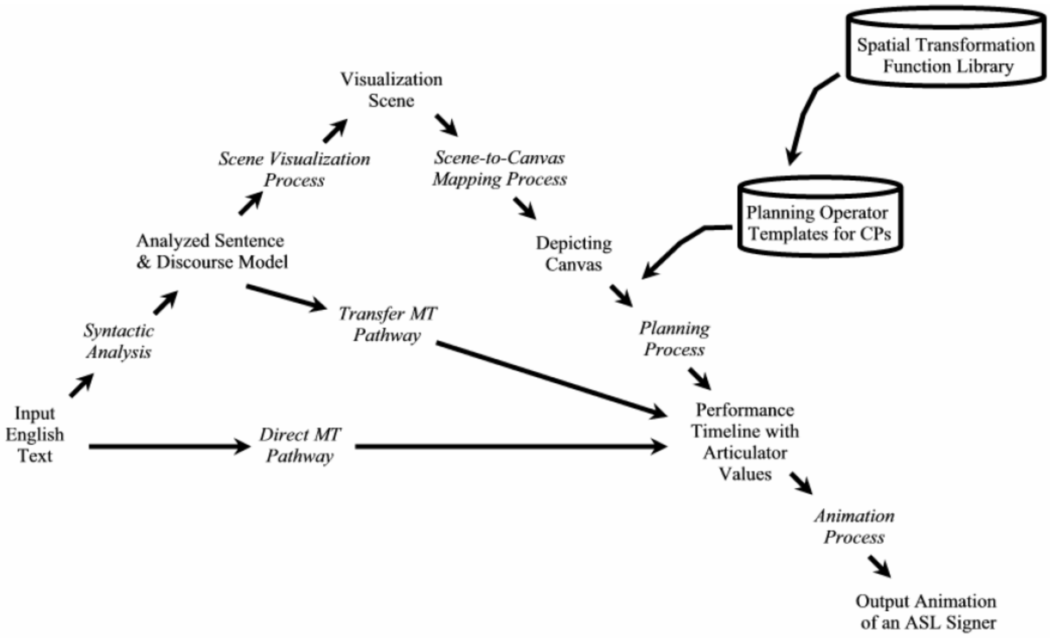

Spatial Transformation Function Library
Visualization Scene
Scene Visualization Process
Scene-to-Canvas Mapping Process
Planning Operator Templates for CPs
Analyzed Sentence & Discourse Model
Depicting Canvas
Transfer MT Pathway
Syntactic Analysis
Planning Process
Input English Text
Direct MT Pathway
Performance Timeline with Articulator Values
Animation Process
Output Animation of an ASL Signer

Figure 3.10: The Multi-path Approach Architecture

Rather than employing linear annotation to represent ASL as most previous approaches have done, Huenerfauth divides the multi-modal ASL string into multiple 'channels' that are hierarchically structured and co-ordinated over time to represent sub-streams of communication. A partition/constituent methodology (Huenerfauth, 2004) is used to create 3–D trees encoding the ASL information sequentially and simultaneously yet not overlapping. SL features such as movement, classifier hand shapes and spatial information are encoded in the representation. A worked example of the model is shown in Figure 3.11 for the sentence "The cat sat next to the house".

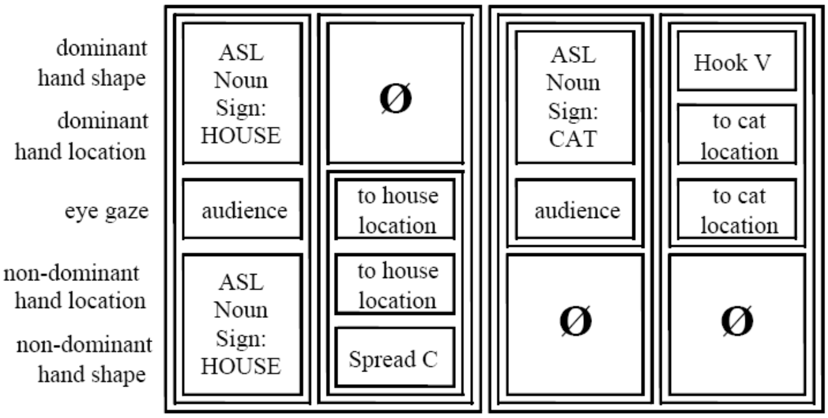


Figure 3.11: An Example of the Partition/Constituent Methodology

The multi-path architecture is used to translate different types of ASL sentences. Focusing on classifier predicates, Huenerfauth proposes an interlingual approach where the interlingua is a 3–D visualisation of the arrangement of the objects in the sentence of English input. In this way the visualisation acts as a semantic representation and a deeper level of analysis than a transfer approach. Discourse entities and CPs are mapped onto the virtual signing space for later generation by a signing avatar.

Generating 3–D visualisation models is computationally expensive so for sentences that would not involve CPs in ASL, a transfer approach is proposed. For input sentences that cannot be handled by the rules in the transfer approach, a direct system is proposed that would produce SEE.

Given that the transfer and direct sections of the system have not been implemented and as the nature of the CP interlingua output is complex, automatic and objective evaluation for comparison with other systems is not possible. Native ASL signers evaluated the CP animations rating the animations on a scale of one to ten for understandability, naturalness of movement, and ASL grammatical correctness. Signers were also asked to compare the output against animations created by a native ASL signer performing CPs wearing a motion-capture suit, animations created by digitizing the coordinates of hand, arm, and head movements during the performance, and using the information to create an animation of a virtual human character. Ten CP animations generated by the prototype system described above were chosen to represent a variety of CPs. 15 participants evaluated these sentences along with parallel videos of 10 Signed English animations and 10 motion-capture animations. Huenerfauth states that the CP animations scored higher than the other two animations for all judging criteria. This is discussed in more detail in Chapter 6

The main focus of Huenerfauth's work is the CP generation component for an MT system that could be layered on top of existing transfer-based systems. While interlinguas are computationally expensive, particularly if they involve 3–D representation models as is the case in this approach, Huenerfauth takes this into account and proposes a back-off method allowing for 'simpler' approaches to be taken when possible. This also adds a robustness to the approach. Given the complexity and language-specificity of the CP interlingual approach, it does mean it falls down in terms of extensibility to new language pairs and there is little room for independent and objective evaluation.

3.3 Data-driven Approaches

The proliferation of empirical technologies in language and computing research, coupled with the failure of rule-based methods to build robust, extensible and broad-coverage translation systems has paved the way for empirical MT approaches. Advancements in computing in terms of speed of processing and the amount of machine readable text available means that is is easy to collate statistics on the empirical approaches to MT problems.

Data-driven approaches began to gain ground in the 1990s and now dominate the research field. This approach, often termed 'corpus-based', can be sub-divided into *statistical* MT (SMT) and *example–based* MT (EBMT). Compared to rule-based approaches, there are fundamental differences in both data-driven processes yet they remain inherently similar. In general, linguistic information and rules are eschewed in favour of probabilistic models collected from a large parallel corpus.

Statistical methods are largely derived from the area of speech recognition (Brown et al., 1988, 1990). Essentially, the translation of a sentence is seen as the probability of the most likely target string (e) given the source string (f): $\Pr(e|f)$. The process of deducing a translation can be performed using the *Log-Linear Model* or the *Noisy Channel Model*.

The Log-Linear Model (Koehn, 2004) estimates $\Pr(e|f)$ directly from the parallel training corpus using the calculation in Equation 3.1.

$$argmax_e \Pr(e|f) = e^{\sum_i \lambda_i h_i(e,f)} \tag{3.1}$$

λ_i defines the model parameters and h_i defines the feature functions such as phrase- and word-translation probabilities.

The Noisy Channel Model (Brown et al., 1990) is a special case of the Log-Linear Model that employs Bayes' Theorem to decompose the translation problem $\Pr(e|f)$ as shown in Equation 3.2.

$$\Pr(e|f) = \frac{\Pr(f|e). \Pr(e)}{\Pr(f)} \tag{3.2}$$

This equation multiplies the likelihood of a source sentence being a translation of the target by the likelihood of the target string being a correct string. Commonly, the denominator ($\Pr(f)$ - the probability of the source sentence) is ignored as it is assumed that this string is correct. Bilingual corpora are employed to provide a source for estimating the probabilities. These probability estimations can be broken down into a *translation model* ($\Pr(f|e)$) and a *language model* ($\Pr(e)$). The translation model estimates its parameters using alignments derived from the parallel training corpus. These alignments can be on a word- or phrase-level. The language model requires only a monolingual corpus and uses n-gram models to calculate the string probability.

The Noisy Channel Model of the Log-Linear equation is modified as shown in

Equation 3.3 (Och and Ney, 2002).

$$argmax_e \Pr(e|f) = e^{1.log(\Pr(f|e)+1.log(\Pr(e))} = \Pr(f|e). \Pr(e) \tag{3.3}$$

In order to perform translation the Viterbi algorithm (Viterbi, 1967) is often employed in the *decoder* in order to search for the most likely candidate translation from the set of candidate strings. Translation of a sentence involves the calculation of the fertility probability (how likely a word is to translate as any number of other words in the target text), the probability of each word pair and the probability of distortion.[14] The string (or *n*-best list of strings) with the highest probability is deemed the candidate translation.

Example-based methodologies, although similar, employ an analogy principle comparing the input sentence against the bilingual corpus looking for candidate translations. Three steps are generally assumed in translation:

1. finding the closest matches for the input against the source side of the parallel corpus and retrieving the corresponding target language aligned translations;

2. finding translation links on a sub-sentential level;

3. recombining target language segments for output translation.

In Chapter 4, for our experiments, we describe a basic EBMT system that uses this architecture and in Chapter 5, we describe a more sophisticated SMT system with EBMT modules.

Both methodologies have their advantages and disadvantages. Both require large amounts of bilingual data to be available from which to extract probabilities or examples. In the case of spoken language translation large amounts of data are becoming increasingly available, although data collection remains a problem for SL MT (cf. Chapter 4). Data-driven approaches tend to be consistent, objective,

[14] 'Distortion' refers to movements of words or strings in a sentence to account for differing word orders in the language pairs, e.g. some languages are verb-initial, some verb-final.

language-independent and involve little human intervention, although they often suffer from sparse data given their dependence on large texts. We will examine these issues next by discussing those data-driven attempts at SL MT which have been published to date.

3.3.1 Translating from German SL

The first statistical approach to SL MT emerged in Germany with the novel translation direction of SL to spoken language text using DGS and German. The language direction here is the reverse of previous approaches due to this system's focus on recognition technology as a component of the broader MT architecture. Within the domain of 'shopping at the supermarket', Bauer et al. (1999) employ gesture recognition technology to a database of signed video for training which is then fed into an SMT system. The architecture of the system is shown in Figure 3.12.

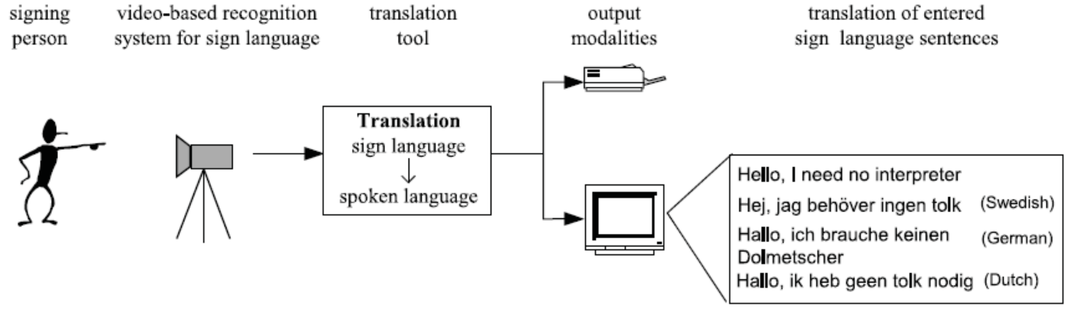

signing person
video-based recognition system for sign language
translation tool
output modalities
translation of entered sign language sentences
Translation
sign language
↓
spoken language
Hello, I need no interpreter
Hej, jag behöver ingen tolk (Swedish)
Hallo, ich brauche keinen Dolmetscher (German)
Hallo, ik heb geen tolk nodig (Dutch)

Figure 3.12: The DGS SL Recognition and MT System Architecture

First, the video-based recognition tool takes 6 hours of signed video for training and extracts features such as the size and shape of the hands and body. All features are then entered into a feature vector that reflects the manual sign parameters. The recognition tool uses these feature vectors from each frame of the video and enters them into a Hidden Markov Model (HMM) (Rabiner and Juang, 1989) where each sign is modelled with one HMM. Based on the trained dataset, an input sentence of a signed video is entered and the best match found resulting in a stream of recognised signs being produced and converted into a meaningful sentence in German. The architecture of the MT engine itself is shown in Figure 3.13.

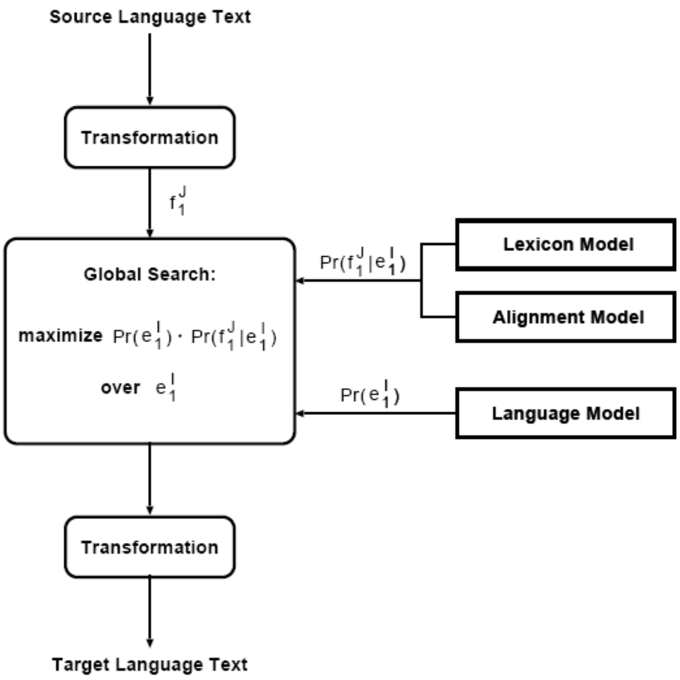


Figure 3.13: The DGS MT System Architecture

The translation process is a straightforward SMT approach with a translation model, consisting of a lexical model and an alignment model, and a language model that extracts probabilities from the target sides of parallel texts. Each model feeds into a global search engine that uses Bayes' decision rule (cf. Equation 3.2) to find the best match. The decoder is based on a dynamic programming algorithm.

No MT evaluations were performed but the recognition tool was tested using one hour of video data within the same domain. Bauer et al. report that for 52 signs they achieve a recognition accuracy of 94% and a score of 91.6% for 100 signs. From this the authors surmise that the translation tool would achieve a word accuracy of 90%. They also deduce that the recogniser is able to handle the word order of SLs and can recognise continuous SL.

The novel reversal of the traditional SL MT translation direction addresses the hearing–Deaf directionality of a bilingual translation system and introduces the addition of an SL recognition module. Given that this is the first reported attempt at data-driven MT for SLs, it is unfortunate that no formal evaluations were carried out, but it must be noted that this research was undertaken before automatic evaluation methods were developed. However, the reported recognition accuracy results indicate that the addition of this module and the reversal of the translation direction should not adversely affect the SL MT component in terms of additional noise regardless of the capabilities of the MT system itself.

3.3.2 RWTH Aachen University Group

Another more recent group working on SMT for SLs is headed by Prof. Hermann Ney at the RWTH Aachen University (Stein et al., 2006). Working primarily with the German–DGS language pair, this group has attempted translation both to and from SLs. Initial experiments on DGS data were performed within the domain of weather reports (Stein et al., 2006) and subsequent work has addressed the more practical domain of airport information announcements[15] (Stein et al., 2007). A gloss annotation is used as a linear semantic representation of the sign language being used. They follow the glossing convention of conveying the meaning of a sign using the uppercase stem form of words in the spoken language. NMFs and features such as repetition are included in this annotation. For the weather data, Stein et al. (2006) report a dataset of 2468 bilingually aligned sentences in German and DGS annotation.

The baseline MT engine used for their SL translation is the phrase-based SMT system developed at RWTH (Matusov et al., 2006). The language model employs trigrams that are smoothed using the Kneser-Ney discounting method (Kneser and Ney, 1995). The translation model is a phrase-based model. Monotone search is used to find the best path as well as various reordering constraints (Kanthak et al., 2005). Such constraints address the variation in word order of different languages and involve acyclic graphs that allow limited word reordering of source sentences. Pre- and post-processing steps are also employed to prepare the data for translation and to augment the raw system output to improve evaluation, respectively. The system architecture diagram used by this group is a copy of the DGS architecture

[15] This is parallel data to the set used in experiments in Chapter 5.

of Bauer et al. (1999) shown above in Figure 3.13. Despite this, the authors claim their system is not derived from the work of Bauer et al..

The progress of their system is assessed in a set of experiments investigating reordering constraints as well as pre- and post-processing steps. The first set of experiments involving morpho-syntactic pre-processing of the data uses the gerCG parser[16] to identify the parts of speech of the input and then transform nouns into stem form, split German compound words and remove German parts of speech not commonly found in DGS. Discarding unnecessary parts of speech reduces the baseline of 48% WER[17] by almost 8% and splitting compound nouns lowers the baseline by approximately 10%. Stemming the nouns lowered the number of out-of-vocabulary words. Post-processing steps that add in discourse entities such as stored people and places are re-entered after translation but these do not affect automatic scores.

Experiments using three variations of reordering constraints showed that the local constraint that allows each word in the sentence to be moved a maximum of w-1 (where 'w' indicates the window size) steps to the beginning or end of the sentence is the most successful approach with a window size of 2 yielding an almost 10% improvement on error rates.

Further reordering constraint experiments were performed in later research, Morrissey et al. (2007b) and (Stein et al., 2007), using air traffic information data in parallel with our own data-driven experiments. This joint research is described in more detail in Chapter 5.

In a separate phase in Stein et al. (2006), a signing avatar was plugged into the back end of the system to visualise the textual output. The avatar was developed as part of the ViSiCAST system (Elliott et al., 2000) (cf. Section 3.2). A subjective manual evaluation was carried out on the signing avatar by deaf adults. 30 test sentences and 30 reference sentences were evaluated on a scale of 1-5. Average

[16] http://www.lingsoft.fi

[17] Word error rate: calculates the number of correct words in the correct word order. See Chapter 4 for a more detailed explanation.

results were 3.3, just slightly above the mean mark on the scale. The authors report that these results correlate with the automatic evaluation scores, but by using a 1–5 scale scoring system it is possible that evaluators chose not to rate the translations either too positively or negatively and take the middle ground. A more discerning score could have been obtained using an even-numbered scale.

This approach primarily investigates SMT applications to SL MT rather than trying to be a complete system. The addition of an avatar in Stein et al. (2006) is a functional yet poorly supported extension. However, the scores of the automatic and manual evaluations may be considered comparable and their approach does manage to take linguistic phenomena into account.

3.3.3 Chinese SL MT

A more recent development by Wu et al. (2007) describes a hybrid transfer-based statistical model for translating Chinese into Taiwanese SL (TSL). In one of the largest datasets noted to date, the authors cite a bilingual corpus of 2,036 sentences of Chinese with parallel annotated sign sequences of corresponding TSL words from which CFG rules are created and transfer probabilities derived. A Chinese Treebank containing 36,925 manually annotated sentences is also noted and both corpora are used to derive a probabilistic context-free grammar (PCFG).

Wu et al. describe a three-stage process in their translation model. This is shown in Figure 3.14. First, the Chinese input is segmented into word sequences, and then analysed and parsed into a set of possible phrase structure trees (PSTs), statistically modelled by the PGFGs of the Chinese Treebank. This gives P(C|$\tilde{T}_C$), where C is the Chinese string and $\tilde{T}_C$ is the Chinese PST. In parallel, a constrained POS back-off model is used to handle out-of-rule problems. In what is termed the ‘translation model’ and is akin to a transfer phase, transfer probabilities (P($\tilde{T}_C$, $\tilde{T}_S$), where $\tilde{T}_S$ is the TSL PST) are derived from the parallel corpus and these are used to weight the Chinese CFG rules. These rules are then used to create the TSL CFGs and subsequently the TSL PSTs. P(S|$\tilde{T}_S$) denotes the TSL sequence-generation

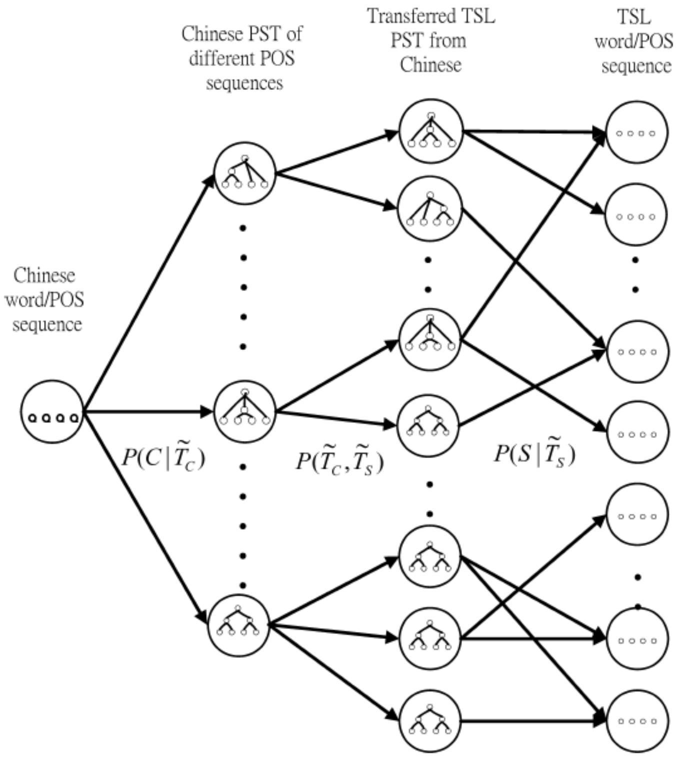

Chinese PST of different POS sequences
Transferred TSL PST from Chinese
TSL word/POS sequence
Chinese word/POS sequence
$P(C|\tilde{T}_C)$
$P(\tilde{T}_C,\tilde{T}_S)$
$P(S|\tilde{T}_S)$

Figure 3.14: The Chinese SL MT Architecture

probability, where S denotes the TSL sentence. Here the Chinese CFG rules are altered to reflect the word order of TSL and the altered rules are re-assigned as TSL rules. Chinese CFG rules derived from the Chinese Treebank that do not occur in the bilingual corpus are directly assigned as TSL rules, notwithstanding possible shifts in word order.

The Viterbi algorithm is used to deduce the optimal TSL word sequence and best translation. This is carried out by calculating the probabilities of the three translation stages, namely the Chinese generation probability given the Chinese PST (P(C|$\tilde{T}_C$)), the PST transfer probability between Chinese and TSL (P($\tilde{T}_C|\tilde{T}_S$)), and the TSL sequence-generation probability for a given the TSL PST (P(S|$\tilde{T}_S$)). There is no described avatar used with this system and the nature of the output is not detailed in their work but it is assumed it takes the form of a sequence of TSL signs/words.

Similar to the other SMT methodologies cited here, the authors have undertaken automatic and manual evaluation of their work. Two sets of experiments were performed. The first concerns the POS back-off model for addressing the sentences that cannot be converted to PST using PCFGs and reports scores of up to 100% when back-off is applied to both nouns and verbs for forests of candidate PSTs of up to 500 trees. For the translation evaluation, four metrics are employed: three automatic metrics, namely Alignment Error Rate (AER) (Och and Ney, 2000), Top-N (Karypis, 2001), and BLEU (Papineni et al., 2002) for an objective assessment of the system's translation; and one manual, subjective evaluation based on mean opinion score (MOS). The system developed by the authors is compared with IBM Model 3 (Och and Ney, 2000) for each evaluation where 80% of the corpus was used for training and the remaining 20% for testing. This system is shown to outperform IBM Model 3 across the board for their Chinese–TSL translation with comparative scores of AER of 0.09% compared to 0.225% for IBM Model 3; for Top-N, correct translation rates of Top-1 achieved 81.6%, compared with the 73.7% of IBM Model 3; BLEU scores showed scores of 0.86 for the proposed model and 0.8 for the IBM Model 3. BLEU scores were calculated using only one reference text.

For the subjective manual evaluations, two groups of 10 people were chosen, one group of fluent TSL users that were hearing and another that were deaf. Evaluation took the form of assigning 'good', 'fair' or 'poor' marks to the 20 chosen sentences. General reading comprehension assessments were also performed with more than three quarters of the sentences deemed comprehensible and approximately half of the output translations considered good by each group.

While attention is given to the use of grammars and fusing transfer and statistical methodologies, no mention is made of addressing any TSL linguistic phenomena nor of any annotation methodologies. It is interesting to note the high automatic evaluation scores given that spoken language translation systems with years of research behind them often do not attain such high scores. One would also question the subjective manual evaluations and how accurate a measure of a translation system can be derived from evaluating unnatural (and unspecified) text versions of TSL. In Chapter 6, we discuss this issue further and show that manual evaluation is best reserved for judging SL animations.

3.4 Summary

In this chapter we introduced the two central paradigms on which SL MT to date has been based: rule-based approaches and data-driven approaches. From Section 3.2 we can see that various rule-based approaches, both transfer and interlingua, have been adapted to tackle SL MT with varying degrees of success. Similar issues to those found in general MT involving this type of approach can be seen to arise here too. The transfer approaches described have shown to be restricted by the need for language pair-specific linguistic rules, as is typical of all transfer approaches. In SL MT, this has posed a problem as there is often little grammar development or linguistic knowledge resources available for the SL being used (cf. (Veale et al., 1998; van Zijl and Combrink, 2006)). The need for language pair-specific development also affects the extensibility of these transfer systems to new language pairs as new sets of transfer rules and lexicons would need to be developed for each language pair making the process computationally expensive. A further drawback to the transfer approaches is the need for human involvement, which, as in the choosing of the correct rules or structures (San Segundo et al., 2007), can slow down the translation process. It also makes the system less user-friendly as users would be required to have linguistic knowledge of the languages at hand.

Huenerfauth (2006) proposes a multi-path approach where interlinguas are only required for SL sentences that have CPs and claims that transfer and direct approaches are sufficient for other SL strings. Interestingly, many of the transfer approaches described propose a deeper processing of source input during analysis, often on a semantic level akin to an interlingual approach while remaining language pair-specific (Grieve-Smith, 1999; San Segundo et al., 2007). This indicates that researchers using transfer approaches find the surface syntactic level of parsing too shallow for dealing with SLs and their multi-modal structure. This may be attributed to the requirement of SLs to be represented and produced on a phonological rather than morphological level, as is the case with languages that have writing systems. Comparatively, interlingual structures seem to be capable of representing SL notation and accompanying simultaneous features such as NMFs. In saying this, using interlinguas to represent SLs and their features does challenge the notion of a truly 'language universal' intermediate representation of languages, particularly when the language used to represent the SL is typically a modified version of the source language. This calls into question the ability of one language to accurately represent another language, and particularly in the case of SLs, to represent another language of a different modality.

In general, the linguistic dependencies of rule-based approaches, whether language-dependent or not, hamper the flexibility and objectivity of SL MT. These dependencies, whether for grammar or lexicon creation, restrict the extensibility of these systems to new language pairs and translation directions.

Furthermore, an accurate assessment of the achievements of these systems is not possible given the lack of objective, and in most cases even subjective, evaluation. As automatic evaluation metrics were not in common usage for *any* MT during the development of these systems, no objective translation scores were available. In many cases, the expected accuracy or usability of the translations produced were not discussed. This makes it difficult to assess the dependability of rule-based approaches for SL MT.

Comparatively, an overview of data-driven approaches to SL MT was given in Section 3.3. The three systems described may be differentiated from each other not only by the approach used but by the fact that they have employed automatic evaluation metrics to assess the accuracy of their translations, with the exception of Bauer et al. (1999) where only the output from the recogniser is formally evaluated.

Little can be deduced from the system described by Bauer et al. (1999) given that no experiments were performed nor any MT evaluations carried out. The Chinese system (Wu et al., 2007) demonstrates that a combined transfer and statistical approach is possible and can lead to high translation quality based on three automatic evaluation metrics. However, the issue of needing to supply language pair-specific knowledge for the transfer modules remains. Such a dependency restricts the flexibility and portability of a potentially successful approach.

The RWTH system (Stein et al., 2006, 2007) is a prime example of the positive features of data-driven systems. Their experiments have shown that SMT systems are flexible and may be easily adapted to new language pairs and changes of translation direction. They are able to handle grammatical features specific to SLs and work on small datasets. As with most SMT systems, their work requires a number of pre- and post-processing steps outside of the translation phase, *per se*, to handle the data.

While the prerequisite of a bilingual corpus can cause some problems for SL MT (cf. experiments in Chapter 4), once a suitable data set is found it does provide a set of reference translations to facilitate automatic evaluation. Furthermore, the flexibility and adaptability of these approaches facilitate the easy extension of such systems to bidirectional and multi-lingual translation, when compared with a linguistics-heavy rule-based approach. Coupled with the successful data-driven experiments described in this chapter using relatively small data sets, we are confident that a data-driven approach is the most suitable and practical methodology for SL MT, as we demonstrate in the following chapters.

A comparative overview of each system and its features is shown in Table 3.1. In the column 'Approach', 'I' refers to an interlingua system, 'T' refers to a transfer system, and 'S' indicates a statistical system. A $\surd$ indicates that the system meets the criteria of the column in which it appears.

The systems described in this chapter are by no means complete but they do show the potential for using MT technology to break down some of the language barriers for SLs. Outlining these systems highlights problems that arise during SL MT that are not adequately tackled by previous approaches. These include:

- **Notation:** adequate notational representation of SLs,

System	Approach	Languages	Direction	Notation	Domain	Evaluation	Human	Bidirectional	Extensible	Animation
				Rule-Based MT Systems						
Zardoz	I	(A,J,I)SL	to SL	Gloss	Doc–patient					√
Albuquerque	T	ASL	to SL	Newkirk Not.	Weather					√
ViSiCAST	T	BSL	to SL	HamNoSys	Narrative		√			√
TEAM	I	ASL	to SL	Gloss	n/a					√
SASL	T	ASL	to SL	Gloss	Doc–patient					√
ASL W'bench	T	ASL	to SL	Move–Hold	n/a		√			
Spanish SL	T	ASL	to SL	Gloss	Wthr/travel		√			√
Multi-path	I&T	ASL	to SL	Part/Cons	n/a	Manual				√
				Data-Driven MT Systems						
German SL	S	DGS	to DE	Feature Vector	Shopping			√	√	
RWTH	S	DGS/ISL	both	Gloss	Wthr/travel	Auto & Man		√	√	√
Chinese	T&S	TSL	to SL	Gloss	General	Auto & Man				√

Table 3.1: Comparison of Related Research

- **Flexibility:** the need for a system to address the multiple language pairings within this visual-gestural modality,

- **Linguistic Phenomena:** the need for a system that can handle the spatial nature of SLs coupled with language-specific features such as NMFS,

- **Data:** the small amount of data and linguistic knowledge available for SLs,

- **Accessible Output:** the need for an appropriate system output that is accessible to the Deaf given the lack of a writing system, namely real SL,

- **Evaluation:** the employment of evaluation techniques to assess and compare the success and usability of each system.

In the remaining chapters of this thesis, we will demonstrate that data-driven methodologies coupled with an appropriate data set are capable of addressing and handling these problems where other systems have failed. We support these claims using formal automatic and manual evaluation experiments.

Chapter 4

Problems for Sign Language Machine Translation

Given the relative novelty of SL MT when compared with MT for spoken languages, applying MT technology that is primarily focussed on text-based data to a visual-spatial language such as an SL will inevitably lead to some problems.

In this chapter, we discuss the main problematic issues involved in SL MT, namely those concerning data and evaluation. We demonstrate where these problems arise in SL MT research by discussing experiments we have performed using a prototype data-driven MT system.

Having outlined the problems, we illustrate how, in the initial stages of our research, we dealt with these issues and were able to overcome some of these problems.

4.1 Data

The manual modality of SLs and the lack of a standardised writing system contributes to the limited availability of SL data both in terms of desired quantity and quality for use in a data-driven SL MT system.

A prerequisite for any data-driven approach is a large bilingual corpus aligned at sentence-level from which to extract training and testing data. The larger 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 strings, which correspondingly increases the chances of improving system output. For translation between major spoken languages, such data is available in large amounts; in the National Institute of Standards and Technology (NIST) evaluation[1], we used approximately 4 million aligned English–Chinese and English–Arabic sentence-pairs to seed our MaTrEx system. 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.

Finding a corpus to suit the data needs of an SL MT system is a difficult task. As noted in Chapter 2, there is no standardised written version of SLs. As a result the standard format for SL data is usually videos of human signers. While much of the video data is made for private use in educational contexts, such as signed video examples for instruction or as deliverables for homework, there are some available in DVD format for learning SLs[2] and others for linguistic analysis purposes (as discussed in the following sections). Although there are some data freely available, generally that which is procurable amounts to only a few hundred sentences or even just isolated words, and nothing on the scale of spoken language data resources. This is partially a result of data collection for SLs being time-consuming and expensive in comparison to resourcing written text data for spoken languages. At the very least, SL data collection involves finding native signers, organising elicitation tasks or interviews and orchestrating videography with correct lighting, visual clarity and recording equipment.

It is conceivable that collating the complete sentences from numerous SL learning DVDs could result in a few hundred sentences from which a database may be created. Furthermore, the tendency of language instruction material on these DVDs to remain

[1] http://www.nist.gov/speech/tests/mt/

[2] http://www.forestbooks.com/pages/Categories/DVD.html

in a restricted domain with few topics (e.g. introducing yourself, giving directions) could justify its use in a data-driven approach to MT. However, the videos of SLs provided in these DVDs is not in a format that is easily accessible. As the description of research in Chapter 3 has shown, MT systems require a textual representation of SLs such as annotation or some other symbolic encoding. While it could be possible to extract the recorded sentences from the DVDs and have them annotated or encoded, it would be a lengthy and expensive process.

In contrast, SL video that that has been created for linguistic analysis purposes is usually annotated in some way. Three such major projects have annotated SL data available that have the potential to seed a data-driven MT system:

- Signs of Ireland Corpus,
- American Sign Language Linguistic Research Project Corpus,
- European Cultural Heritage Online Corpus.

4.1.1 Signs of Ireland Corpus

The Signs of Ireland corpus (Leeson et al., 2006) was developed as part of the "Languages of Ireland" programme at the Centre for Deaf Studies in the School of Linguistics, Speech and Communications, Trinity College, Dublin. Over a period of 3 years, from 2004 to 2007, video data of 40 Deaf ISL users was collected by a native Deaf woman involved in both the Irish Deaf community and the data collection project. The signing participants comprised both male and female native signers of ages varying between 18 to 65 from all over Ireland. Participants were asked to tell a personally selected narrative, a children's story called "The Frog", and an elicitation task based on that devised by (Volterra et al., 1984) that uses pictures to obtain certain types of utterance. While the final amount of data in terms of the number of sentences is not known, early estimations following video collection noted approximately 20 hours of video. Over the course of the project, the videos were annotated using ELAN software (described in Section 4.1.3 below) using glossed notation of signs, NMFs and an English translation.

4.1.2 American Sign Language Linguistic Research Project Corpus

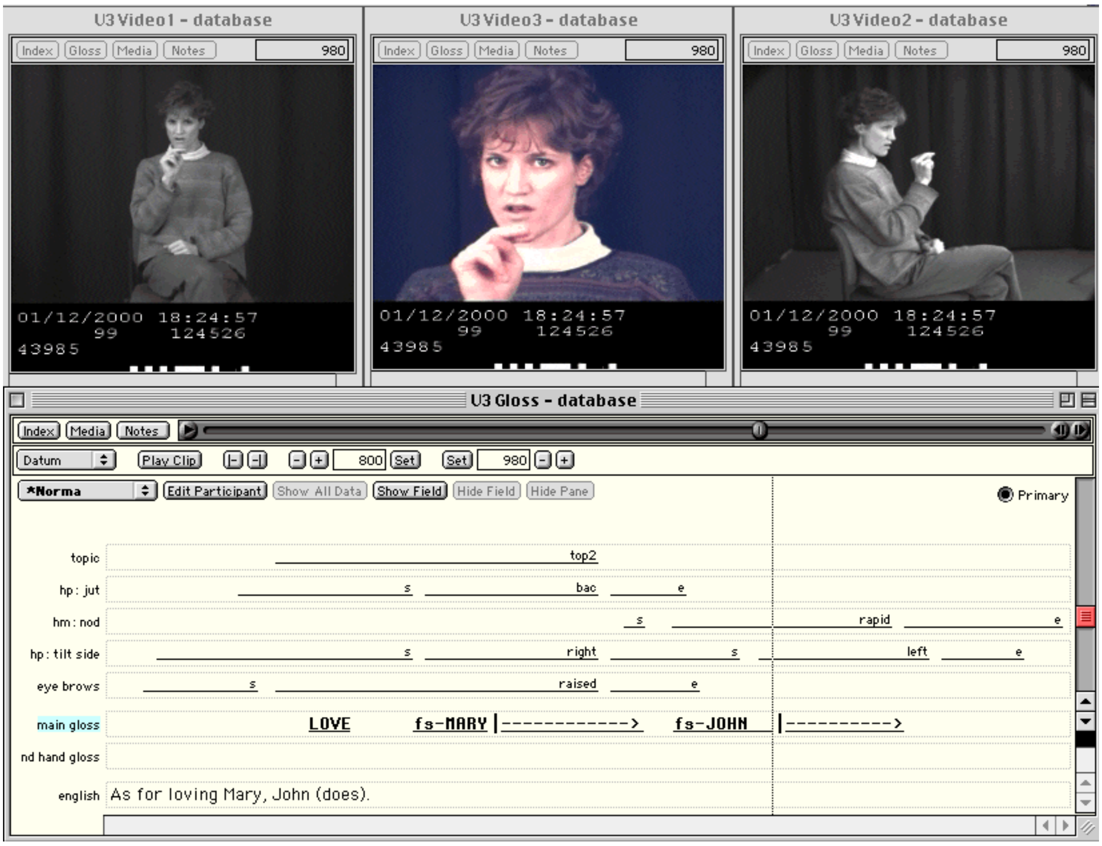
A corpus for linguistic analysis has also been created as part of the American Sign Language Linguistic Research Project (ASLLRP)[3] led by Prof. Carol Neidle of Boston University. ASL video data collection began in 1999 in Boston University and Rutgers University. Signers were captured on digital video cameras in 4 different simultaneously recorded views. The data consists of elicited sentences aimed at obtaining certain sentence structures in ASL, sentences within a fixed vocabulary for computer vision research, short stories, dialogues between two signers, and multiple views of different hand shapes. The project is still under development with over 3,200 elicited utterances currently available as well as the short stories. Most of the data available is annotated using SignStream[4] (Neidle et al., 2001), an interface that has been created and developed within the ASLLRP for manual annotation of video data. The interface is shown in Figure 4.1.

The annotation tool allows the user to view different videos of the same sequence (appearing in the top half of the screen) simultaneously with an accompanying annotation (in the lower half of the screen). The annotation fields may be chosen by the user and appear on the lower left-hand side of the screen with accompanying annotations added along a time-line adjacent to and in line with each annotation field. The data currently annotated includes glosses of the ASL signs, NMFs, some grammatical markers and an English translation.

Up to the Autumn of 2007, the only annotated data available was 1,600 elicited sentences of unspecified domain. SignStream and its annotated data is available on CD-ROM or to download online with a standard fee being charged for the tool itself

[3] http://www.bu.edu/asllrp/

[4] http://www.bu.edu/asllrp/SignStream/



topic: top2
hp : jut: s bac e
hm : nod: s rapid e
hp : tilt side: s right s left e
eye brows: s raised e
main gloss: LOVE fs-MARY |------------> fs-JOHN |---------->
nd hand gloss:
english: As for loving Mary, John (does).

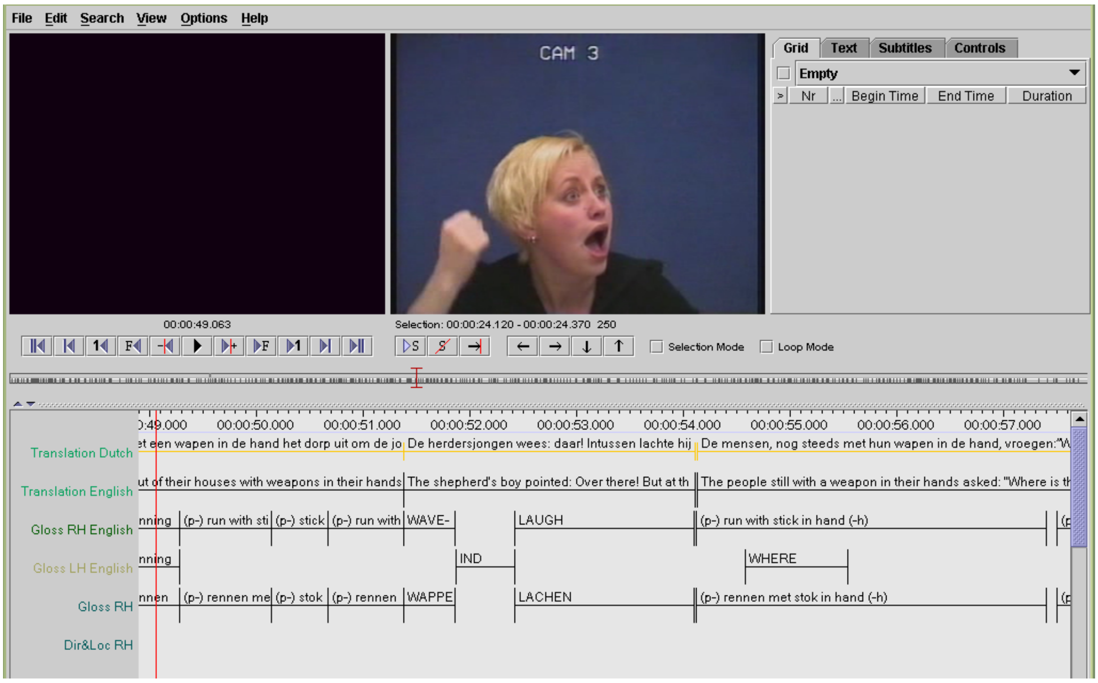
Figure 4.1: Screenshot of the SignStream Annotation Interface

and any additional annotated data. SignStream is currently restricted to a Mac operating system.

4.1.3 European Cultural Heritage Online Project Corpus

A third project that provides fully annotated digitised SL video data suitable for use in MT systems is Case Study 4 of the European Cultural Heritage Online (ECHO) project.[5] This is an EU-funded scheme based in the Netherlands that has made fully annotated digitized corpora of Dutch, British and Swedish SLs freely available on the Internet. The project began in 2003 and has collected annotated SL videos from the Netherlands, UK and Sweden. The data consists of the same five children's stories signed in each language as well as other vocabulary lexicons, poetry, and some interviews. For most data segments there are two movie files available that have been shot simultaneously, a whole upper body video and a close up of the face.

[5] http://www.let.kun.nl/sign-lang/echo/



Translation Dutch: et een wapen in de hand het dorp uit om de jo | De herdersjongen wees: daar! Intussen lachte hij | De mensen, nog steeds met hun wapen in de hand, vroegen:"W

Translation English: ut of their houses with weapons in their hands | The shepherd's boy pointed: Over there! But at th | The people still with a weapon in their hands asked: "Where is th

Gloss RH English: nning | (p-) run with sti | (p-) stick | (p-) run with | WAVE- | LAUGH | (p-) run with stick in hand (-h) | (p

Gloss LH English: nning | IND | WHERE

Gloss RH: nnen | (p-) rennen me | (p-) stok | (p-) rennen | WAPPE | LACHEN | (p-) rennen met stok in hand (-h) | (p

Figure 4.2: Screenshot of the ELAN Annotator Interface

The children's stories number approximately 500 sentences in each language, with the other data varying in content and amount. These data are manually glossed (cf. Section 2.3.4) using a tool for gesture researching: the EUDICO Linguistic Annotator (ELAN).[6] This tool is the standard for creating and developing SL corpora, and was specifically designed for the language and gesture analysis (Leeson et al., 2006). The annotation software's interface is shown in Figure 4.2.

The videos are shown on the top half of the screen with the annotation panel below and a time-line in between. The annotation panel allows for multiple levels of transcription depending on the needs of the user. The different annotation fields appear vertically to the left of the panel and their respective annotations stream horizontally. Similar to SignStream, this tool allows the user to view SL video and annotations simultaneously. Previously described annotations may be edited and new SL videos can be annotated from scratch with the user deciding on the granularity of the coverage of the annotation fields. The same transcription conventions

[6] http://www.mpi.nl/tools/elan.html

are maintained for each dataset and each language.[7]

4.1.4 Choosing Suitable Data for SL MT

The primary data-related problems facing corpus-based MT are that of sparseness, inaccessibility and unsuitable formats. The three projects outlined above circumvent these problems by each providing at least a couple of hundred SL sentences made available on the Internet complete with transcribed annotations. While the cumulative data in each of these projects is respectively small when compared with the large datasets currently feeding spoken language, the data is sufficiently large to seed a prototype data-driven MT system as we will show in Section 4.3. Requirements for data-driven MT should include the following criteria:

- an adequate amount of data to seed a system,
- occur in a restricted domain to maximise the repetition of phrases and words for the development of statistical weights,
- the data should be consistently annotated using transcription conventions,
- the data should be in an easily accessible format for sentence extraction in the pre-processing phase.

Each of the projects described have enough data to begin data-driven MT experimentation with a view to increasing the amount of data as and when more becomes available. Both the Signs of Ireland and the ASLLRP projects contain data that is without domain restriction and practically any topic could be anticipated. The ECHO project more is suitable for data-driven MT, therefore, given that most of its data is restricted to 5 children's stories, involving a small vocabulary with frequent repetition and short sentences.

All three projects have employed a system of transcription conventions for annotation. Each of them differs but is based on the same principle of linguistic analysis.

[7] http://www.let.kun.nl/sign-lang/echo/docs/ELANtranscr_conv.pdf

The ASLLRP project appears to provide a more heavily-coded linguistic annotation with more information fields than the other two projects. The ECHO project data, while it is linguistically rich, containing NMF and some deictic information, is more straightforward and simple in its annotation. In terms of annotation accessibility for sentence and information extraction, all systems provide text-based file formats that only require some basic pre-processing before they are ready for use in a data-driven MT system.

As we are developing an SL MT system in Ireland, it would be appropriate to make use of ISL data with a view to facilitating and supporting both the Irish Deaf community and SL research in Ireland. Unfortunately, despite the anticipated suitability of the ISL Signs of Ireland project corpus, the data was unfinished and unavailable for use in our research.

The ASLLRP corpus provides the most extensive range of annotated sentences. The data does, however, lack any kind of fixed domain, consisting instead of sentences elicited for their structure rather than content. Furthermore, the corpus is impractical for our purposes as it is only suitable for use on a Mac operating system as annotations are not yet available outside of the SignStream format.

The ECHO project data may not provide as much data in terms of sentence numbers as the two previous corpora, but the domain of the data is somewhat restricted making it more suitable for use in a corpus-based MT system given the increased likelihood of repetition of phrases and words. The project also includes the same data in multiple languages using consistent annotation transcription conventions which could facilitate easy extensibility of the MT system. The annotation data in this project is in an easily accessible XML format making pre-processing of data a reasonably straightforward task. The data is also freely available and annotated using the most widely used transcription tool for SL linguistic analysis. Furthermore, the annotation includes NMFs, classifier and deictic references (cf. SL linguistic phenomena described in Chapter 2). For these reasons we selected this data as the most suitable candidate for our initial SL data-driven MT experiments.

4.2 Evaluation

Within the last 7 or so years, MT system development has typically involved the use of automatic evaluation metrics to assess and compare the accuracy of the system. In some cases, manual evaluation by linguists or native speakers is also carried out. Consequently, developers of SL MT systems should seek to evaluate them in the same fashion. However, the visual-spatial nature of SLs and the lack of a writing system can make this less straight-forward.

There are a number of automatic evaluation metrics that are frequently used within MT research for evaluating text-based output. String-based matching metrics such as BLEU (Papineni et al., 2002), and metrics that calculate the rate of errors such as Word Error Rate or Sentence Error Rate are among the most commonly used. Manual evaluations carried out by a panel of human judges are also often used to assess translation quality as a supplement to automatic evaluation, or where automatic evaluation is not possible, e.g. if no reference text is available.

Translating from an SL into spoken language text should afford no further problems for automatic evaluation nor manual evaluation than would occur for non-SL MT systems as the output of both systems is text-based. But there is the problem of there being less of a demand for this directionality. There is more requirement for spoken language text to be translated into SLs, than the other way around. Translating from spoken language text into any SL, on the other hand, can cause problems, given the non-standardised format of SLs. Output can take the form of annotated glosses or of avatar videos, each posing a potential evaluation problem. The multiple fields of the annotated format require that 'gold standard' reference translations match the candidate translations in terms of annotation granularity and scope. For avatar videos there are currently no automatic methodologies for evaluation. In light of these issues, we next outline automatic and manual evaluation methodologies and discuss their suitability for SL MT.

4.2.1 Automatic Evaluation

Automatic evaluation metrics are designed to assess linear text output, requiring the provision of at least one gold standard version of the testing data as a reference for comparison. The majority are string-based matching algorithms that do not take syntactic or lexical variation into account. Some of the more widely used metrics include:

- **Bi-Lingual Evaluation Understudy (BLEU)** score: 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. A higher score indicates a more accurate translation.

- **Sentence Error Rate (SER)**: computes the percentage of incorrect full sentence matches by comparing the system's candidate translations against the reference translations. With all error rates, a lower percentage score indicates better candidate translations.

- **Word Error Rate (WER)**: computes the distance between the reference and candidate translations based on the number of insertions, substitutions and deletions in the words of the candidate translations divided by the number of correct reference words.

- **Position-independent Word Error Rate (PER)**: computes the same distance as the WER without taking word order into account.

- The **Meteor** Automatic Machine Translation Evaluation System (Banerjee and Lavie, 2005): performs two stages of comparative matching for candidate and reference translations. Exact matching of unigrams, and stemmed matching, where remaining unmatched words are decomposed into stems using the Porter stemmer and subsequently form matches. Stem matching and synonym matching are based on WordNet models. Scores are obtained by calculating the sum of n-gram matches.

- The **General Text Matcher (GTM)** (Turian et al., 2003): bases its evaluations on accuracy measures such as recall, precision, and F-measure.

- **Dependency-based evaluation** (Owczarzak et al., 2007): employs LFG dependency triples using paraphrases derived from the test set through word/phrase alignment with BLEU and NIST. It evaluates translations on a structural rather than string level and allows for lexical variance.

For spoken language text output such as English sentences, the evaluation process is simple. The user need only apply the required candidate and reference translation files and literally click a button to get an estimated calculation of the accuracy of the system's candidate translations.

This process is more complex when dealing with SL output. The prerequisite of a gold standard text can pose problems for this directionality (Morrissey and Way, 2006) (cf. experiments in Section 4.3.4). An increased number of reference texts increases the chances of better evaluation scores as there are more variations of translations to compare. Due to the time-consuming and laborious annotation process, the chances of sourcing more than one set of reference translations for SL data is decidedly low.

Furthermore, all gold standard reference texts would need to be of the same granularity of annotated description to be able to form a fair comparison. If the candidate translations had only glosses of the hands but the reference translations included NMF or phonetic information, an accurate evaluation score could not be calculated as the texts would not be comparable. Given that SL annotation can contain multiple fields of description, as described in Chapter 3, even if both candidate and reference data contain the same fields, it is doubtful that an accurate evaluation could be carried out as the data would not be in a linear format, but rather multi-levelled. The metrics mentioned above are equipped to handle linear strings of words and not multi-level composite SL data and are, therefore, only useful for SL MT if the output is of a consistent linear format that is comparable with the format of the reference translations. This is possible if only SL glosses are used, as we shown in Chapter 5.

While automatic evaluation metrics have limited usage for annotated output, these methodologies are neither suited nor equipped to handle non-text data, namely a signing mannequin in avatar format. Unfortunately, comparative 'quick and easy' automatic gesture recognition and evaluation research is not yet sophisticated enough to offer a reliable evaluation tool for SL MT avatar output. For this reason human evaluation is a more viable option.

4.2.2 Manual Evaluation

While there are obvious issues involving the use of automatic MT evaluation metrics with SL MT output, it is, of course, still possible to evaluate the candidate translations manually. A panel of human evaluators with native knowledge of the target language can be asked to assess the output translations based on a prescribed set of criteria noting scales of accuracy and fluency. Manual evaluation is described in more detail in Chapter 6 where we employ this methodology to assess our animated avatar translations.

One of the main problems with manual evaluation is that it can be a laborious process. Rather than being able to have evaluation scores in a matter of minutes at the click of a button, system developers must find willing and able evaluators and draw up evaluation criteria, score sheets and possibly a questionnaire. Using labelled or unlabelled annotation output for evaluation requires the evaluators to be familiar with not only the target language, but the linguistics of that language. They must also have some familiarity with the transcription conventions to be able to best evaluate the translations. There may be difficulty in deciding if the coded output could potentially be an accurate SL sentence or whether it is just good coded output in comparison with other coded output.

Manual evaluation of a signing avatar, on the other hand, is somewhat more reliable as it is the real language that is being evaluated and not just a representation. It can also be quicker as evaluators are only required to assess what is in front of them, not what could possibly be produced from what is in front of them. In addition, all that is needed is a good knowledge of the target language. In saying this, manual evaluation of an avatar is not without its problems. For the most part, avatar technology is merged with MT technology to produce signing mannequins from the MT translations. This combination of systems could increase the risk of errors in the data to be evaluated as any MT errors would be multiplied by any errors from the avatar production. Furthermore, the evaluators approval or not of the avatar itself can also be a factor, as we show in our evaluation experiments in Chapter 6. For example, the MT data may have the potential to be close to accurate but the shortcomings of the avatar technology may mean the sentence could receive a lower score than it deserves. Therefore, manual evaluations cannot be seen as evaluating just the MT system itself, but rather the larger system. In saying that, this type of evaluation does permit an evaluation of the real sign language, as opposed to a textual representation. We suggest that asking someone to manually evaluate annotated output to assess if it is good SL would be akin to asking someone to look at English text output and determine if it would produce good speech. For this reason, we contend that each component of an SL MT system should be evaluated using the methodology most appropriate to the output produced. In the following sections, we discuss automatic evaluation in practice.

4.3 A Prototype EBMT System for Sign Language Machine Translation Illustrating Data and Evaluation Problems

In order to assess the suitability of the chosen ECHO data for data-driven MT and to demonstrate the evaluation problems, we describe the prototype EBMT system we designed and constructed for this purpose (Morrissey and Way, 2005, 2006). We first specify the system itself, then explore data and evaluation problems through a series of experiments.

4.3.1 Dataset Construction

To construct our dataset, we assembled Dutch Sign Language/Nederlandse Gebarentaal (NGT) data from the ECHO project as described previously. The data consists of a selection of the children's stories *Aesop's Fables* signed by various native signers as well as some NGT poetry. At the time of construction this was the largest annotated dataset available with 561 sentences and an average sentence length of 7.89 words (min. 1 word, max. 53 words). The sign language side of the corpus consists of annotations that describe the signs used in the video in gloss format, as well as NMF detail and some grammatical information. As the English translation annotation field and the other annotation fields are time-aligned according to the video sequence, sentence alignments were easy to extract automatically. The raw annotation text file is pre-processed and the data is sifted into a file of SL annotations and a corresponding file of English translations. These form the bilingual datasets which are split into training and testing sets.

4.3.2 System Description

The prototype system we developed was a basic EBMT system (cf. Chapter 3) that required a bilingual dataset (Morrissey and Way, 2005). The system used a similarity metric based on Levenshtein's distance metric (Levenshtein, 1966) to search the source side of the bitext for 'close' matches and their translations as well as for determining sub-sentential links on the target side. The retrieved target language substrings were then recombined into an output translation of the source string. A diagram of this process is shown in Figure 4.3.

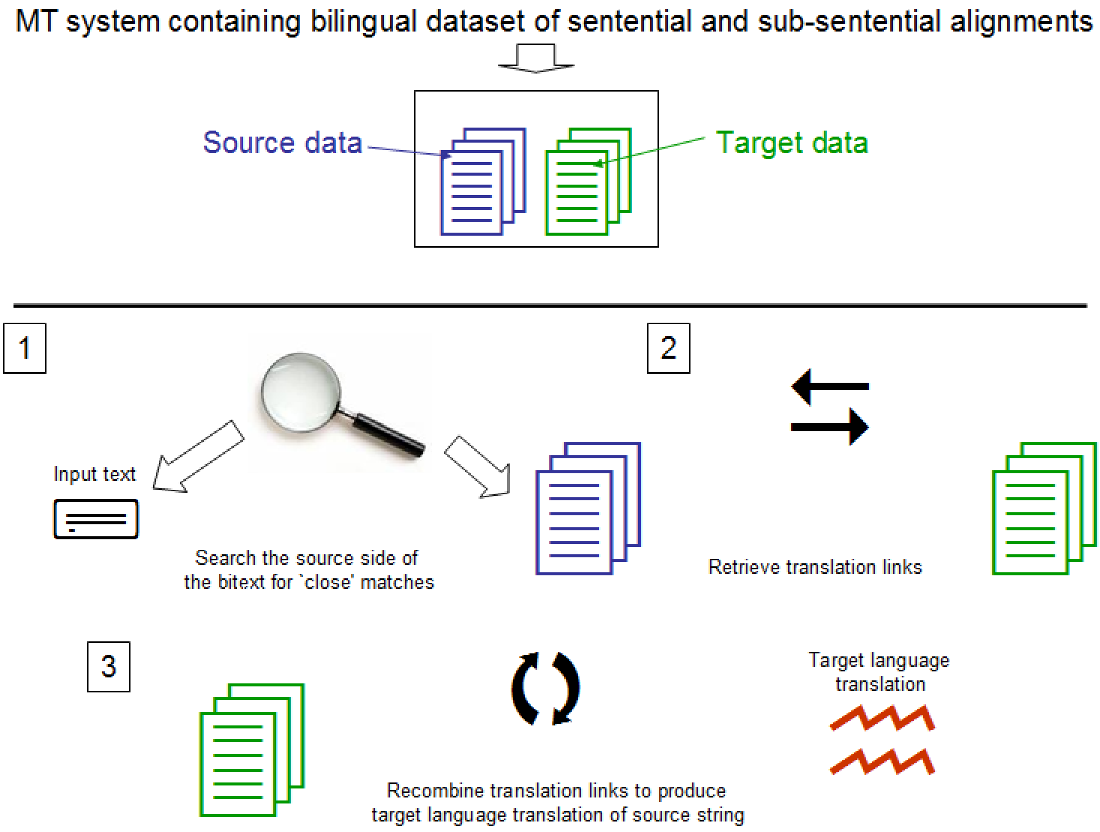

MT system containing bilingual dataset of sentential and sub-sentential alignments
Source data
Target data
1
Input text
Search the source side of the bitext for `close' matches
2
Retrieve translation links
3
Recombine translation links to produce target language translation of source string
Target language translation

Figure 4.3: Outline of the Translation Process of Our Prototype EBMT system

The processes were based on the system of (Veale and Way, 1997; Way and Gough, 2003, 2005) whose work employs the 'Marker Hypothesis' (Green, 1979) to sub-sententially segment data. The Marker Hypothesis is a universal psycholinguistic constraint which posits that languages are 'marked' for syntactic structure at surface level by a closed set of specific lexemes and morphemes. In a preprocessing stage, Gough and Way (2004b) use 7 sets of marker words for English and French (e.g. determiners, quantifiers, conjunctions etc.), 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.

In order to describe this in more detail, we revisit an example from Gough and Way (2004a), namely:

(2) each layer has a layer number ⟹ chaque couche a un nombre de la couche

(2) shows sententially aligned English and French sentences. The Marker Hypothesis tags marker words in each string with their relevant POS label, as shown in (3).

(3) <QUANT> each layer has <DET> a layer number ⟹ <QUANT> chaque couche a <DET> un nombre <PREP> de la couche

Each source chunk is then aligned with a corresponding target chunk based on the POS tags and lexical similarities. These sub-sentential chunk alignments are shown in (4), where we can see that n:m alignments are permitted.

(4) a. <QUANT> each layer has: <QUANT> chaque couche a

b. <DET> a layer number: <DET> un nombre de la couche

In addition, marker templates can also be produced in this process. A marker template is a chunk where the marker word has been replaces with its tag. This increases the likelihood of matches where there are minor differences, e.g. *the layer number* and *a layer number* would have produce the same marker template of *<DET> layer number*. This also serves to increase the robustness of this approach. An example of generalised templates produced from (4) are shown in (5).

(5) a. <QUANT> layer has: <QUANT> couche a

b. <DET> layer number: <DET> nombre de la couche

On the sign language side it was necessary to adopt a different approach as a result of the sparseness of the English closed class item markers in the SL text. This is normal in SLs, where often closed class items are not signed, as is the case with many determiners, or are subsumed into the sign for the neighbouring classifier as is sometimes the case with prepositions (Emmorey et al., 2005). Initially experiments were performed on different divisions of the SL annotations. The NGT gloss field was segmented based on the time spans of its annotations. The remaining annotations in other fields were then grouped with the NGT gloss field annotations within the appropriate matching time frame. In this way, these segmentations divided the SL corpus into concept chunks. Upon examination these concept chunks were found to be similar in form to the chunks that were formed using the the Marker Hypothesis on the English text and suitable for forming alignments, thereby providing a viable option for chunking the SL side of the corpus.

The following example shows segments from both data sets and their usability for chunk alignment. (6) shows the results of the different chunking process on both sentences, (6a) being taken from the English chunking process and (6b) from the SL chunking process. In both cases angle brackets denote a chunk demarkation. The text in round brackets in the SL text denotes the field name from which the annotation is taken. In the SL chunks there is a certain amount of information encoded in letters and symbols, for example: (p-) indicates a classifer sign, '`closed-ao`' in the ***(Mouth)*** field indicates the aperture of the lips and mouth, '`p`' in the ***(Cheeks)*** field indicates a puffing of the cheeks.

(6) a. `<DET> the hare takes off <PREP> in a flash.`

b. `<CHUNK> (Gloss RH English) (p-) running hare :`

`(Mouth) closed-ao :`

`(Mouth SE) /AIRSTREAM/ :`

`(Cheeks) p :`

`(Gloss LH English) (p-) running hare :`

`(Gloss RH) (p-) rennen haas :`

`(Gloss LH) (p-) rennen haas :`

<CHUNK> (Gloss RH English) FLASH-BY :

(Gloss RH) VOORBIJ-SCHIETEN :

(Mouth) closed, forward :

(Mouth SE) /PURSED/ :

(Eye gaze) rh

(7) shows specific chunks that can be successfully aligned following the chunking process, (7a) being taken from the English chunked text and (7b) from the SL chunked text. Angled brackets contain the markers, round bracketed text names the field, the remaining text is the annotation content of that field and each field is separated by a colon.

(7) a. <DET> the hare takes off

b. <CHUNK> (Gloss RH English) (p-) running hare :

(Mouth) closed-ao :

(Mouth SE) /AIRSTREAM/ :

(Cheeks) p :

(Gloss LH English) (p-) running hare :

(Gloss RH) (p-) rennen haas :

(Gloss LH) (p-) rennen haas

The main concept expressed in (7a) and (7b) is the running of the hare. The English chunk encapsulates this concept with the words *the hare takes off*. This same concept is expressed in the SL chunk in the combination of annotations. The '`Gloss RH English`' and '`Gloss LH English`' show the running of the hare and the additional semantic information of the effort involved in *takes off* as opposed to running at ease is expressed in the NMF fields with the indication of puffing of the cheeks (`p` in the `Cheeks` field) and the closed mouth with breath being exhaled (`closed-ao` and `/AIRSTREAM/` in the `Mouth` and `Mouth SE` fields respectively). Despite the different methods used, they are successful in forming potentially alignable chunks.

4.3.3 Experiment One: Assessing the Suitability of a Small Dataset

At the beginning of the experimental phase of the system development, to assess the progress, we prepared some preliminary tests. Test sets were manually constructed in four groups of ten sentences. This is unusually small for a test set when compared to larger MT systems for spoken language and the usual 90:10 training:testing division. However, it serves the purposes of our experiments and subsequent manual evaluation. The groups are as follows:

(i) full sentences taken directly from the corpus,

(ii) grammatical sentences formed by combining chunks taken from different parts of the corpus,

(iii) sentences made of combined chunks from the corpus and chunks not in the corpus,

(iv) sentences of words present in the corpus but not forming alignable chunks and of words not in the corpus.

Each sentence was run through the MT system and the resulting output manually evaluated based on the alignments of the corpus. The results we evaluated and divided into four categories depending on their quality: *'good'*, *'fair'*, *'poor'* and *'bad'*. We now provide an explanation of the metric employed with examples using the sentence *it was almost dark*.

Good: contains all or most of the correct grammatical information (i.e. adverbs, prepositions that provide detail about the concept) and content (i.e. head noun or verb) information.

(8) `Gloss RH English`: DARK

`Gloss LH English`: DARK

`Mouth`: ‘donker’

`Brows`: f

`Eye Aperture`: s.

Fair: contains the correct content information but is missing some of the grammatical detail.

(9) `Gloss RH English`: DARK

`Gloss LH English`: DARK

`Mouth`: ‘donker’

(no brow or eye movement shown, which alters the meaning of the phrase)

Poor: contains only some correct content information and either lacks grammatical detail or contains the incorrect grammatical detail.

(10) `Gloss RH English`: DARK

`Eye Aperture`: c.

Bad: contains an entirely incorrect translation.

(11) `Gloss RH English`: WHAT

Results and Discussion

The manual evaluations performed on the test results show that the system is competent in translating sentences that occur fully intact in the corpus as would be expected from any EBMT system. These results also show that more than half the translations of sentences made up of chunks from the corpus provide reasonable-to-good translations. The system is able to segment the input and find adequate matches in the corpus to produce coherent translations for 60% of the sentences tested from (ii). This is also the case for almost a third of test sentences where data consists of combined corpus and external chunks (sentence type (iii)). The more data that is not present in the training set that is introduced in the test set, the lower the rating, as can be seen from the results of type (iii) and (iv), where an increased amount of material not present in the corpus is tested. In these cases, translations are still produced but are of poor to bad quality. For sentence type (iii), only a third of the sentences were of fair quality. For sentence type (iv), more than two thirds of the translations were considered bad and the remainder poor. These results are shown in Table 4.1.

	Good	Fair	Poor	Bad
Sentence Type (i)	40	0	0	0
Sentence Type (ii)	4	20	16	0
Sentence Type (iii)	0	12	20	8
Sentence Type (iv)	0	0	12	28

Table 4.1: Manual Evaluation Scores for NGT Data

In order to demonstrate the translation quality for each category, we provide examples in (12), (13), (14) and (15). In each case, the (a) shows the English test sentence and (b) shows the NGT gloss translation. The Dutch translation glosses have been removed for ease of reading and an explanation of the annotation codes are included in round brackets after each occurrence.

(12) a. the tortoise wins

b. `Eye Aperture`: b (‘b’ indicates a blink)

`Mouth`: ‘schildpad’

`Eye Gaze`: c (eye gaze is directed to the camera)

`Gloss RH English`: TORTOISE

`Gloss LH English`: TORTOISE

`Mouth`: ‘win’

`Gloss RH English`: WIN

`Eye Gaze`: rh (eye gaze is directed to the right hand)

`Gloss RH English`: (g-) pu (this indicates a 'palms up' gesture)

Example (12) obtained a 'good' score as it contains all the correct information.

(13) a. leaving a cloud of hot air

b. `Gloss RH`: LEARN

`Gloss RH English`: CLOUD OF DUST

`Cheeks`: p ('p' indicates a puffing of the cheeks)

`Mouth`: /CHEEKS/

`Gloss RH English`: HOT AIR

`Gloss LH English`: HOT AIR

Example (13) obtained a 'fair' evaluation mark as it contains mostly correct content information but displayed some errors.

(14) a. John saw Mike but he ran away

b. `Gloss RH`: DOG

`Gloss RH`: SAD

`Gloss RH`: MAKE

`Gloss RH English`: QUICKLY RUN AWAY

A 'poor' evaluation mark was given to (14) as it lacks both correct content information as well as grammatical details such as NMFs.

(15) a. John went to the village in the rain

b. `Gloss RH`: DOG

`Gloss RH`: TEN

`Gloss RH`: I

`Gloss RH`: DIE

`Gloss RH English`: SUN

The final example in (15) obtained a 'bad' evaluation mark as it completely mistranslated the sentence, producing no correct information.

It was noted in these initial experiments that the Marker Hypothesis proved a sufficient approach for segmenting the English data but it was necessary to employ a time-frame-based technique to segment the SL annotations. We found that employing these segmentation approaches provided us with chunks of a similar format from which adequate alignments could be constructed for use in the translation process.

Despite the small corpus, initial results for this simple prototype system are promising and indicate further development is worthwhile even on a dataset of only a few hundred sentences. As with data-driven MT systems in general, were the corpus to be larger and to contain a richer word-level dictionary, the system would probably be able to produce closer, if not exact, matches for an increased number of chunks and words, thus improving the scores.

4.3.4 Experiment Two: Assessing Evaluation

To more objectively assess the translation capabilities of our system, we formalised our testing for the next set of experiments (Morrissey and Way, 2006). 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 was entered into the system and a translation produced based on the 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. An example translation is shown in (16). For viewing clarity, the Dutch glosses have been removed but the mouth patterns, also in Dutch, are included as they are a necessary part of the translation. The English test sentence is shown in

(a) and the candidate translation in NGT annotations is shown in (b).

(16) a. the tortoise saw the sheep nibbling the grass

b. `(Gloss RH English) TORTOISE`

`(Gloss RH English) SHEEP`

`(Gloss LH English) SHEEP`

`(Mouth) 'schaap'`

`(Gloss RH English) (p-) nibbling grass`

`(Eye gaze) c`

`(Eye gaze) rh`

This example shows the multiple field complexity that can occur in SL annotations. As noted in Section 4.2, this limits the applicability of automatic evaluation metrics as field information adds noise, and fair evaluation is only possible if the gold standard is in the same format. Furthermore, such multi-tiered annotation is difficult for an untrained eye to discern the correctness of the output, making manual evaluation a challenging task.

In light of this issue, for this particular data set, we chose to reverse the translation process taking in annotations as input and producing English. Spoken language output 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 somewhat artificial in that there is much less demand for translation from SL to spoken language. However, the task is still relevant given that data-driven MT systems have the functionality to be bidirectional, so a reversal of the language direction should provide us with a reliable estimate of the competence and accuracy of the system in general. Moreover, this task forms an important role in a bidirectional Deaf-hearing communication system.

From the change in direction we were able to obtain automatic evaluation scores and had reference translations against which to measure the output. However, as

SLs by their very nature contain few closed class lexical items (cf. p.75) (meaning that 'MOUSE' in the source SL text is more likely to be aligned with 'mouse' rather than 'the mouse' from the target English text for example), the output was sparse in terms of lexical data and rich only with respect to content words. This resulted in decidedly low evaluation scores.

The system was evaluated for the language pair NGT–English using the traditional MT evaluation metrics BLEU, SER, WER and PER. For the 55 test sentences, the system obtained an SER of 96%, a PER of 78% and a WER of 119%.[8] Due to the lack of closed class words produced in the output, no 4-gram matches were found, so the system obtained an overall BLEU score of zero.

Experiment Two Extended

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.

Discussion

An example of the candidate translation capturing the central content words of the sentence may be seen in (17a) compared with its reference translation in (17b).

(17) a. mouse promised help

b. '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

[8] 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.

evaluation metrics, the 'distance' between the output in (17a) and the gold standard in (17b) will cause the quality to be scored very poorly. Furthermore, the presence of only one gold standard reference means the candidate translation can only achieve a good score if it is similar to the single reference translation, i.e. there is no room for any variance here.

In these experiments, for our purposes, we concentrate mostly on the 'GLOSS' field, but 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 as important details may be omitted and thus affect the evaluation scores.

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.

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 SL MT 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 is its evaluation. Only one '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. The poor scores shown above indicate that we struggled to use mainstream string-based MT evaluation metrics such as BLEU, WER and PER for this exercise. While the small dataset may have contributed to the unsatisfactory scores, other pertinent issues include having only one reference text, the problem of the lack of closed class lexical items in SLs and the focus on the gloss field for the task at hand. While no formal manual evaluations were carried out on the data, it is assumed that higher scores would have been obtained using such methodologies.

While the ECHO data proved useful for aiding the development of a prototype system, 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 as it is an open domain with much descriptive, non-repeated content where quite possibly any word never encountered before could arise. Such a corpus is also likely to contain colloquial terms and quotations to further hamper the MT process. Furthermore, its practical uses are limited as there is little demand for spoken language poetry or prose to be automatically translated into any SL.

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 results in data sparseness; for any data-driven approach, the larger 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 the input string, which correspondingly increases the chances of improving system output. We investigate this further in Chapter 5, where we will show that it is the domain and quality of the data and system rather than quantity that affects translation quality the most in our experiments.

4.4 Summary

In this chapter we discussed the problematic issues that arise when integrating SLs into data-driven MT systems. The primary problems concern data and evaluation issues.

In Section 4.1 we discussed the issue of data as a prerequisite for data-driven MT systems and outlined the difficulties in obtaining SL data of sufficient quantity and in an accessible format for use in this MT approach. We showed how we circumvented this problem by finding three corpora that had the potential to seed a data-driven MT system: the Signs of Ireland project, ASLLRP and ECHO project. The latter proved the most viable choice, given its availability, domain restriction, and multiple languages, despite the fewer number of sentences.

In Section 4.2 we discussed the issue of evaluating SL MT systems. We outlined how automatic metrics are suited to spoken language text output for MT systems but the potential complexity of annotations can hinder accurate automatic evaluations. It was also noted that the time-consuming and subjective annotation process leads to a lack of gold standard reference translations being available, thus decreasing the chances of the candidate translation obtaining a successful evaluation. We then addressed the comparative evaluation methodology of manual analysis. We outlined its prohibitively labour-intensive and knowledge-dependent aspects for analysing annotations and described it as a more suitable methodology for the evaluation of avatar videos.

Having discussed both these processes, we argued that in order to accurately assess the capabilities of an MT engine, the annotated output should ideally be evaluated using automatic metrics before avatar production to avoid the potential addition of errors from the avatar technology. As the visual production of the SL as the final output is necessary for the development of a fully functioning system, manual evaluation should be carried out on the avatar. In general, we contend that an evaluation metric most suited to the output at each stage should be used to properly assess an SL MT's systems capabilities in each area.

To set these issues in context, we continued by describing a prototype EBMT system developed to test data and evaluation choices. We first described the simple system and ECHO dataset, and then discussed experiments performed to test the suitability of our chosen data in terms of quantity and quality and also evaluation.

In the preliminary exploratory experiment in Section 4.3.3, we selected sentences with varying degrees of novel sentence information to test the system and perform informal manual evaluations. The initial examination of output showed that even using a simple system, the central concepts were present and data-driven MT was possible for annotated SL data with as small a dataset as the ECHO corpus.

Formalising the process for the evaluation experiments in Section 4.3.4, we took advantage of the bidirectional functionality of data-driven MT systems and reversed the language direction for the NGT–English dataset. This produced linear English sentence output, as opposed to multi-tiered annotations, which allowed us to use automatic metrics for evaluation. The evaluation scores proved lower than anticipated, yet manual inspection shows that the central concepts were largely translated correctly. We contend that a number of linguistic issues, such as a lack of closed class lexical items and NMFs affected the outcome, as did the lack of reference texts. We concluded, therefore, that the domain of children's stories and poetry was not a desirable context for either MT system development or practical usage by the Deaf.

The experiments carried out in this section demonstrate that despite difficulties in sourcing SL data suitable for data-driven MT, translation is possible with a sentence set as small as 561 sentences. Low evaluation scores demonstrate that further development is required, and we project that given a larger dataset in a more restricted domain and using a more sophisticated system, there are grounds for the development of a successful SL data-driven MT system.

Evaluation is an integral part of MT that enables developers to judge the success of the system. We are confident that further system development will improve evaluation scores for both automatic and manual metrics, and that both approaches are necessary for an accurate assessment of a full system's capabilities.

In the next chapter, we will show that an improved data set (a purpose-built ISL corpus) and a more sophisticated system (namely the MaTrEx system) does serve to improve evaluation scores. We also address the issue of evaluation on multiple output formats and show how data that is annotated more simply than the ECHO data can more easily be evaluated and achieve impressive scores.

Chapter 5

Data-Driven Sign Language Machine Translation: Experiments Using the MaTrEx MT System

In this chapter, we describe experiments in data-driven SL MT on a purpose-built ISL data set using the MaTrEx System, the data-driven MT system developed at the National Centre for Language Technology,[1] Dublin City University. We first introduce the ISL corpus and discuss the collection and annotation processes carried out. Then we introduce the MaTrEx system, explaining its component modules in the context of the data-driven system methodology described in Chapter 3. The main body of this chapter describes the set of experiments we have performed using this system, originally developed for spoken language data, for ISL MT. We demonstrate that despite the small amount of data and initial problems concerning data, as described in Chapter 4, data-driven SL MT is not only possible, but can achieve automatic evaluation scores comparative to mainstream spoken language MT.

[1] http://www.nclt.dcu.ie/mt/

5.1 Data

In the previous chapter, we concluded from our experiments that the unrestricted domain of children's fables and poetry was neither conducive to data-driven MT, given the wide vocabulary, nor to the development of practical technology for the Deaf community, given the topic. In light of this, we chose to construct our own dataset, with a practical and restricted domain considered a necessity.

Our choice of domain arose from conversations with Deaf colleagues about circumstances in which translation technology could be useful to them. The area of travel was suggested and we began searching for a suitable corpus. We found two suitable datasets: the ATIS corpus (Hemphill et al., 1990) consisting of 595 utterances and the SunDial corpus (Peckham, 1993) consisting of a further 852 sentences. The ATIS (Air Travel Information System) corpus is a dataset of transcriptions from speech containing information on flights, aircraft, cities and similar related information. The SunDial corpus consists of dialogues of flight information requests and responses.

These corpora are particularly suited to our MT needs as they are within a closed domain and have a small vocabulary. Having originated from speech, we believe that the corpora are particularly suitable for translation into SL, given that signing may be considered the equivalent of speech, both being a direct and person-to-person means of communication. Furthermore, the domain itself has a potentially practical use for Deaf people. 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 later if at all. It is also not displayed in the first and preferred language of the Deaf. For this reason, many Deaf people find themselves uninformed about changes to schedules and gates 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. This system could be extended to accommodate SLs without too much difficulty. 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 the video screens for the Deaf to view.

In the following sections, we show how we have begun to tackle this problem by discussing our data selection, collection and annotation processes.

5.1.1 Data Translation

One of our primary concerns for the translation of the data into ISL was the creation of a parallel corpus that was as authentic to natural ISL as possible. We feel it is important to be guided by the potential users of the technology we are developing (Morrissey and Way, 2007). For this reason, we engaged the assistance of the Irish Deaf Society to find two native ISL signers to translate the English data and to advise us on ISL grammar and linguistics in the area of corpus development.

Over a period of 4 days, we recorded the two signers (one male, one female) signing the 1,447 utterances. The signers worked in tandem, translating the English and discussing it. Then one would sign in front of the camera while the other acted as a monitor, suggesting any changes to ensure the final sentences were as authentic to ISL as possible and in standard ISL (cf. ISL variations in Chapter 2). Some alterations were made to the English data to facilitate signing, e.g. foreign place names were changed to locations in Ireland. Many Irish locations have specific signs given to them in ISL that can be articulated smoothly with the rest of the sentence, such as *Dublin* (made by joining the index finger and thumb and touching the chin with the rest of the fingers closed, then extending the thumb and index finger and touching the chin again with the base of the index finger). Many international place names included in the original corpus, such as *Newark*, must be finger-spelled which can be a laborious process if the location name is long and occurs repeatedly. It was also felt that this adjustment would make the corpus more relevant to the Irish Deaf Community.

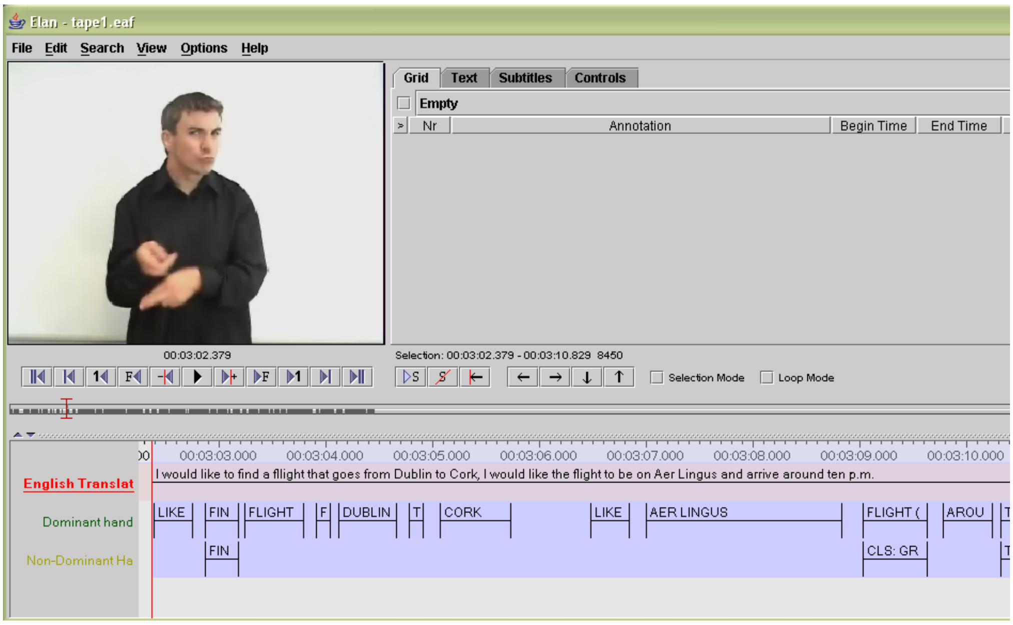
5.1.2 Data Annotation

As discussed in Chapters 2 and 4, it is necessary to annotate the SL videos in some way to facilitate SL MT. We chose to use manual annotation for the following reasons:

- we have previously worked with annotated data and found the format to be conducive to our data-driven MT experiments,
- it does not require knowledge or skills in symbolic notation formats such as HamNoSys, meaning anyone who knows the SL can annotate it,
- it allows for flexibility of the granularity of annotation, so that annotations can be as simple as glossing the signs on the hands or as complex as including phonetic features for later avatar development,
- trends in SL linguistic analysis seem to favour annotating data (cf. discussion of corpora in Section 4.1) which increases the likelihood of corpora being available for MT use in the future.

To annotate our data, we made use of the ELAN annotation software described in Section 4.1.3 as it is easy to use and freely available. Following the transcription conventions of the ECHO corpus (Nonhebel et al., 2004) and the ASLLRP corpus (Neidle, 2002), we divided our annotations into 'fields', also known as 'tiers'. We chose to keep it relatively simple in the beginning with a view to progressing to a more complex description, so initially focused on three fields:

1. Gloss for the dominant hand, (e.g. the right hand if the person was right-handed),
2. Gloss for the non-dominant hand, and
3. the original English sentence.



Selection: 00:03:02.379 - 00:03:10.829 8450

English Translat: I would like to find a flight that goes from Dublin to Cork, I would like the flight to be on Aer Lingus and arrive around ten p.m.

Dominant hand: LIKE | FIN | FLIGHT | F | DUBLIN | T | CORK | LIKE | AER LINGUS | FLIGHT (| AROU | T

Non-Dominant Ha: FIN | CLS: GR | T

Figure 5.1: Screenshot of the ELAN Annotation Interface with ISL Video and Accompanying Annotations

An example of the ELAN interface showing the video segments and three-field annotation is shown in Figure 5.1. This is a replication of the annotation interface shown in Figure 2.6. Where before it was shown to demonstrate a generic annotation interface, here we include it as an example of our own use of ELAN to annotate our ISL videos.

There are clearly limitations when using one language to describe another. Cultural and modality issues can present themselves as lexical gaps, where the describing language (English, in our case) simply does not have a direct equivalent word for the SL sign (for example, there may be one sign for *'the flight left the airport'*). Neidle suggests that, because of this, the best annotation choice would be to describe phonological features of the signs. Describing the 5 phonological features of each handshape is an extremely time-consuming process. Given that glosses provide the reader with a description of the units forming a sentence (Ó'Baoill and Matthews, 2000), and that the semantic meaning and syntactic structure remain intact, we feel glossing is sufficient for our purposes at this stage. ELAN facilitates the addition of further fields, so phonological features could be added at a later stage. Furthermore, glossing signs allows easy identification of the signs produced and provides a means to demarcate the beginning and end of a sign, a feature useful for gathering parallel information within the time boundaries of the sign.

Conventional glossing of sign languages, such as those described in Baker and Cokely (1980), Smith et al. (1988) and Ó'Baoill and Matthews (2000), use words of spoken language as glosses, to describe what is being signed by the hands. These spoken language words are usually in their root form and presented in capital letters. Where multiple words of the description language are used to represent one sign, these words are hyphenated. An example of this taken from our data is *HOW-MUCH*, where there is a single sign in ISL used to articulate this.[2]

There is generally only one gloss used to cover each sign, regardless of whether the dominant, non-dominant or both hands sign the concept. Following Nonhebel et al. (2004) and Neidle (2002), we also chose to have separate glosses for each hand in order to capture different meanings in the movements of each hand. For example, if the non-dominant hand is in a hold position pointing at an object and the right hand is signing a verb relative to that object, each hand will receive a separate and exclusive gloss.

We kept the glosses themselves quite simple using basic root forms of English words, and unlike those of the ECHO corpus, we refrained from adding extra linguistic information such as marking classifiers, or when a two-handed sign is made with one hand, for example. One exception to this was for dealing with lexical gaps resulting from how spatial deictics[3] are dealt with in visual-manual languages such as SLs (cf. Section 2.2.3). In the context of our data, place names (i.e. Dublin) would often be signed, placed in the signing space with a pointing gesture and then referred back to later in that sentence using a pointing gesture. To annotate the difference between these gestures and to encode their meaning relative to the previously signed noun, we annotated the first pointing gesture as "BE-DUBLIN", indicating that is

[2] 'How much' in ISL is signed by rubbing the thumb over the tips of the fingers on both hands with the palms facing toward the signer. It can be likened to the somewhat international hand gesture for 'money'.

[3] Spatial deictics describe word forms "whose use and interpretation depend on the location of the speaker and/or addressee within a particular setting" (O'Grady et al., 1999):297.

where it is henceforth located, and the second pointing gesture as "REF-DUBLIN", indicating a reference back to the same location in the signing space and therefore the same object.

We undertook the task of data annotation ourselves. Employing multiple persons to annotate one dataset can lead to inter-annotator disagreement and reduce the standardised annotation format. One person annotating the 595 sentences of the ATIS corpus, using the 3 fields described above, took approximately three months. At this stage it was felt that further annotation, both in terms of quantity (including the 842 sentences of the SunDial) and granularity (the inclusion of more descriptive fields, such as NMFs and phonological features) would be prohibitively time-expensive. For this reason, we chose to cease annotation at this point and begin our experiments using the complete annotated ATIS corpus.

Returning to the linguistic features of SLs, as described in Chapter 2, we can see that this form of annotation does address some of these features. As described above, deictic references are included, and classifiers are also present in the glosses. As noted above, NMFs are absent. Missing these grammatical and semantic additions could indeed affect the translations, however as we show in Chapter 6, NMFs are successfully included in the animation stage to partially compensate for this.

Comparing this new data set against the criteria for data of MT systems listed in Chapter 4, we can see that the ATIS ISL corpus has an adequate amount of data to seed a system (as we will show in our experiments in the following sections), has a more restricted domain than our previous choice, is consistently annotated using a fixed number of fields and only one annotator, and has an easily accessible format for data extraction.

The 595 sentences of the English (EN) ATIS corpus were also translated into German (DE) and then DGS gloss annotation. This provided us with four parallel corpora, already sententially aligned, with the potential to work with four translation pair types:

(i) from SL to spoken language (ISL–EN, ISL–DE, DGS–EN, DGS–DE),

(ii) spoken language to SL (EN–ISL, DE–ISL, EN–DGS, DE–DGS),

(iii) spoken language to spoken language (EN–DE, DE–EN),

(iv) and the novel translation pairings of SL to SL (DGS–ISL, ISL–DGS).

Each data set underwent a pre-processing step to extract the aligned sentences and annotations in preparation for the next phase: translation.

5.1.3 Testing the New Data

In order to test our theory that a more closed domain would facilitate improved translation, we set up an experiment using the newly prepared ATIS corpus. To get a more accurate approximation of the improvements from a change in data alone, and in order to fairly compare this experiment with the ones using the NGT data, we ran the experiments using the prototype system described in Chapter 4.

At the time of running this experiment 400 sentences had been annotated. The dataset was divided into approximate 90:10 training:testing sets with a test set comprising 44 randomly selected sentences. In order to obtain comparative results with previous experiments, we maintained the SL-to-spoken language translation direction and used BLEU and error rate measures to score the resulting output.

A direct comparison of evaluations is shown in Table 5.1. Sample output from the translations are shown in (18) and (19). (a) in each example shows the gold standard and (b) shows the candidate translation produced by the system.

	BLEU	SER	WER	PER
ECHO Data	0	96	119	78
ATIS Data	6	95	89	55

Table 5.1: Comparison of Automatic Evaluation Scores for ATIS Data and ECHO Data

(18) a. a couple of hours later he suddenly wakes up and looks around, where is the tortoise?

b. on and on time passes awake I hare look tortoise where look

(19) a. Departing Thursday mornings before nine o'clock

b. thursday in the morning to leave before 5 pm

Even on this small training and testing set, even smaller than the ECHO data set, the system obtained higher scores across all metrics. The new data scored a SER of 95%, a WER of 89%, a PER of 55% and a BLEU score of 6%. While there is little change in the SER (a mere 1% improvement), the WER and PER are significantly improved by 30% and 23% respectively. This shows a relative increase of approximately 25% for WER and 30% for PER. The presence of a BLEU score alone shows that even switching to a more suitable data set can prompt dramatic improvements to evaluation scores. These improvements are reflected in the sample output translations, where we can see that the translation for the ATIS data, compared with the gold standard (shown in (19)) is a better translation than the one shown for the ECHO corpus in (18). This also serves to highlight the importance of a suitable domain for data-driven MT.

Although the error rate scores are still quite high and the BLEU scores low, in subsequent experiments we will show that the introduction of a more sophisticated MT engine, namely the MaTrEx system described in the next section improves results further.

5.2 The MaTrEx System

Our preliminary experiments, discussed in Chapter 4, employed a basic EBMT system for performing experiments. However, to better assess the translation potential of our new data source, a more sophisticated data-driven system is required.

MaTrEx (Machine Translation using Examples) is the data-driven MT system developed at the National Centre for Language Technology, DCU (Stroppa and Way, 2006). It is a hybrid system that avails of both Example-Based MT (EBMT) and SMT approaches (Armstrong et al., 2006) by combining the resulting chunk- and phrase-alignments to increase the translation resources.

The system is modular in design consisting of a number of extensible and reimplementable modules that can be changed independently of the others. This modular design makes it particularly adaptable to new language pairs and experiments can be run immediately with new data without the need to create linguistic rules tailored to the language pair. It also facilitates the employment of different chunking methods. This system has been developed using established Design Patterns (Gamma et al., 1995). An overview of the system architecture is shown in Figure 5.2 taken from (Armstrong et al., 2006). The main modules are described in the following sections.

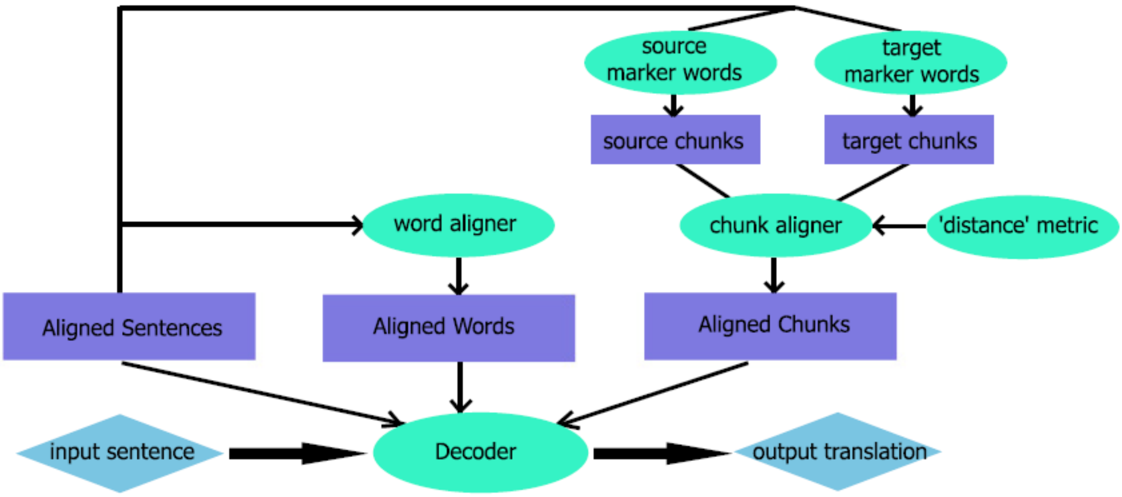


Figure 5.2: The MaTrEx Architecture

5.2.1 Word Alignment Module

The word alignment module takes the aligned bilingual corpus and segments it into individual words. Source words are then matched to the most appropriate target word to form word-level translation links. These are then stored in a database and later feed the decoder.

Word alignment for the system is performed using standard SMT methods, namely Giza++ (Och, 2003), a statistical word alignment toolkit employing the "refined" method of Koehn et al. (2003). The intersection of the uni-directional alignments sets (source-to-target and target-to-source) provides us with a set of confident, high-quality word alignments. These can further be extended by adding in the union of the alignments. Only one-to-one word alignments are produced here. Probabilities for the most likely translation alignments are estimated using relative frequencies.

5.2.2 Chunking Module

The aligned bilingual corpus is also taken by the chunking module to be segmented into sub-sentential components, so-called 'chunks'. The primary chunking strategy employed for our language pairs in this system is based on the Marker Hypothesis (Green, 1979) (cf. Section 4.3 for description). Marker-based chunking has the advantage of being easily adaptable to new languages by simply providing the system with a relevant list of marker words. This simplicity keeps training and linear complexity to a minimum. The module works on both the source and target aligned sentences, producing source and target chunks that are fed into the next module.

5.2.3 Chunk Alignment Module

The chunk alignment module works on a sentence-by-sentence basis, taking the chunks formed in the previous module one sentence at a time and forming translation links between source and target language chunks.

An 'edit-distance style' dynamic programming alignment algorithm is employed to align these chunks. The source and target chunks with the least 'distance' between them in terms of insertions, deletions and substitutions form translation links. The distance metrics used are:

1. Distance based on Marker tags,
2. Chunk minimum edit-distance (word-based distance and character-based distance),
3. Cognate information (i.e. words that have similar roots in both languages),
4. Word translation probabilities,
5. Combinations of the above.

The algorithm can also be adapted to allow for 'jumps' or block movements to take into account the possible differences in word order of the source and target languages (Morrissey et al., 2007b). For example, should we be translating a language that is verb-final into one that is verb-initial, the system can be configurated to allow the search for chunk alignments to extend to the the length of the full sentence to ensure a good match.

The resulting aligned chunks are then combined with the SMT phrasal alignments. The two alignment styles are merged to help produce translations of a higher quality than the respective baseline systems following the research of (Groves and Way, 2005b,a).

5.2.4 Decoder

Source language sentences are translated into target language sentences via the decoder. The decoder chooses the best possible translation by comparing the input against the source side of the bilingual databases of aligned sentences, EBMT chunks, SMT-phrases and words that feed it. Translation links are retrieved for these matches and they are recombined to produce a candidate target language translation string. The decoder in MaTrEx is a wrapper around Moses (Koehn et al., 2007), a phrase-based SMT decoder.

5.2.5 Non-SL Experiments using MaTrEx

The MaTrEx system is primarily used to translate between texts of spoken languages and has successfully participated in translation competitions such as the International Workshop on Spoken Language Translation (IWSLT) in 2006 and 2007. In order to demonstrate the translation capabilities of this system, we show in Table 5.2 the BLEU, WER and PER scores for experiments involving Arabic-to-English, Chinese-to-English and Japanese-to-English (Hassan et al., 2007) as well as Italian-to-English (Stroppa and Way, 2006). In order to assess the translation potential of a small corpus using the MaTrEx system, we ran an experiment using the German and English spoken language ATIS data. Despite having only a fraction of the training data of the other experiments shown in Table 5.2, the German–English ATIS language pair produced the best results, and show over 13% improvement on BLEU score alone over the next best, Arabic–English.

Language Pair	**BLEU**	**WER**	**PER**
Arabic–English	47.09	n/a	n/a
Chinese–English	27.37	n/a	n/a
Japanese–English	39.59	n/a	n/a
Italian–English	34.67	49.64	37.44
German–English	60.73	26.59	22.16

Table 5.2: Evaluation Scores for MaTrEx Spoken Language Experiments

We can see from looking at the BLEU scores that accuracy ranges from between 60.73% for the German data and 27.37% for the Chinese data. Stroppa and Way (2006) state that these scores are competitive with other state-of-the-art systems. Furthermore, output from Hassan et al. (2007) was ranked first according to human evaluations.

5.3 Experiments

In this section we describe the experiments carried out using the data-driven MT engine described in Section 5.2 seeded by the ATIS data set described in Section

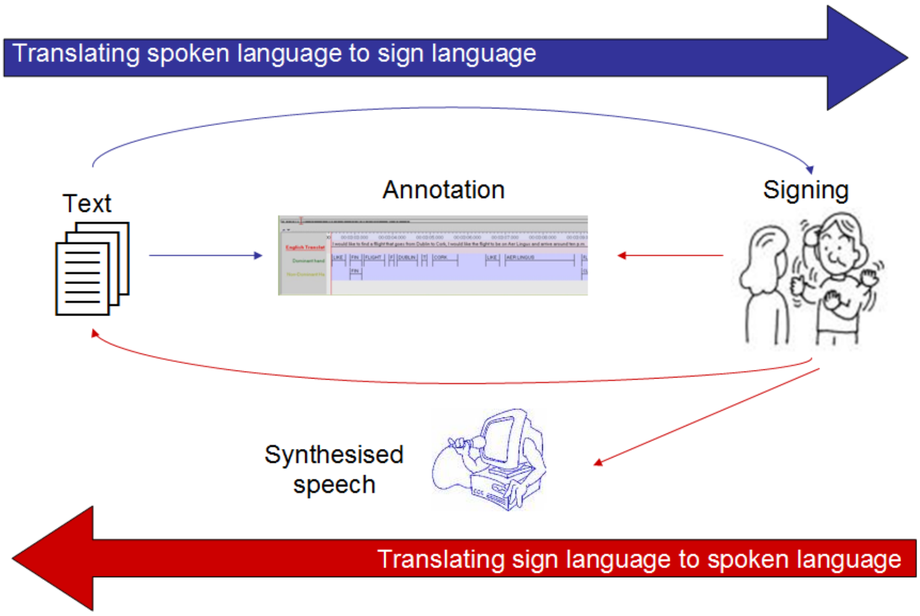

Translating spoken language to sign language
Text
Annotation
Signing
Synthesised speech
Translating sign language to spoken language

Figure 5.3: MaTrEx Translation Directions

5.1. The bidirectional ability of our MT system allows us to translate to and from both languages with ease. Five sets of experiments will be described, exploring these different translation directions, namely:

- gloss-to-text,
- SL-to-text,
- gloss-to-speech,
- text-to-gloss,
- text-to-SL.

A schema of these translation paths is shown in Figure 5.3. The narrow lines indicate experiments performed in those directions.

5.3.1 Translating Sign Language Gloss to Spoken Language Text

In order to compare experiments using MaTrEx with those described in the last chapter and the change of data experiment in Section 5.1, we maintain translation

		EN	ISL
Train	no. sentences	418	
	no. running words	3008	3028
	vocab. size	292	265
	no. singletons	97	71
Dev	no. sentences	59	
	no. running words	429	431
	vocab. size	134	131
Test	no. sentences	118	
	no. running words	999	874
	vocab. size	174	148

Table 5.3: Overview of ATIS Corpus Divisions: training, testing and development

in the direction of spoken language text.

In this experiment, we translate from the annotated gloss version of the ISL data into English text. We divided the 595 sentences of the ATIS corpus into training, development and testing sets of 418 sentences, 59 sentences and 118 sentences respectively. An overview of the corpus breakdown is given in Table 5.3.

Within this experiment we carry out four sub-experiments:

1. Baseline,
2. Introducing EBMT chunks: Type 1,
3. Introducing EBMT chunks: Type 2,
4. Changing the Distortion Limit.

Baseline

In order to have a baseline against which to compare further experiments, we used the most basic functions of the MaTrEx system, namely the modules described in Section 5.2 with the exception of EBMT chunks. The candidate English translations produced by the system were automatically evaluated using BLEU, WER and PER.

Even using the most basic version of the system on first run produced dramatic improvements in automatic evaluation scores for the dataset compared with the old system. The candidates, when compared against the 'gold standard' withheld for this purpose, obtained a BLEU score of 51.63% and a WER and PER of 39.32% and 29.79% respectively. This signifies that over 70% of the translated words are correct and over 60% are in the correct order based on the gold standard. Table 5.4 compares these scores against those obtained by the old system on the same data set as well as the old system with the NGT data set.

	System and Data	**BLEU**	**WER**	**PER**
Gloss to Text	MaTrEx Baseline with ISL	51.63	39.32	29.79
	Prototype with ISL	6	89	55
	Prototype with NGT	0	119	78

Table 5.4: Comparison of Evaluation Scores: MaTrEx baseline with ISL data, prototype with ISL data and prototype with NGT data

Comparative translation samples of the ATIS corpus taken from the output of the MaTrEx system and the prototype system are shown in (20). (a) shows the gold standard reference translation, (b) shows the prototype system's candidate translation and (c) shows the MaTrEx system's candidate translation.

(20) a. What flights from Kerry to Cork on Saturday?

b. to london on saturday whats do to kerry to go cork

c. on saturday what flights from kerry to cork

From these translation samples, we can see a clear improvement in the translation produced by the MaTrEx in comparison with that of the prototype system. The scores, coupled with the sample translations show that, while changing to a more suitable data set already improved scores, employing the MaTrEx system significantly enhanced translation even further. As stated in Chapter 4, changing the dataset incurred increases of 6%, 30% and 23% for BLEU, WER and PER respectively. Here we can see the combined increases incurred by choosing a more suitable corpus and a more sophisticated MT engine: BLEU scores are improved by 51.63%, WER by 70.68% and PER by 48.21%. Furthermore, a manual examination of the sentences shows the MaTrEx translations are more intelligible and comparable to the reference translation.

Introducing EBMT chunks: Type 1

Having developed a baseline for the MaTrEx system using ISL data, we began experimenting with the addition of EBMT chunks. These EBMT chunks can potentially improve translation results by bolstering the SMT-phrasal alignments with additional sub-sentential alignment examples.

For the first experiment, we employed the Marker Hypothesis to segment the source and target aligned sentences. Despite the reduced number of closed-class lexical items in SLs, compared with spoken languages, there are some still present. We chose to take this into account for our first experiments. We collated closed class lexical items in English to create our marker lists. As the ISL data was annotated in English, we were able to use the same list for segmentation of both source and target languages. The resulting segments created by the Chunking Module were fed into the Chunk Alignment module to form translation links. These were then added to the SMT phrasal alignments.

Using the same divisions of the data set as were used in the baseline experiments, the MaTrEx system translated the ISL glosses into English and automatically evaluated them. The candidate translations obtained a BLEU score of 50.69%, a WER of 37.75% and a PER of 30.76%. This shows an improvement in scores for WER (by 1.57%) but not for BLEU or PER (poorer results by 0.94% and 0.97% respectively) when compared with the baseline. A comparision of these results is shown in Table 5.5.[4]

It is likely that this deterioration in evaluation scores is a result of the lack of closed class lexical items in SLs. The Marker Hypothesis is based on the presence of these closed class lexical items, using them as markers for segmentation. With these lexical items, such as determiners, being largely absent from the ISL, it is less

[4] It should be noted at this point that statistical significance testing has not been performed on the experiments described in this thesis.

	System	BLEU	WER	PER
Gloss to Text	MaTrEx Baseline	51.63	39.32	29.79
	Baseline + Type 1 Chunks	50.69	37.75	30.76

Table 5.5: Comparison of Evaluation Scores: MaTrEx baseline alone compared with the addition of Type 1 chunks

likely that the contents of ISL chunks will correspond to that of the English chunks. This is reflected in the poorer BLEU and PER scores. The seemingly inconsistent increase in the WER score shows that there is an increase in the number of correct words in the correct order. This signifies that the addition of EBMT-style chunks improves the likelihood of producing at least parts of sentences in correct word order. In order to experiment further with EBMT chunks, we next investigate an alternative chunking methodology.

Introducing EBMT chunks: Type 2

Given the natural lack of closed class lexical items in SLs, as discussed in the previous section, we propose a different chunking methodology. During our examination of the source and target language texts, it was noted that frequently one gloss annotation taken from the ISL data corresponded to a whole marker-based chunk from the English data. An example of this is shown below. (21) shows the English sentence and the corresponding ISL gloss. (22) shows the English marker chunks on the first line and ISL word boundary chunks on the second line and (23) shows the correspondences between these chunks indicating potential alignments. Angled brackets (< >) indicate the beginning of a chunk in each case. The tag within the angle brackets has been removed for ease of reading.

(21) a. I'd like a flight

b. LIKE FLIGHT

(22) a. <> I'd like <> a flight

b. <> LIKE <> FLIGHT

(23) a. <> I'd like = <> LIKE

b. <> a flight = <> FLIGHT

In order to create a new set of chunks for the source and target, we used the same Marker-based methodology for the English data and segmented the ISL data using spaces between glosses as delimiters so that each individual gloss became a chunk, as shown in (22b). The resulting evaluation scores, compared with both the baseline and Type 1 chunking methodology results are shown in Table 5.6.

	System	**BLEU**	**WER**	**PER**
ISL-EN	MaTrEx Baseline	51.63	39.32	29.79
	Baseline + Type 1 Chunks	50.69	37.75	30.76
	Baseline + Type 2 Chunks	49.76	39.92	32.44

Table 5.6: Comparison of Evaluation Scores: MaTrEx baseline alone compared with the addition of Type 1 chunks and addition of Type 2 chunks

As shown on the comparison table, this chunking methodology produces worse scores across the board, both in comparison with the Type 1 chunks and with the baseline, with differences of between 0.6% and 2.65%. This indicates our chunking methodology did not have the expected effect. This may be because there are still some closed class lexical items in the ISL data such as the preposition 'after'. Conflicts in chunk alignment could arise when, for example, the ISL sign 'after' is aligned with the chunk 'after five o'clock' in the English chunk set and the ISL chunk 'five o'clock' is also aligned with 'after five o'clock'. A large number of partially incorrect alignments such as this increases the amount of repeated and incorrect information in the English chunks chosen to produce the candidate translations.

Comparative sample candidate translations produced by each experiment are shown in (24). (a) shows the reference translation, (b) shows the baseline translation, (c) shows the Type 1 chunks translation and (d) shows the Type 2 chunks translation. To better illustrate the different translations, the example shown here is different from that used previously in (20).

(24) a. Which are the morning flights?

b. which is the morning flights

c. which of these flights morning

d. which flights which the morning

The above candidate translations, while all similar and capturing the gist of the translation, we can see the differences that reflect the evaluation scores. Although this is only a sample set of translations, it illustrates how even slight deviations from the reference translation can affect evaluation scores.[5]

Assessing the Benefits of Allowing Jumps by Changing the Distortion Limit

The distortion limit function in the system (Morrissey et al., 2007b) allows for jumps or block movements to occur in translation to account for the differences in word order of the languages being translated. The limit is set to 0 jumps as default. Given the differences between SLs and spoken language grammar in terms of word order, particularly the sentence-initial positioning of time references and similar grammatical structures in SLs, we experimented using varying limits.

We found that allowing a distance range of 10 jumps for block movements when decoding produced the most significant increase in scores. Against the baseline, BLEU score improved by 0.55%, WER by 0.84% and PER by 0.12%. When used with Type 1 chunks, the BLEU score decreased by 0.62%. The error rate scores improved slightly by 0.36% for WER and 0.13% for PER. For Type 2 chunks, the BLEU score also decreased, this time by 0.56%, and again the error rates improved slightly by 0.36% and 0.12% for WER and PER. A possible reason for the discrepancy in these scores is the methodology behind the evaluation metrics. The change in the distortion limit to 10 may increase the number of correct words found, thus increasing the improvement in error rates, but this may decrease the number of cor-

[5] It must be noted that punctuation errors are factored into scores.

rect n-grams in the candidate compared to the gold standard. One way of improving this would be to have multiple gold standard reference texts. Table 5.7 shows the comparative scores of the above experiments including the difference that altering the distortion limit made to each.

		BLEU	WER	PER
ISL–EN	baseline	51.63	39.32	29.79
	Dist. Limit = 10	*52.18*	*38.48*	*29.67*
	Type 1 Chunks	50.69	37.75	30.76
	Dist. Limit = 10	*51.31*	*37.39*	*30.63*
	Type 2 Chunks	49.76	39.92	32.44
	Dist. Limit = 10	*50.32*	*39.56*	*32.32*

Table 5.7: MaTrEx Evaluation Results for ISL–EN Gloss-To-Text Experiments

Comparing System Performance on SL Data with Spoken Language Data

Having performed our first set of experiments on the ISL data, we were satisfied that our system was doing well and that the ATIS annotated data format was a satisfactory choice for our experiments. In order to assess this and put our SL experiments in the context of more mainstream spoken language MT, we compared the best ISL–EN scores against the scores obtained by the MaTrEx system for our ATIS German–English spoken language data experiment described on page 100. Table 5.8 shows the comparative scores with the SL experiments shown in **bold face**.

Language Pair	**BLEU**	**WER**	**PER**
German–English	60.73	26.59	22.16
ISL–English	**52.18**	**38.48**	**29.79**

Table 5.8: Evaluation Scores for MaTrEx Spoken Language Experiments compared with SL experiments

Comparing these scores, we can see that our ATIS SL experiments obtain BLEU evaluation scores less than 10% lower than the spoken language experiment. Note that, from our previous comparision of the German–English pairing with larger spoken language corpora in other languages attains the highest BLEU score of all language pairs. This is interesting, as the German–English pairing experiments use the 595 sentences of the ATIS data as opposed to the millions of sentence pairs used in the other IWSLT experiments. The German–English and ISL–English ATIS data experiments with the MaTrEx system show that, not only is data-driven MT possible for small data sets, it is possible to achieve evaluation scores comparable with mainstream spoken language experiments of much larger data resources. Furthermore, it demonstrates that the annotated SL data is suited to data-driven MT.

Comparative Experiments Using DGS and German Parallel Data

Our main focus has been the translation of ISL data, enabling us to develop a system that is useful to our national Deaf community. However, having had the ISL data translated into both German and DGS, we ran comparative experiments in order to more broadly assess the translation capabilities of the MT engine for SLs.

The language pairs used in these experiments are as follows:

- ISL–DE,
- DGS–DE,
- DGS–EN.

Each experiment for each language pairing was run exactly the same as described in the previous section: baseline, Type 1 chunking methodology and Type 2 chunking methodology. The results are shown in Table 5.9.

Comparing the baseline scores of all language pairings, we can see that the ISL–EN pairing produces the best scores despite the system not being trained specifically for this language pair. There is an improvement in BLEU score of 6.38% compared with the next best: DGS–EN. At baseline level all systems score within a 16% range of each other, which shows that the MaTrEx system is capable of achieving

		BLEU	WER	PER
ISL–DE	baseline	38.18	48.52	38.79
	DL = 10	*39.69*	*47.25*	*38.47*
	T1 chunks	40.67	46.72	38.58
	DL = 10	*42.13*	*45.45*	*38.16*
	T2 chunks	38.54	46.93	38.05
	DL = 10	*40.09*	*45.66*	*37.63*
DGS–EN	baseline	45.25	48.85	32.08
	DL = 10	*48.40*	*41.37*	*30.88*
	T1 chunks	44.74	50.66	31.72
	DL = 10	*47.22*	*44.14*	*31.12*
	T2 chunks	44.34	49.93	33.17
	DL = 10	*47.43*	*42.82*	*32.20*
DGS–DE	baseline	38.66	55.28	39.53
	DL = 10	*42.09*	*50.31*	*39.53*
	T1 chunks	34.86	56.65	39.53
	DL = 10	*39.38*	*51.37*	*38.79*
	T2 chunks	35.63	55.81	39.74
	DL = 10	*40.29*	*50.31*	*38.90*
ISL–EN	baseline	51.63	39.32	29.79
	DL = 10	*52.18*	*38.48*	*29.67*
	T1 chunks	50.69	37.75	30.76
	DL = 10	*51.31*	*37.39*	*30.63*
	T2 chunks	49.76	39.92	32.44
	DL = 10	*50.32*	*39.56*	*32.32*

Table 5.9: MaTrEx Comparative Evaluation Results for All Gloss-To-Text Experiments

satisfactorily comparable evaluation scores to mainstream spoken language experiments, as documented in Section 5.2, regardless of the language pairing. Results in general are more favourable for pairings involving ISL. This may be attributed to the closer links between source and target representations for alignment; intuitively, alignments are more likely to occur between English words and English glosses.

The Type 1 chunking methodology improved evaluation scores for the ISL–DE pairing only, with a significant BLEU score increase of 2.49% and WER improvement of 1.8%. This chunking method was not as successful for the DGS pairings where scores decreased by up to 3.8% BLEU score for DGS–DE.

The Type 2 chunking methodology, using the Marker Hypothesis for the German and English and the work-by-word chunking for the SL only serves to improve scores for the ISL–DE pairing. Decreases in scores for the DGS–EN are not much more than 1%, but more significant differences are visible for the DGS–DE pairing, particularly in the BLEU score which shows a decrease of 3.03%.

While the discrepancies in chunking methodology scores show that further investigation is required to tune the chunking methodologies to the language pair at hand, we can see that there is the potential for the addition of EBMT chunks to enhance translations and their resulting scores from the ISL–DE pairing in Type 1 experiments.

Increasing the distortion limit to allow jumps of 10 places improves scores across the board with an average improvement for the three new language pairs of 2.66%. The alteration to the limit showed the most significant average improvement for the DGS–EN pairing (3.6%). This improvement is most likely a result of the distortion limit allowing for differences in word order between DGS and EN, the language pairing with the most significant difference.

System Comparison: MaTrEx vs. RWTH Aachen

Having tested that the system could perform at a level comparative to mainstream systems for multiple language pairs, we decided to compare it with other systems to more broadly assess its SL MT capabilities. Previous collaboration with the RWTH Aachen University for the creation of parallel texts in German and DGS was extended to assess the general translatability of our ATIS data set. Using the the same data sets, we ran parallel experiments on four gloss-to-text language pairs (Morrissey et al., 2007b),[6] namely:

- ISL–EN,
- ISL–DE,
- DGS–EN,

[6] These experiments were performed by the RWTH MT group.

- DGS–DE.

The MaTrEx experiments using these language pairs have been described previously in Section 5.3.1. An outline of the RWTH system and its experiments on reordering constraints is given in Section 3.3.2. Comparative evaluation scores for these experiments are shown in Table 5.10. The MaTrEx scores are shown in **bold face**.

		BLEU	WER	PER
ISL–EN	RWTH baseline	50.72	39.44	30.27
	inv-IBM reord.	52.62	37.63	28.34
	MaTrEx	**52.18**	**38.48**	**29.67**
ISL–DE	RWTH baseline	40.36	47.25	38.90
	inv-IBM reord.	40.40	46.40	38.58
	MaTrEx	**42.13**	**45.45**	**38.16**
DGS–EN	RWTH baseline	40.10	51.62	36.55
	inv-IBM reord.	43.16	46.32	31.36
	MaTrEx	**48.40**	**41.39**	**30.88**
DGS–DE	RWTH baseline	32.92	55.07	40.69
	inv-IBM reord.	35.69	49.15	38.68
	MaTrEx	**42.09**	**50.31**	**39.53**

Table 5.10: Comparison of RWTH Evaluation Results with best of MaTrEx

Comparing the results obtained from RWTH experiments on the ATIS data with the best scores obtained for each language pair by the MaTrEx system, we can see that similar to our own distortion limit experiments, the reordering techniques of the German system have also served to improve scores across the board. The German system also obtained the best scores of their own experiments for the ISL–EN language pair. There are minimal score differences (less than 1% average) between RWTH and MaTrEx results for this pairing but more significant differences are shown for the other three language pairings where the MaTrEx system shows markedly betters scores than the RWTH system. In the case of the DGS–DE language pairing the MaTrEx system has a 6.4% better BLEU score than the RWTH system. These results are also displayed on a graph in Figure 5.4.

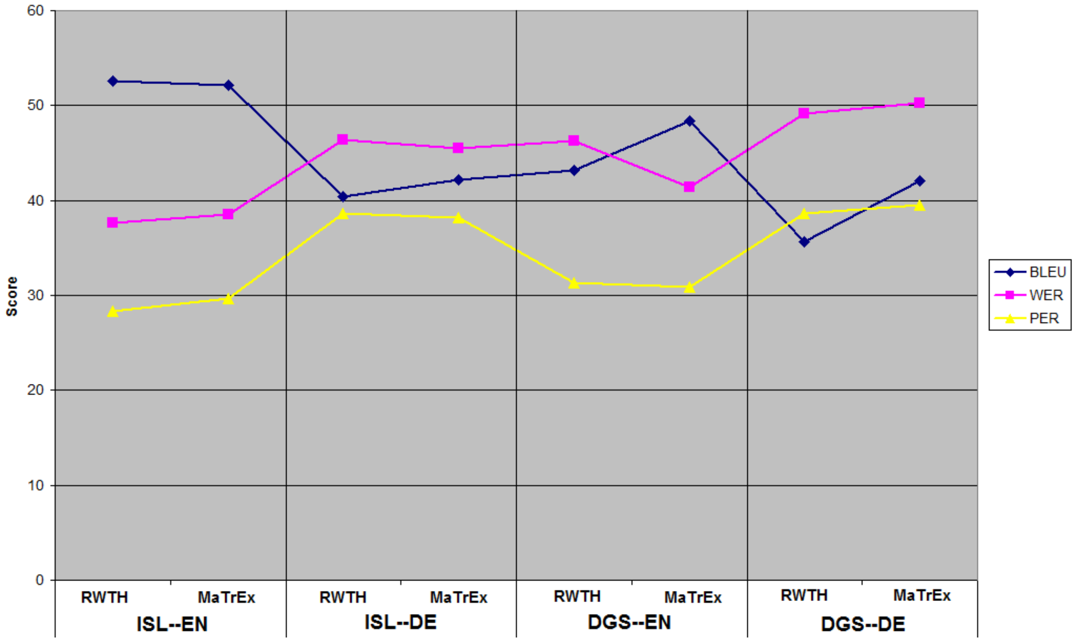


Figure 5.4: Graph Comparison of best RWTH results with MaTrEx results

While the MaTrEx system proved to be the better system for the majority of experiments, all scores for both systems were roughly within the same range. Currently the MT systems described here are relatively similar in design, considering their basic SMT make-up. In order to assess the exact components of each system that contribute to improved translations (with a view to developing a hybrid MT system), further experimentation is required. However, what this does show is that, regardless of which data-driven system is employed on this ATIS data, it is possible to achieve similar and satisfactory evaluation scores. Furthermore, our MaTrEx system is capable of achieving translations parallel with, and in most cases, better than, the RWTH system.

5.3.2 Translating Sign Language Video to Spoken Language Text

Having achieved satisfactory evaluation scores for the data through the various experiments described in the previous section, we next sought to expand the system to make it more practical with the addition of SL recognition technology. SL recognition and SL MT systems already exist but little work has been done combining the two (cf. (Bauer et al., 1999) in Section 3.3.1). Individually, these systems employ an intermediate notation system that is not directly intelligible for untrained users and is, therefore, of relatively little use to the intended user group. An SL system that only produces notation and an MT system that only accepts notation as input are not usable systems by themselves, but when combined, they have the potential to contribute to the development of a fully automated SL-to-spoken language text system. Such a system would greatly contribute toward the development of a full bidirectional SL–spoken language communication system, facilitating both Deaf and hearing users. In the following sections we introduce SL recognition technology and discuss its use in SL MT.

Sign Language Recognition

As the primary focus of our work is the translation component, we cooperated with the Sign Language Recognition group in RWTH Aachen University, Germany, to extract data in a glossed format from the same ISL videos we manually annotated in Section 5.1. Their automatic sign language recognition (ASLR) system is based on an automatic speech recognition system (Dreuw et al., 2007) with a vision-based framework. It is signer-independent and does not require any special equipment other than a standard video camera for data capturing so our videos could be used quite straightforwardly.

The system uses annotation glosses as whole-word models, where each word model consists of one-to-three 'pseudophonemes' for each sign that models the average word length. Bayes' decision rule (cf. Equation 3.2 in Chapter 3) is employed by the system to choose the best word sequence for the input observation based on the pre-trained word-model inventory and the language model. This word sequence is considered the result of the recognition process, which then feeds the translation system. Based on previous experiments using the RWTH-Boston-104 corpus[7], the

[7] http://www-i6.informatik.rwth-aachen.de/~dreuw/database.html

system boasts a WER of 17% (Dreuw et al., 2007).

Our ATIS video data was taken for recognition experiments using the RWTH system (Stein et al., 2007). The experiments focussed on the dominant hand glosses and consisted of a basic feature setup to begin with. Despite the system's success with the Boston data, a similar result was not obtained using our ATIS videos. The preliminary recognition of the videos had an error rate of 85% (consisting of 327 deletions, 5 insertions and 175 substitutions out of 593 words). This extreme difference in score is attributed to the ATIS corpus being a more sophisticated data set than the RWTH-Boston-104 corpus with an increased number of words occurring only once in the data set. Furthermore, the experiments using the RWTH-Boston-104 corpus underwent an increased number of tuning operations on the development set and had an increased number of features used for recognition. It is expected that significant improvements in results could be attained with a corresponding amount of adaptation time spent on the ATIS data.

Reviewing the attempts made by Bauer et al. (1999) in their work on recognising DGS video data, it may be assumed that there was more time spent tuning the recognition system to the data given their data set of up to 100 signs in comparison to the 593 of the ISL data. The resulting outcome of joining their recognition results with their translation tool is not discussed, merely estimations of a word accuracy of 90% are made. Given the significantly differing data sizes and time spent tuning each system to the data are very different, it is not possible to compare these systems. However, based on the description of the recognition performance combined with the performance of the Boston corpus mentioned above, it can be assumed that recognition is quite possible indeed for SL video once enough time and tuning is performed.

Given the initial poor score for the ISL data and the fact that combining any two systems will introduce additional error sources, it is apparent that to use such inaccurate data to seed an MT system would be unrealistic at this time. Further recognition experimentation on the ATIS data was ceased at this point and will hopefully resume at a later stage in order to complete a more successful SL-to-text system.

5.3.3 Translating Sign Language to Synthetic Speech

Having already explored many of the translation possibilities for SL and spoken language, we decided to further exploit the possible uses for our MT engine by adding on a speech synthesis module (Morrissey et al., 2007a). In the context of a fully-functioning, bidirectional SL MT system, speech, as opposed to text, is a more natural and appropriate output for the spoken language. This is because speech is more akin to signing than text as they are both direct means of communication, produced face-to-face.

In order to explore this avenue, we collaborated with the Multilingual Ubiquitous Speech Technology: Enhanced and Rethought (MUSTER) research group in University College, Dublin. The MUSTER group has developed the Jess system for synthesising speech in various languages (Cahill and Carson-Berndsen, 2006). It is a modular system that allows for different synthesiser algorithms to be plugged in and tested using the same source and target data. Voice data is stored in four formats: utterance, word, syllable and phoneme. Text can be input in orthographical form and the system estimates a phonetic transcription for it then calculates the best pronunciation output.

For our experiments, we provided the MUSTER group with some of the English text output from the experiments described in Section 5.3.1. This was fed into the Jess system to produce English speech. Taking only minutes to complete the entire process, this was an easy task and seeds further collaboration between our groups, which will take place as part of the CSET project on Next Generation Localisation (cf. introduction to Chapter 3).

5.3.4 Translating Spoken Language Text to Sign Language Gloss

Given the satisfactory results of translating SLs into English and German text, we decided to exploit the bidirectional functionality of our data-driven system and reverse the translation process. Translating spoken language into SLs has the potential to be directly useful to the Deaf community, enabling them to access information, particularly for private matters such as legal or medical information, without the need for an interpreter.

Translating from English to ISL, our initial experiments in this direction took English text as input and produced ISL glosses. Just as we did for the ISL–EN experiments, we ran three sub-experiments: baseline, Type 1 chunking methodology and Type 2 chunking methodology. Given the linear format of the output annotations, we were able to apply the same automatic evaluation metrics: BLEU, WER, and PER. The resulting scores for this experiment are shown in Table 5.11.

	System	**BLEU**	**WER**	**PER**
EN–ISL	MaTrEx Baseline	38.85	46.02	34.33
	Baseline + Type 1 Chunks	39.11	45.90	34.20
	Baseline + Type 2 Chunks	39.05	46.02	34.21
	ISL–EN best scores	*52.18*	*38.48*	*29.67*

Table 5.11: Evaluation Scores for EN–ISL experiments: comparison with best ISL–EN scores shown

Examining the evaluation results for these text-to-gloss experiments, we can see that, contrary to the ISL–EN experiments, the chunking methodologies have served to improve the evaluation scores but only by a small amount. Neither of the chunking experiments have increased the baseline score by more than a fifth of a percentage point. This is further testament to the fact that further investigation is required for EBMT chunking methodologies.

We also note from these results that the overall scores are not as good as the reversed language direction, with BLEU scores alone showing a difference of between 13.07% and 14.13%. The MT framework and translation methodologies are unchanged for these experiments, and while there may be some variation in translation alignments when translating in the direction of SLs, it is more likely that the evaluation metrics do not adequately capture the intelligibility of the output translations but more assess their fidelity to the single gold standard. It is likely that a trained human judge could better evaluate the output for intelligibility. Furthermore, producing ISL glosses as a 'translation' does not facilitate the wider Deaf community as it is still not in their native language. For these reasons, we feel it is necessary to further develop our research and produce animated SL from the translated glosses.

5.3.5 Translating Spoken Language Text to Animated Sign Language

Translating into SLs has a significant practical use for the Deaf community, but a system that produces gloss output is not of much use to a Deaf person and is more likely to be confusing by providing spoken language stem words in an SL syntax. For this reason, the next natural step is to produce real SL output in the form of an animated computer figure or avatar.

As the subject of avatar production and evaluation is somewhat outside the scope of this chapter, these experiments will be discussed in detail in Chapter 6.

5.4 Summary

In this chapter we have shown, through numerous sets of experiments, that data-driven MT for SLs using annotated video data can achieve automatic evaluation scores comparable to mainstream spoken language MT.

In Section 5.1 we discussed our collection and annotation methodologies for the ATIS ISL corpus. We highlighted the fact that, despite being the most suitable approach, annotation can be a subjective and time-consuming process. We showed that we were correct in our assumption that an improved data set in a small domain would help improve translations by comparing experiments run on our prototype system using both the NGT and ISL data. A simple change of data set showed improvements in scores of 30% for WER, illustrating that improved translations can be facilitated by an appropriate choice of data domain.

In Section 5.2 we addressed the conclusion drawn at the end of Chapter 4, namely that a more sophisticated MT system would serve to further improve translations and evaluation scores. We suggest that the MaTrEx data-driven MT system is capable of this and describe its architecture and component modules. In order to demonstrate the capabilities of the MaTrEx system, we highlight spoken language experiments from IWSLT competitions in which the system has competed. For sets of experiments translating from Arabic, Japanese, Italian, Chinese and German into English, we compare automatic evaluation scores pointing out the authors' assertion that these scores are comparable with other state-of-the-art MT systems and have been successful in the aforementioned competitions.

We continue addressing our hypothesis that the MaTrEx system can improve on the ISL experiment of the prototype system in Section 5.3. The gloss-to-text experiments discussed in this section showed that employing a sophisticated data-driven MT engine such as the MaTrEx system can dramatically improve the translations produced and resulting automatic evaluation scores, even at the baseline level. While, in theory, the addition of further sub-sentential information to the system, in the form of EBMT chunks, should improve the translations, they have in fact deteriorated. While initial experiments using two different chunking methodologies have proved unsuccessful, we are confident that further parallel examination of the language pair and their respective linguistic constituents would lead to the development of a suitable chunking methodology. Despite the disappointing chunk experiment results, our experiment of altering the distortion limit to cater for differing word order was successful and increased scores across the board.

Furthermore, we show that the evaluation scores for the baseline experiments alone are sufficient to warrant satisfactory evaluations that are comparable in scoring with the spoken language experiments outlined in Section 5.2. This level of success indicates two important things: annotated data is a suitable representation of ISL for data-driven MT, and that satisfactory data-driven MT is not only possible for SLs using the MaTrEx system but it can be achieved with data sets of only a few hundred sentences.

Further experiments in gloss-to-text translation were carried out to demonstrate that data-driven MT is possible for other SL language pairs. Through experiments on German and DGS using parallel ATIS corpora, we showed that, while ISL–EN translation still produced the best scores, DGS–EN, DGS–DE and ISL–DE pairings produced scores comparable with the spoken language experiments. The Type 1 chunking methodology was shown to have improved scores for the ISL–DE pair only, but by a significant amount of 2.49%. Distortion limit changes also served to increase scores across the board illustrating the need for such a provision in the system to account for differences in word order between the langauges.

We also compared our ATIS experiments with a competitive SL MT system, that of RWTH Aachen University in parallel experiments. The MaTrEx system proved to be the better system for the majority of experiments, although both scores for both systems were roughly within the same range. This compounds our hypothesis that data-driven MT is indeed possible with small amounts of SL data.

While various other measures could be taken to further improve the translation results, we sought at this stage to extend the system to make it more practical and useful for Deaf–Hearing communication by introducing SL recognition technology.

In Section 5.3.2, we looked to expand the system to include SL recognition technology. Previous experiments by Dreuw et al. (2007) and Bauer et al. (1999) on ASL and DGS data respectively indicate that SL recognition is a viable option for feeding data to an MT system. Experiments with the ISL data lacked the necessary language training periods, and initial poor results of 89% WER prevented us taking this collaboration any further.

Another practical extension to the MT process was discussed in Section 5.3.3, where we outlined collaboration with the MUSTER Speech Synthesis group who produced audio speech output from our ISL–EN translations. This effective collaboration further enhances the usability of a bilingual MT system for communication between Deaf and hearing people.

The predominant functionality of our MT experiments lies in the development of a system that can assist communication and understanding for the Deaf community. In Section 5.3.4 we discuss how we reversed the direction of previous experiments, translating this time from English text into ISL glosses. We exploited the bidirectional capabilities of MaTrEx and produced glossed output for English text. Although gloss output is not a suitable output for the Deaf, it enables us to perform evaluations to assess the system's capabilities in comparison with our previous experiments. While the scores for this directionality were not as impressive as the ISL–EN experiments, they were still within the boundaries of satisfactory translation scores when compared with non-SL MT. Furthermore, the chunking methodologies improved scores for this direction, illustrating that further language- and direction-specific research would be of benefit to the system.

Our final experiment, addressed in Section 5.3.5, outlines our final extension to our MT system: producing animated ISL to make the translations intelligible to and useful for Deaf people. This will be discussed in depth in the next chapter.

Finally, this chapter has shown that SL MT using data-driven methodologies is possible, in that competitive results with both mainstream spoken language MT systems as well as other SL MT systems have been achieved with only the provision of a data set of a few hundred sentences.

Chapter 6

Creating and Evaluating Sign Language

As noted in Chapter 5, for an SL MT system to be of practical use to the Deaf community it requires real SL to be produced rather than annotation. In this chapter, we describe the animation creation and evaluation processes we employ for our system. We first review some possible methodologies for SL creation and discuss the approaches taken by the MT systems mentioned in Chapter 3. We then discuss our chosen process, detailing the animation software and the animated signing mannequin creation process. The second half of this chapter deals with manual evaluation of animations. We outline general manual evaluation methodologies and describe the procedure we chose to best evaluate our animations. In a set of experiments employing native Deaf ISL monitors, we show that almost half of the sentences evaluated are awarded the highest rating. We conclude by discussing the results of the questionnaire that accompanied the evaluations, illustrating more broadly the positive impact of our ISL MT system.

6.1 Generating Real Sign Language

As concluded in the previous chapter, annotation is not a suitable format for MT output in SLs. The alternative is to produce 'real' SL. There are two possible generation options for this:

1. real human signing,
2. animated avatar signing.

6.1.1 Real Human Signing

From a series of experiments comparing human video, animations and SignWriting notation on static and real-time scenarios,[1] Naqvi (2007) shows that BSL users have a preference for human video over animations and SignWriting in both cases. The participants felt that animations and SignWriting failed to adequately capture important characteristics of signing, such as facial features.

Producing videos of humans for either static or real-time SL production is, unfortunately, a time-consuming and impractical option for SL MT. While most SL MT is domain-specific, the provision of a set of pre-recorded human signing videos reduces MT to little more than a look-up table for selecting the correct video to correspond with the English input. In terms of real-time production, there are consistency and smoothing issues for joining video segments of people signing different words together into one sequence. Such a process would involve the segmentation and filing of videos of SL words where the same signer and conditions are maintained. This is simply not a practical option, so the findings of Naqvi (2007) are rendered somewhat less applicable to the task of real-time SL production.

[1] 'Static' refers to pre-recorded video animation or notation of consistent, non-stop signing, whereas 'real-time' refers to separate segments of video, animations or annotation combined together in a series.

6.1.2 Animated Avatar Signing

Avatar animation, on the other hand, circumvents this consistency issue by providing a character, surroundings and other features that can be configured to remain the same throughout the animation process. Computer-generated animations are also better equipped to facilitate smoothing between real-time video segments. In recent years, avatar development has produced characters that are increasingly human-like in appearance and range of movement. This makes avatar animation the most practical choice for SL output from MT systems. While the matter of consistency is easily handled by animation software, there is the requirement of such technology to include the important NMFs of the SLs to ensure maximum comprehension by users.

As avatar animation is the most practical and flexible option, but not the most preferred choice for Deaf users, there are certain criteria an animation should strive to meet in order to achieve the most comprehensible avatar, namely:

- **Realism:** the avatar character should be as realistic and as close to a real human as possible,

- **Consistency:** all features, cameras, characters and any other variables should be consistent throughout,

- **Functionality:** the avatar should have realistic functionality of body movement, facial features and individual fingers on hands and be able to articulate NMFS,

- **Fluidity:** all movement, particularly that of the hands, should flow smoothly within signs and throughout the whole utterance.

In the next section we discuss the SL production processes employed by some of the previous SL MT approaches outlined in Chapter 3.

6.2 Previous Approaches to Sign Language Generation

Of the SL MT systems described in Chapter 3, seven produced real sign language and all used some kind of signing avatar as opposed to videos of human signers. We outline in the following sections the processes used by these approaches and compare them against the criteria listed in the previous section. Evaluation experiments for each are noted and figures of the animated character are provided, where possible.

6.2.1 The Zardoz System

The developers of the Zardoz MT system for SLs (Veale et al., 1998) describe a sign synthesis methodology in earlier work (Conway and Veale, 1994). Their focus is on the importance of describing the internal phonological structure of SLs as an essential factor for the synthesis of native SLs in order to generate fluid signing and allow for inflectional variation. The framework described employs lexical, phonological and phonetic stages. The architecture of the system is shown in Figure 6.1.

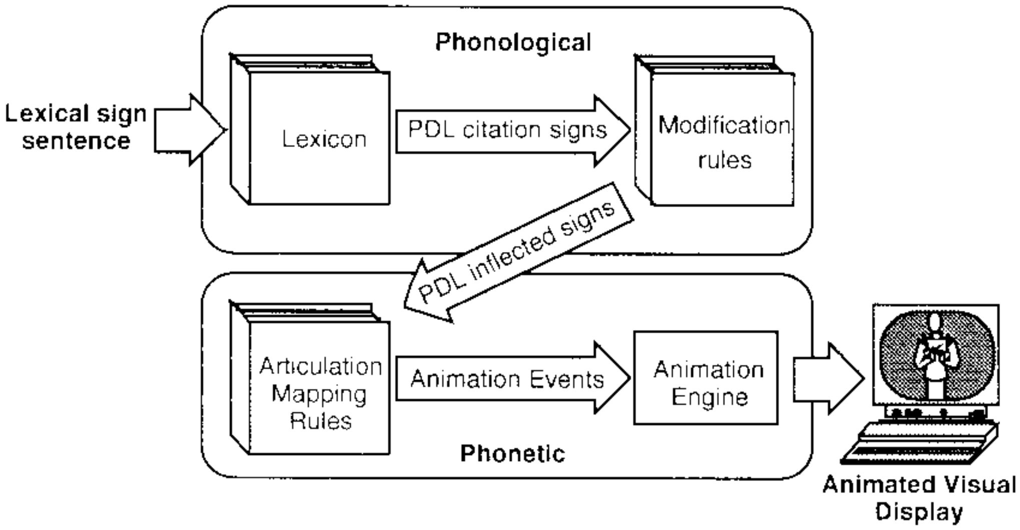


Figure 6.1: The Zardoz Animation Architecture

A series of procedures are employed for SL generation. First the interlingua structure is transformed into a flat stream of sign tokens using heuristic measures to map the concept tokens in the structure to sign tokens using a look-up table. Syntax agencies are also employed here to ensure the correct ordering of tokens. This stream of tokens forms the basis of the 'Doll Control Language' that manipulates an animated doll to articulate the sequence. This stream of tokens is encoded in a glossing methodology borrowed from (Liddell, 1980) that allows for inflectional markers as well as NMFs and noting simultaneous signing. This lexical sign input is fed into the phonological phase where phonological representations of appropriate citation forms of the input signs are taken from a lexicon and grammar modifications are applied to add context- and sign-dependent inflectional information. This intermediate representation of the signed sequence is in a phonological description language (PDL) composed of state and transition representations. A state is described as a snapshot of a sign, and a transition refers to a change of parameters and movement. These representations describe body location, hand configuration and movement phonemes. These representations are then fed into the phonetic phase, where articulatory mapping is used to create phonetic descriptions of the signs. This takes the form of a detailed script of movement required to articulate the input sentence. Each pose is defined using a set of parameters (including facial expression (chosen from fixed set), head orientation, upper body orientation, shrug angle, hand position, palm orientation, elbow raise, hand shape), which are fed in parallel into the animation engine. This in turn appropriately animates a signing avatar creating fluid SL.

Despite detailed descriptions of the processes, the system is not fully implemented but rather a framework for the synthesis of any SL, not just ISL. For this reason, no example avatar is included. This approach does require the provision of a lexicon of citation form signs in phonological representation format as well as a further database of parameters such as inflectional information. The authors stress the importance of including phonological representations in sigh synthesis to accommodate the interpretation of directional verbs and deictic information, for example, as well as allowing fluid, consistent and natural signing.

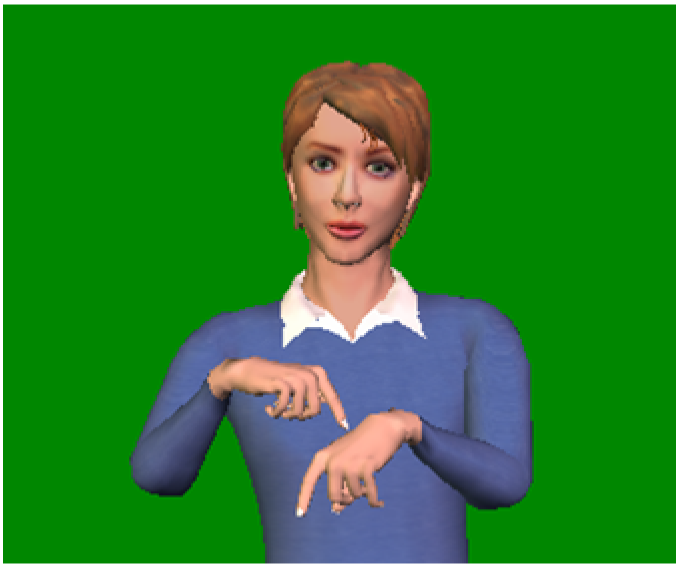


Figure 6.2: The ViSiCAST Avatar

6.2.2 The ViSiCAST Project

The ViSiCAST project (Marshall and Sáfár, 2002; Sáfár and Marshall, 2002; Marshall and Sáfár, 2003) produced HamNoSys as output from the translation system. This symbolic notation is converted to an animated avatar using an interface developed at the University of East Anglia. There is an in-built grammar that consists of a series of constraints that the HamNoSys sequence must satisfy in order to be considered a valid sequence. There is a BSL dictionary of 250 lexical signs, some of which have a fixed descriptions of the signs and are fully instantiated, where others allow for variability, e.g. directional verbs that require loci in the signing space. The resulting description is 'translated' by the interface into a virtual signing human using an avatar illustrated and developed by Televirtual, Norwich, UK. While HamNoSys consists of a linear string of described signs, thereby allowing for fluid SL production, provision is not made for NMFs. The avatar used in this project is shown in Figure 6.2 which shows a realistic female character from the waist up with individual fingers and facial features distinguishable. No manual evaluations of the signing avatar were discussed in their work.

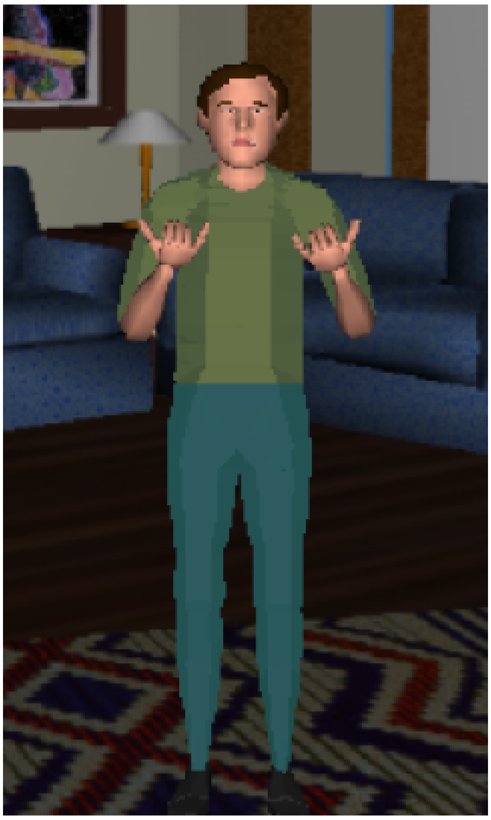


Figure 6.3: The TEAM Project Avatar

6.2.3 The TEAM Project

The intermediate representation from this interlingual MT approach (Zhao et al., 2000) is fed directly into the sign synthesiser. High-level descriptive parameters, which are later converted to low-level qualitative parameters, control the avatar. Default motion templates are provided in an ASL dictionary of signs that are then combined with the parameters from the representation to form each animated sign. In order to smooth the movement between each sign produced, parallel transition networks are employed. An example of the avatar used is shown in Figure 6.3. The male figure depicted is of a basic human form against a 'sitting room' background. Facial features are visible as are the individual hands and fingers. No manual evaluations were carried out.

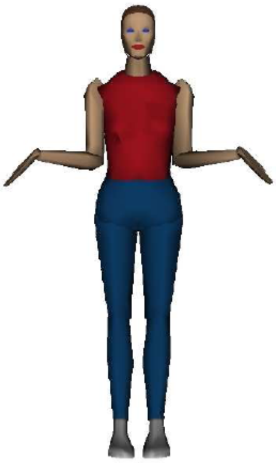


Figure 6.4: The SASL Nancy Avatar

6.2.4 The South African SL Project

van Zijl and Combrink (2006) and van Zijl and Fourie (2007) have developed translation and avatar modules respectively for their work on SASL translation. The modules have not yet been combined to allow the notational parse trees from their MT output to be fed into their signing avatar. For this to happen, the notation from the parse trees would have to be manually transcribed into a script that describes the motor movements of the joints of the signing avatar in order to create each sign. Each script would then be fed into nested queues of concurrent and sequential sign production information, that is in turn fed into the signing avatar and rendered to produce the animation. This queued system of scripted notation allows for fluid signing. The avatar requires further development in order to incorporate facial expressions and the production of phrases. The focus of the animation development of this work is on the creation of a generic signing device. The background animation generation has been plugged into two avatar systems, a cartoon-based character and a more human-like avatar called the Nancy Avatar.[2] The human-like avatar is shown in Figure 6.4 and shows a female figure with basic facial features, with a body that, while clearly human-like, displays basic geometric shaping.

[2] The Nancy Avatar: http://www.ballreich.net/vrml/h-anim/h-anim-examples.html

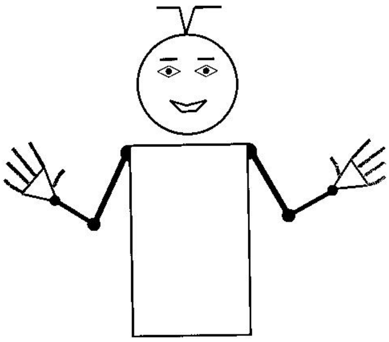


Figure 6.5: The Spanish Sign Language Project Avatar

6.2.5 Spanish SL Project

Unlike the other avatar generation processes described in this section that introduce various 3-D and expressly human-like characters, San Segundo et al. (2007) have developed a 2-D avatar in an attempt to reduce the effort involved in gesture production. An example of the signing mannequin is shown in Figure 6.5. Their character is composed of geometric shapes, such as a rectangle for the body and a circle for the head, and lines forming the arms, fingers and facial features, for example. The SSL semantic concepts produced by the MT system are aligned in n:m alignments with the SSL gestures. A basic set of body positions and facial features are described while continuous signing is achieved via interpolation strategies between these basic positions. To further emphasise the facial expressions, the avatar has been equipped with two antennae-like strands of hair that move in accordance with the facial features. Human evaluations have been carried out on the SL generation of this system by native Deaf signers, but only for the production of letters, as opposed to words or utterances. Evaluations showed that less than 30% (approximately 7 letters) were difficult to understand on first viewing. Although experimentation is in its early stages for this avatar development, the 2-D character is the least human-like figure of the selection described in this section. The character is almost cartoon-like in composition, paricularly the moving hair. Given that Deaf people tend to prefer human signing, this questions the usefulness and comprehensibility of such gesture production even if developed to sign fluidly.

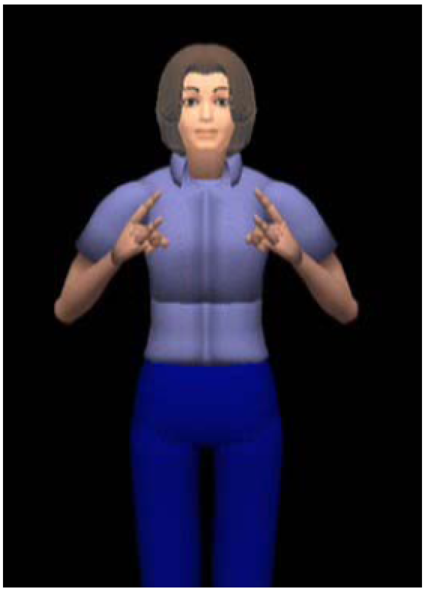


Figure 6.6: The Multi-Path Approach Avatar

6.2.6 The Multi-Path Approach

Huenerfauth's multi-path approach focuses solely on the generation of classifier predicates (CP) and a subset of these are produced in animated form. The output of the CP planning process in the MT system is an ASL surface-form representation that is encoded in the Partition/Constituent formalism mentioned in the description of this system in Section 3.2. Previously calculated discourse models, predicate-argument structures and the visualisation scene are stored in a look-up table for each English sentence along with a library of ASL hand shapes, orientations and locations. These representations are then fed into an animation system developed by the Centre for Human Modelling and Simulation at the University of Pennsylvania. An example of the avatar used is shown in Figure 6.6. It shows a female figure from the knees up that is reasonably natural in appearance. The facial features are distinguishable as are the hand and fingers, which seem a little larger than normal for the size of the figure.

Evaluations were carried out on signed output employing 15 native ASL signers. The evaluation experiment is divided into three comparative experiments where the CP animations were compared with SEE animations and motion-capture anima-

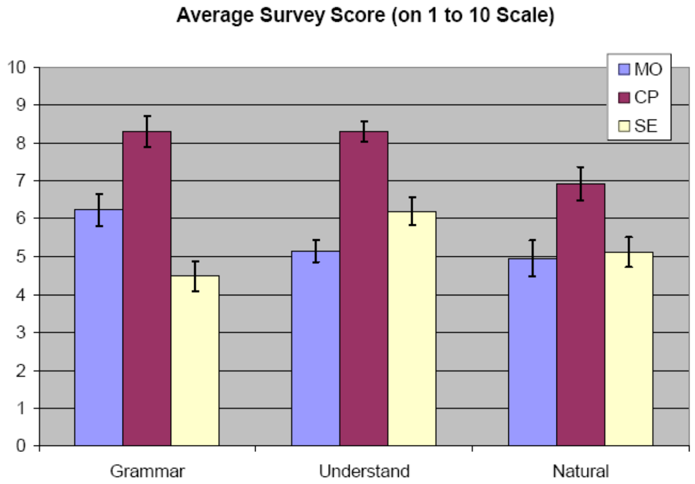


Figure 6.7: Average Scores of Multi-path Animations for Understandability, Naturalness and Grammaticality

tions[3] of the same data. Each was evaluated for understandability, naturalness of movement and ASL grammaticality on a scale of 1 (negative)–10 (positive). Comparative results are shown on the bar-chart in Figure 6.7. The CP animations produced by the prototype system scored higher than both other animations in terms of grammatical correctness, naturalness of movement and understandability. The CP animations attained an average score of just over 8/10 for grammaticality and understandability and almost 7/10 for naturalness. In all cases these are at least 2 scores higher than the other animations.

6.2.7 RWTH Aachen University Group

The SL synthesis phase for MT experiments performed by Stein et al. (2006) is separate to the main MT process. The annotated output produced is fed into the interface used by the ViSiCAST system described previously. Here the annotated output is 'translated' into HamNoSys and the same signing avatar is produced. Unlike the ViSiCAST system, the signed output was manually evaluated by native Deaf signers. In a set of experiments, 30 of the candidate translation sentences and 30 of the reference translation sentences were made into signing videos and

[3] Motion-capture animations involve digitising the co-ordinates of a person in a motion-capture suit and generating animations based on these co-ordinates and movements.

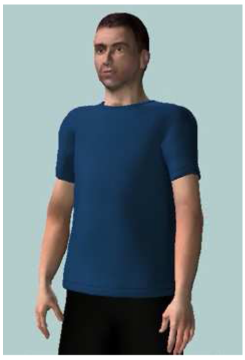
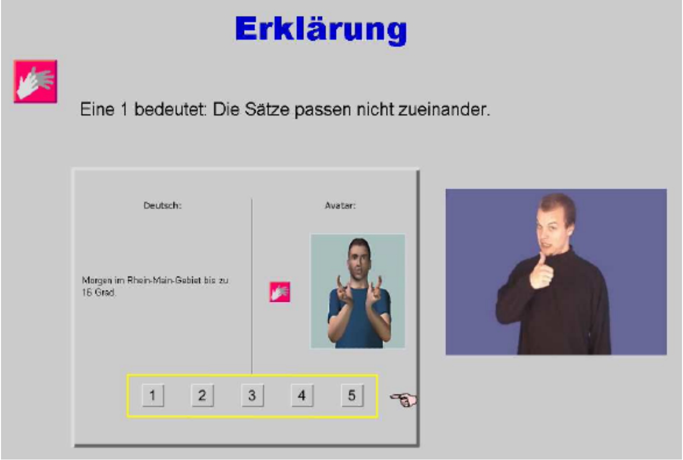


Figure 6.8: The RWTH Aachen University Avatar

the resulting sequences were evaluated. The two evaluators were asked to rate the coherence of the DGS sentence signed by the avatar. The German equivalent was provided in parallel. The authors note that the avatar was poorly supported scoring an average of 3.3 on a scale of 1 (incomprehensible)-to-5 (perfect match). The results, however, are comparable to the automatic evaluation scores of (38.2% WER). The avatar is shown in Figure 6.8. It is similar in form to the ViSiCAST character but male in gender. An example of the evaluation interface is shown in Figure 6.9. The interface displays the avatar in the centre of the screen with the 1–5 rating system in buttons below it and the German sentence to the left. All instructions are provided in SL via a video stream of a human signer.



Erklärung

Eine 1 bedeutet: Die Sätze passen nicht zueinander.

Deutsch:

Morgen im Rhein-Main-Gebiet bis zu 16 Grad.

Figure 6.9: Manual Evaluation Interface Used in RWTH Experiments

6.2.8 Comparing Previous Animation Approaches

From the above descriptions of various animation processes, we can see that by using animated figures, each approach was able to produce SL output using a *consistent* character and format.

There was a wide variety in the forms of animation avatars used. The ViSiCAST project produced the most *realistic*, human-like mannequin, while other systems varied from basic human figures like the TEAM project to the more robot-esque SASL avatar to the linear cartoon-like character chosen by the Spanish team. Given the preference of Deaf people for human signing over animations, we suggest that the ViSiCAST avatar would be the most appealing to a Deaf user, although no such side-by-side comparison has been performed.

All systems can be deemed *functional*l in the sense that facial features and hands and fingers are visible, but the majority of systems fall down in that they do not facilitate NMFs including facial movement in their animations. Given the importance of NMFs and natural body movement in human signing, the translation can only be at a loss for the lack of these features.

In terms of *fluidity* between signs, five of the seven systems report some measures to smooth the transition between signs. While these methods do allow the creation of fluid animations in real-time, the manual scripting of parameters and dictionary entries is a laborious and time-consuming process.

Furthermore, it is difficult to properly assess the usefulness of an animated avatar and the translations it signs if it has not been evaluated. Only three of the seven systems discussed performed a manual evaluation of their animations. Each evaluation shows that the animations have performed well, scoring better than other formalisms Huenerfauth (2006), comprehensibility scores similar to automatic evaluation scores Stein et al. (2006) or the majority of basic signs being understood San Segundo et al. (2007). However, none of these systems detail whether the evaluators liked the animations, whether they found them useful or whether the lack of NMFs affected their judgment.

	Realism	Consistency	Functionality	Fluidity	Evaluated
Zardoz		√	√	√	
ViSiCAST	√	√		√	
TEAM	-	√		√	
SASL	-	√			
Spanish		√	√		√
Multi-Path	-	√		√	√
RWTH	√	√			√

Table 6.1: Comparison of MT Animation Approaches

Table 6.1 compares each system according to the criteria listed in Section 6.1.2. A ‘√’ indicates that the system meets the criterion, a ‘-’ indicates that the system meets the criterion but poorly, and an empty box indicates the system does not meet the criterion in question.

6.3 Our Approach to Sign Language Animation

For our own SL generation module, we chose 50 randomly selected sentences from the output of our EN-ISL experiments described in Chapter 5. Each annotated gloss, or ‘token’, from the MT output was made into a separate video and these videos were joined to form SL sentences signed by our 3–D avatar. The animation process is described in the following sections.

6.3.1 Animation Software

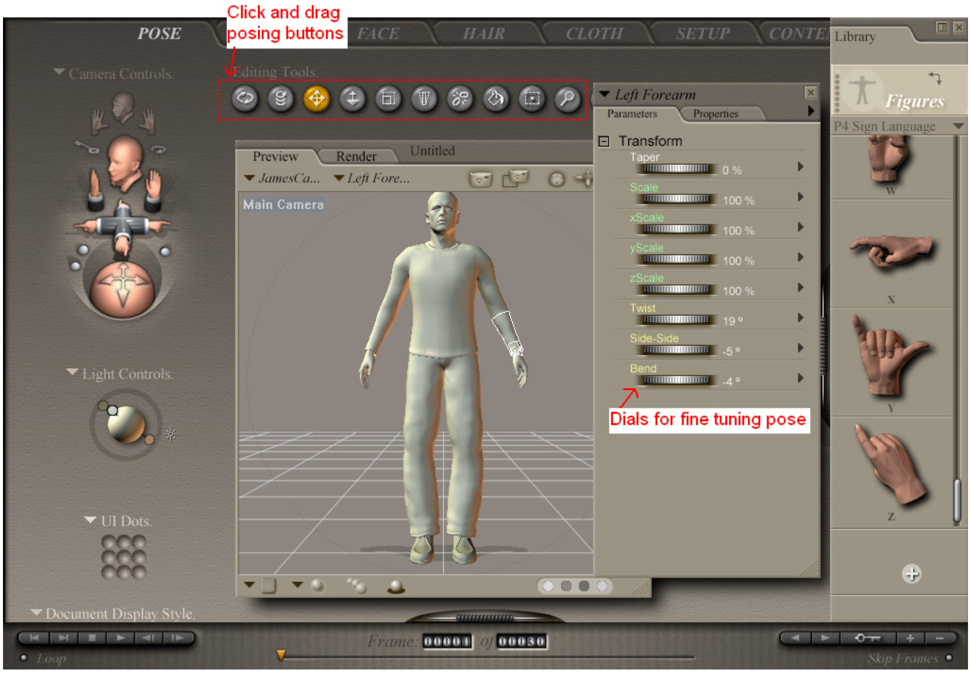
Subsequent to the experiments of Naqvi (2007) detailing Deaf people’s preference for human signing, but given the impracticalities of producing human signing for our MT output, we sought to source the most realistic, human-like animation avatar. As our research is primarily focussed on MT rather than animation production, so in order to achieve the most life-like avatar, we chose a commercial animation tool, as opposed to developing our own. We found suitable characters in Poser Animation Software, Version 6.0.[4] The Poser software tool enables the animation of 3–D human figures. The tool facilitates the creation of new mannequins but also has a library of pre-created characters of various age, gender and feature specifications. Comprehensive libraries are included encompassing various figures, poses, hand shapes and camera configurations. Poser allows the user to create animations using two methodologies: feeding a pre-coded script in the Python programming language of parameters and movements for all features of the animation or manual posing of the figure on a framed basis.

Creating Python scripts describing the various parameters for the animation, including basic figure positions, camera and lighting set ups as well as the phonetic features of the signs to be articulated, facilitates consistency across all animations. Furthermore, combining these signing scripts and feeding them to the avatar can facilitate fluid, natural signing. While this would be ideal for our purposes, early experimentation proved that the process is prohibitively time-expensive for creating the movements required for even semi-fluid signing. For instance, in order to move a figure a script would need to specify the co-ordinates of each joint of the figure for each posed frame. Ensuring correct positioning of the figure in each frame is an awkward and laborious process, even without co-ordinating NMFs.

The alternative is manual posing of a chosen figure. Manual posing involves the positioning of the avatar and its body parts directly in the interface using functions such as twist, rotate, up/down, side-side from buttons on the Poser interface, or by altering the degrees of these functions. These allow for precise positioning of the figure, in particular the facial features and the fingers. The animation interface for Poser 6 is shown in Figure 6.10. The camera and light controls are shown to the left of the screen, while the mannequin appears in the centre with posing buttons above, and fine-tuning posing dials to its right. At the base of the screen is a frame counter together with buttons for playing, pausing and changing the frame sequence.

To create animations using this manual method, the user sets up appropriate

[4] http://www.curiouslabs.de/poser6.html?&L=1



Click and drag posing buttons

Left Forearm
Parameters Properties
Transform
Taper 0 %
Scale 100 %
xScale 100 %
yScale 100 %
zScale 100 %
Twist 19 °
Side-Side -5 °
Bend -4 °

Dials for fine tuning pose

Frame: 00001 of 00030

Figure 6.10: The POSER 6 Animation Interface

lighting and camera features and poses the mannequin. The pose can be altered at different frame intervals, depending on the speed of the movement desired. One of the advantageous functions of POSER is that it interpolates the transitions between fixed poses in order to create natural, human-like movement. For example, if at Frame 1 the mannequin is posed with his right arm raised in the air and at Frame 10 it is posed with his right arm by his side, POSER will alter the intermediate frames (2–9) to gradually and sequentially move the arm from point A in Frame 1 to point B in Frame 10. This is shown in Figure 6.11.

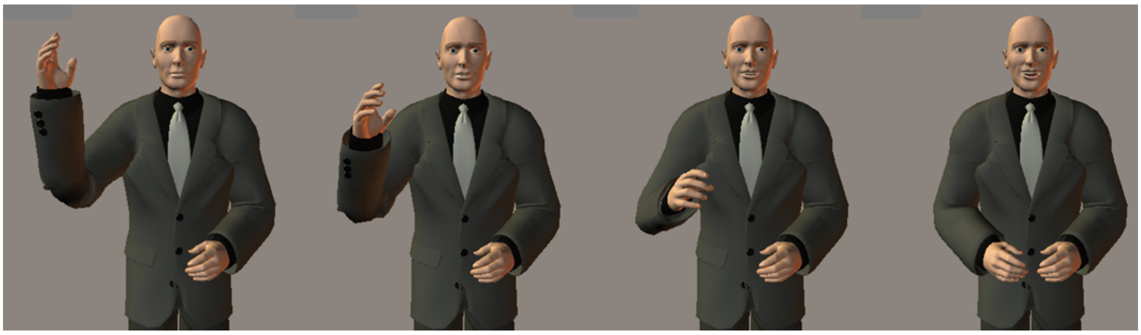


Figure 6.11: Example of Interpolation Between Frames in POSER

The manual creation of individual signs can be a laborious process. However, when compared with the time taken to produce Python scripts for each sign, manual animation is a more practical process. Furthermore, manual posing enables the user to produce detailed movements and articulations of the mannequin, particularly the fingers and the facial-features, something that could take days to articulate via Python scripts. Manual creation also allows the user to see exactly what is happening to the mannequin as it is being manipulated. For these reasons, we opted to manually create the ISL signs. Having previously worked with manual animation of SLs using Poser 4, we were able to further speed up the animation process.

6.3.2 Creating the Animations

Taking the 50 candidate annotation translations from our EN–ISL experiments, we segmented the data into individual words, giving 246 tokens with 66 individual types.[5] Each of the 66 ISL tokens was then created into an individual video using the manual animation process described above.

Rather than developing our own mannequin, we chose the business man figure from the selection of pre-created characters as he was one of the formal characters, more appropriate for MT translation than some of the more video-game-like characters available. The mannequin, who we named Robert,[6] is shown in Figure 6.12.

Ideally, each animated sign would blend seamlessly with the next in fluid, natural articulation. As each sign produced by our system is animated manually and individually, this can pose some fluidity problems when joining the signs to form sentences. It is not yet possible to join the manually created signs together in Poser in real-time to avail of the interpolation technique and thus avoid 'jumping' between animations. In order to overcome this problem, to produce real-time signing and

[5] 'Type' refers to unique words and 'token' refers to instances of words. For example in the sentence *The dog saw the cat*, there are five tokens but only four unique types as *the* is repeated.

[6] This name was given to the mannequin following previous work where the figure was used in a rule-based MT system, hence the name RoBerT

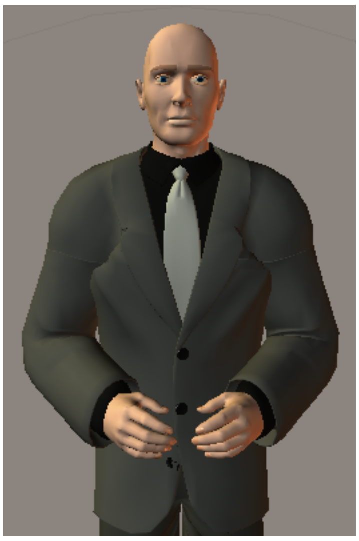


Figure 6.12: Robert: The Poser 6 Avatar

to minimise jumping between signs in a sentence, we formated each animation so that the mannequin began and finished each individual sign animation in a neutral position with his hands resting in front of and close to his body and his face relatively expressionless (cf. Figure 6.12). While it is not natural to pause in this neutral position between each sign in normal discourse, this smoothing methodology is adequate for our purposes at this stage. Future development of this animation approach will seek to remedy this issue.

Although we have ISL competencies ourselves, for the SL animation we sought to further support our knowledge by examining the original ISL videos created for the corpus to confirm the correct articulation of each of the 66 signs. These videos were also employed to provide examples of natural body movement and NMF detail, particularly mouth and eyebrow patterns. This enabled us to improve the already human-like mannequin by adding in natural body movement making Robert less stiff and robotic in nature. We considered seeking the assistance of native ISL signers to help create and review our animations, but finding alternative candidates to those of our evaluation set proved difficult. As we did not want our evaluators to be biased in their judgment from having helped develop the animations, we chose not to overlap their duties and instead relied on examples provided in the ISL videos.

Within each sign animation, the figure was posed at the start and end points of signs allowing the interpolation technique to fill in between. In some cases, particularly for more complex signs, intermediate posed frames were added to ensure a fluid flow of movement, e.g. signs that involve a circular motion. In order to generate signing at the most natural speed possible, we set poses ten frames apart, allowing Poser to create the interim movements. Employing this process allowed us to generate more natural, human-like signing possible.

The 66 videos took approximately a week to format and review, which was assisted by the Poser library of pre-defined ASL letter hand shapes. This was particularly useful for quick configuration of the mannequin's hands as 18 of the 26 letters in ASL are common ISL hand shapes. Other basic pre-defined hand shape libraries were also used. This helped us to speed up the animation process as well as ensuring accurate positioning of fingers for signs.

Videos of the combined frames were compiled into .avi format using the Poser software. The files in this format were quite large, which would cause them to take longer to load and play, as well as causing storage issues. In order to decrease the file size for each video, STOIK Video Converter 2[7] was used to convert the videos to smaller .wmv format without losing out on picture quality.

6.3.3 Comparing Robert Animations with Previous Approaches

From the above description of our animation processes, we can see that our approach meets the four criteria listed in Section 6.1.2:

- **Realism**: our choice of the Poser animation software ensures our signing mannequin, Robert, was as realistic and human-like as possible,
- **Consistency**: setting up a basic format that we use throughout the animations ensures consistency of all pertinent features,
- **Functionality**: allowing detailed manipulations of Robert's facial features and fingers as well as the interpolation between signs creates a functional

[7] http://www.stoik.com/products/svc/svc_main.htm

mannequin with NMF ability and natural, human-like movement,

- **Fluidity**: interpolation between posed frames ensures fluidity within each sign animation, and the inclusion of a neutral pose at the start and end of each sign smooths the transitions between signs when joined in sentence format.

Table 6.2 below compares our approach (in **bold** face) with the previous approaches discussed in Section 6.2.

	Realism	Consistency	Functionality	Fluidity	Evaluated
Zardoz		√	√	√	
ViSiCAST	√	√		√	
TEAM	-	√	?	√	
SASL	-	√			
Spanish		√	√		√
Multi-Path	-	√		√	√
RWTH	√	√			√
MaTrEx	√	√	√	√	√

Table 6.2: Comparison MaTrEx Animation with Other MT Animation Approaches

We can see that where other systems fail, particularly in the areas of *functionality* and *evaluation*, the animations produced using our processes show an improvement on past approaches. In the next section we discuss the evaluation process.

6.4 Human Evaluation of Machine Translation

As noted in Chapter 4, it is only within the last decade that automatic evaluation metrics have been employed for measuring the correctness of MT output. Prior to this, evaluation of translations was undertaken manually by human evaluators.

Human evaluation is an important contribution to MT, particularly so in SL MT, given that the evaluators are generally the prospective beneficiaries of the system. This out weighs the potentially negative aspects of this type of evaluation, namely the subjectivity of using humans and the time-consuming process of recruiting evaluators, setting up procedures and the evaluation itself. Ultimately, no matter how good a score is obtained using automatic metrics, if the end-users do not consider the output as correct, or understandable and helpful, the automatic evaluation is rendered somewhat less relevant.

6.4.1 Judging Criteria

Given that subjectivity is a factor in human evaluations, it is important to standardise the evaluation process as much as possible. Results cannot be considered significant if evaluators are not given criteria and a scale for scoring translations. Standardising manual evaluation procedures dates back to the work of Carroll from his study in the ALPAC[8] report of 1966 (Pierce et al., 1966). In his study, Carroll describes experiments using three judgment criteria:

1. intelligibility of the output, independent of the source text,

2. fidelity of the output as a translation of the source text,

3. the reading/rating times of the judges.

Comparatively, the later work of van Slype (1979) outlines intelligibility and fidelity as important methods for rating translations as part of what he calls a ‘superficial evaluation’, assessing the acceptability of the MT system in question. Termed as ‘fluency’ and ‘adequacy’ respectively, intelligibility and fidelity arise again in by White et al. (1994) and LDC (2005) as chosen features for evaluation. Therefore, it may be considered that, since the beginnings of manual evaluation of MT output, an assessment of both the quality of the translation in its own right as well as the quality of the translation when compared against the source or reference is necessary for proper human judgment. As automatic evaluation metrics generally only compare against reference texts, they are somewhat flawed in their methodologies.

[8] ALPAC (Automatic Language Processing Advisory Committee) assessed the progress and prospects of human language technology from its inception in 1964 under the auspices of the US Government.

6.4.2 Scoring Scales

Past manual evaluation procedures have employed various scales in order to standardise the process, at least within the context of the experiment in hand. Experiments carried out by Carroll employed two comprehensive scales for scoring evaluations: a 9-point scale for intelligibility and a 10-point scale for fidelity. Both included descriptions of criteria for each point on the scale. A 9- or 10-point scale does provide an in-depth analysis of translations, but it further increases the time incurred to perform the evaluation. Later work by van Slype (1979) employed a 4-point scale, while the more recent scale developed by LDC employs 5-points with descriptions included. The latter has become the most frequently employed scale for human judgments in shared MT tasks such as the NIST Open Machine Translation Campaign.[9]

6.4.3 Number of Evaluators

For reasons of inter-rater variance, it is broadly considered that a number of evaluators should be selected, although there is a difference of opinion when considering the number of evaluators necessary for adequate manual evaluation. Carroll asserts that at least 3 or 4 evaluators are required, whereas van Slype segregates the scoring methods, stating that intelligibility should be assessed by 4-10 evaluators, but fidelity requires only 1 or 2. This highlights the necessary trade-off in manual evaluation between the number of evaluators against the time evaluations take. A sufficient number of evaluations should be performed to show some level of significance for the scoring, yet too many evaluators or translations can make the task even more time-consuming than it already is. In the next section, we discuss how we have used these methods in our evaluations.

[9] http://www.nist.gov/speech/tests/mt/

6.5 Manual Evaluation of Sign Language Video Output from the MaTrEx System

6.5.1 Evaluation Methodology

Setting up evaluation procedures for our data, we took matters of methodology, evaluation scales and number of evaluators into account. As noted in the previous section, the consensus for manual evaluation is for an assessment of both the intelligibility/fluency and the fidelity/adequacy of candidate translations. In line with this, we chose to use both methods (herein referred to as intelligibility and fidelity) and decided on a scale of 1-4 with descriptions, as follows:

Intelligibility:

1. Incorrect or too confusing to grasp the meaning

2. Difficult to understand but I grasp the gist of it

3. Understood but somewhat incorrect

4. Understood and correct

Fidelity:

1. Completely incorrect translation

2. Basic concepts are correct but mostly incorrect or information missing

3. Good translation, a few things incorrectly translated or missing

4. Excellent translation, no errors

Despite a 5-point scale being the most common choice for manual evaluation tasks, we chose a 4-point scale, similar to van Slype. The decision to opt for an even-numbered scale, requires the evaluator to choose the positive or negative side of the scale and "provides no 'common middle ground' " for judges (Owczarzak, 2008). In order to qualify the ratings and assist evaluators, we provided short descriptions of the criteria which a translation should match in order to be given that score. We described the four categories of this simple scale to offer as fair a set of options as possible. For both intelligibility and fidelity, the evaluators have two 'negative' ratings and two 'positive' ratings. Roughly speaking, this allows the evaluator to attribute a completely negative, mostly negative, mostly positive or completely positive rating to each translation. While using a small scale for evaluation reduced the analysis and decision-making time of the evaluator, it may also be considered less reliable as evaluators can group sentences together in under the one evaluation that might other wise receive quite different scores if the scale was much broader. For this reason, its results might be deemed less significant than a larger more detailed and analytical scale. However, we feel that the scale and methods we have chosen are adequate for the task we are performing, given that these are the first human judgments of our first proper SL MT experiments.

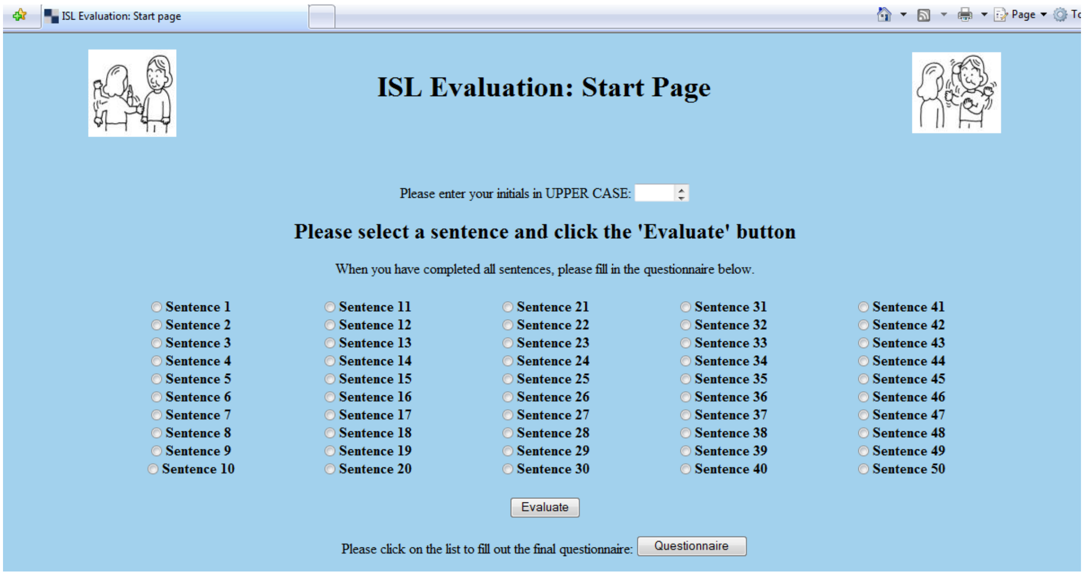
It has become customary to only provide sample sets of the translation data for evaluation, e.g. Paul (2006) in the IWSLT 2006 evaluation and Callison-Burch et al. (2007) in the ACL 2007 shared task. Given the time-consuming process of video development for SL animations, we chose to follow suit and use a subset of our candidate translations. Of the 118 translations produced, 50 (just over 40%) were selected as representatives of the whole.

Following the work of Pierce et al. (1966) and van Slype (1979), we chose to recruit 4 ISL evaluators, or 'monitors' as they are referred to, for the task. With only approximately 5,000 members in the Irish Deaf Community, finding a subset of this already small group that were willing to help proved a difficult task. Despite assurances that the task could be completed by the monitors on their home computer at their own convenience and would only take approximately 3 hours to complete, our initial communication with the Irish Deaf Society (IDS) to assist us in our search resulted in only two potential candidates. Subsequent requests by the IDS and the two candidates resulted in a further two candidates. Luckily, each candidate was equipped with the necessary computer facilities for the task and were willing to help.

6.5.2 Evaluation Interface and Format

In order to facilitate the ISL monitors, we sought to make the task as simple as possible. We developed a web-based format for the evaluations where the monitors were provided with some background, a description of the project and detailed instructions for the task at hand. While it was necessary for the monitors to have Internet Explorer and Windows Media Player on their computers, these are standard features on most home computers so did not prove to be a problem. The framework could easily be extended at a later stage to encompass other browsers and video players.

Following a brief testing session to ensure the videos were playing correctly, the monitors were guided toward the main task. A screen shot of the central navigational page is shown in Figure 6.13.



ISL Evaluation: Start Page

Please enter your initials in UPPER CASE:

Please select a sentence and click the 'Evaluate' button

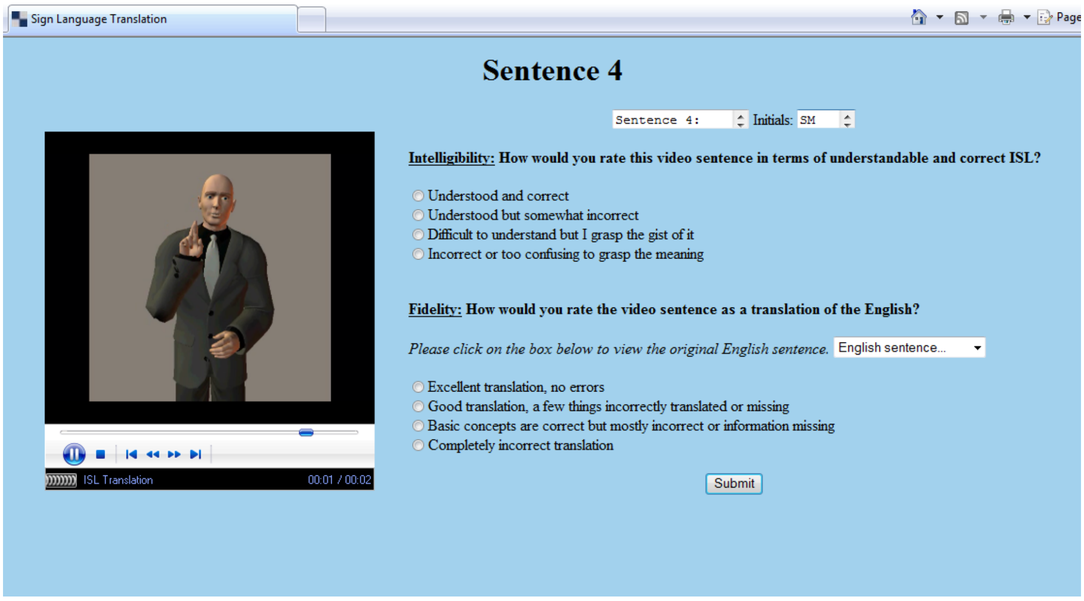
When you have completed all sentences, please fill in the questionnaire below.

○ Sentence 1
○ Sentence 2
○ Sentence 3
○ Sentence 4
○ Sentence 5
○ Sentence 6
○ Sentence 7
○ Sentence 8
○ Sentence 9
○ Sentence 10
○ Sentence 11
○ Sentence 12
○ Sentence 13
○ Sentence 14
○ Sentence 15
○ Sentence 16
○ Sentence 17
○ Sentence 18
○ Sentence 19
○ Sentence 20
○ Sentence 21
○ Sentence 22
○ Sentence 23
○ Sentence 24
○ Sentence 25
○ Sentence 26
○ Sentence 27
○ Sentence 28
○ Sentence 29
○ Sentence 30
○ Sentence 31
○ Sentence 32
○ Sentence 33
○ Sentence 34
○ Sentence 35
○ Sentence 36
○ Sentence 37
○ Sentence 38
○ Sentence 39
○ Sentence 40
○ Sentence 41
○ Sentence 42
○ Sentence 43
○ Sentence 44
○ Sentence 45
○ Sentence 46
○ Sentence 47
○ Sentence 48
○ Sentence 49
○ Sentence 50

Please click on the list to fill out the final questionnaire:

Figure 6.13: ISL Evaluation Central Navigational Page

This page consists of a text box in which the monitor types his/her initials, a set of 50 radio buttons numbered Sentence 1–50, a button marked ‘Evaluate’ and a button marked ‘Questionnaire’. By providing his/her initials each monitor ensured that all evaluations were catalogued in a file specific to him/her. In order to commence evaluation, the monitor clicks on the relevant sentence number and then on the evaluate button. This opens a new window for each sentence so that



Sentence 4

Sentence 4: Initials: SM

Intelligibility: How would you rate this video sentence in terms of understandable and correct ISL?

- ○ Understood and correct
- ○ Understood but somewhat incorrect
- ○ Difficult to understand but I grasp the gist of it
- ○ Incorrect or too confusing to grasp the meaning

Fidelity: How would you rate the video sentence as a translation of the English?

Please click on the box below to view the original English sentence. English sentence...

- ○ Excellent translation, no errors
- ○ Good translation, a few things incorrectly translated or missing
- ○ Basic concepts are correct but mostly incorrect or information missing
- ○ Completely incorrect translation

Figure 6.14: ISL Evaluation Sample Sentence Evaluation Page

when they return to the central navigation page, they can see what sentence number they have just evaluated by looking at the radio button. An example of one such evaluation page is shown in Figure 6.14.

Each evaluation web-page takes the same format; on the left hand side of the page a video screen appears that automatically streams the video appropriate to the sentence selected on the navigation page. The ISL sentence can be played as many times as the monitor wishes. On the right hand side, intelligibility and fidelity ratings are shown. In order to judge intelligibility the monitor is asked:

(25) Intelligibility: How would you rate this video sentence in terms of understandable and correct ISL?

This is followed by the 4 intelligibility ratings shown in Section 6.5.1. This is rated without viewing either a reference or source text so as not to bias the judges. Below this is the fidelity rating:

(26) Fidelity: How would you rate the video sentence as a translation of the English?

The source English sentence is provided in a drop-down menu, with the sentence itself hidden from view until the monitor reveals it by clicking on it. This is followed by the 4 fidelity ratings also shown in Section 6.5.1.

There are two possibilities when measuring the fidelity of a candidate translation: comparing it with a reference translation and comparing it with the source text. This can depend on the focus of the evaluations and the language knowledge of the evaluators. If the evaluators are bilingual, they can assess the fidelity of the candidate as a translation of the source text. If the evaluators do not have language competencies in the source language, reference translations can be used for comparison. This raises an interesting issue for SL MT evaluation. The monitors are considered bilingual, in that they have competencies in both languages, but, as noted in Section 2.2.5, they are not predominantly fluent in their second language, English. The natural assumption is to provide reference translations as the alternative. A reference translation, in the case of SL MT could take three forms:

1. an annotation,
2. a video of a human signing,
3. a 'gold standard' animation.

Annotation is not an appropriate choice for comparison given that the majority of Deaf people would not be familiar with this type of notation for their language. Furthermore, it could be likened to comparing synthesised speech with text to see if it is a good translation, rather than comparing like with like. The video of a human signing the sentence could be an ideal reference text, but given the propensity for Deaf people to favour human signing over animated signing (Naqvi, 2007), it is thought that such a comparison would be unfairly biased against the animations and therefore not provide an adequate analysis of fidelity. There is also a difficulty in finding separate signers and evaluators. The third option and the most appropriate comparative reference is a gold standard animation. Unfortunately the labour-intensive manual process of video creation for individual videos, never mind whole sentences, is prohibitively expensive for our experiments. This is, however, the ideal reference translation for SL comparison in fidelity evaluations, being the least biased and most similar to[10] the candidate translations.

Consequently, we decided that a comparison with the English source text would be the most appropriate means of rating the fidelity of the ISL translations. Contrary to van Slype (1979)'s argument that a fidelity evaluation only requires one evaluator, we feel that given the variances in English fluency of the ISL monitors, it is necessary to have all 4 individuals assess the fidelity of the translation.

Once the monitors have chosen their ratings and clicked on the appropriate radio buttons, they click on the 'Submit' button at the bottom of the page. This registers their evaluations for that sentence in their own file and replaces the evaluation page with an overview of the rating chosen for that sentence and instructions to close this window and return to the central navigation page. The monitors have the option of returning to any sentence at any time to alter their evaluations. Once finished, the monitors are asked to complete the questionnaire.

On the basis that this is the first human evaluation experiment for the MaTrEx system using SL generated by a Poser avatar, we chose to include a questionnaire to assess more broadly the animations and to help direct our future work. It included questions on the speed of signing, the naturalness of the animations and the potential use of a system such as ours for the Deaf community. Most questions required the monitors to choose a 'yes', 'no' or multiple rating choice for the question combined with space for comments or qualifying their choice.

6.6 Manual Evaluation Results

Intelligibility and fidelity scores along with the answers to the questionnaires were collected from the monitors' individual files. We first show and discuss results for each evaluation method separately. This is followed by an overview of the general

[10] By 'similar to' we mean in terms of format, as opposed to linguistic similarity which could unfairly affect the ratings.

information gleaned from the questionnaire.

6.6.1 Intelligibility

Figure 6.15 shows a bar-graph detailing each monitor's intelligibility scores. The *x-axis* illustrates the score given (1–4) and the *y-axis* shows the number of sentences that received that mark. Each monitor is shown using a separate colour.

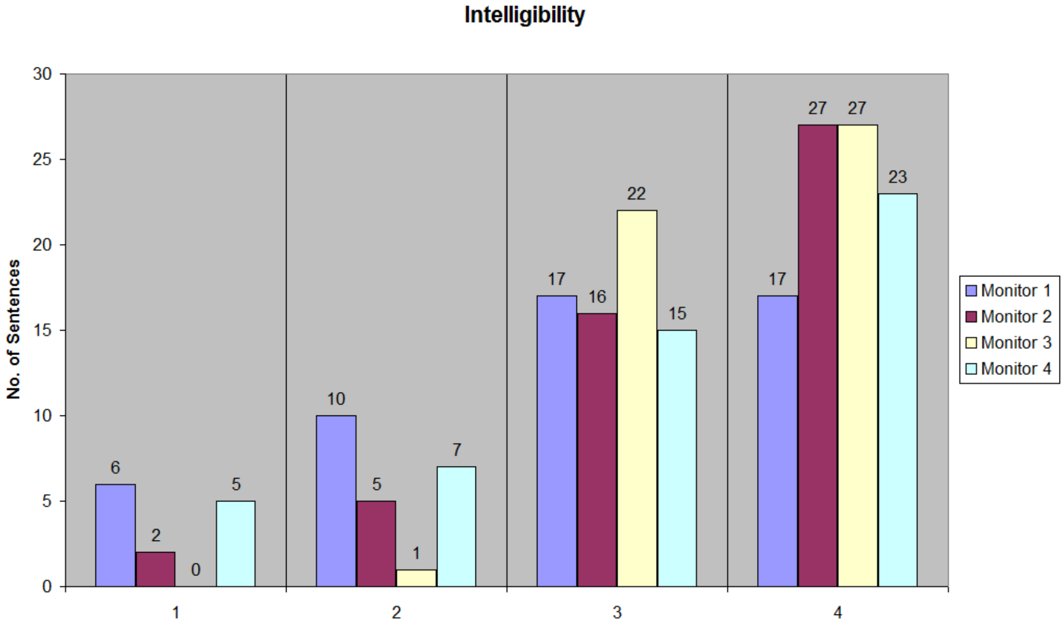


Figure 6.15: Bar-Graph Showing Manual Evaluation Intelligibility Scores

From this graph we can see that, broadly speaking, the monitor's evaluations correlate across all categories. The lowest score is given to the least number of sentences, with sentence numbers steadily increasing and reaching a peak at the highest score.

The number of sentences per evaluation category range between:

- 0–6 (3.25 average): incorrect or too confusing to understand,
- 1–10 (5.75 average): difficult to understand but I get the gist of it,
- 15–22 (17.5 average): understood but somewhat incorrect,
- 17–27 (23.5 average): understood and correct.

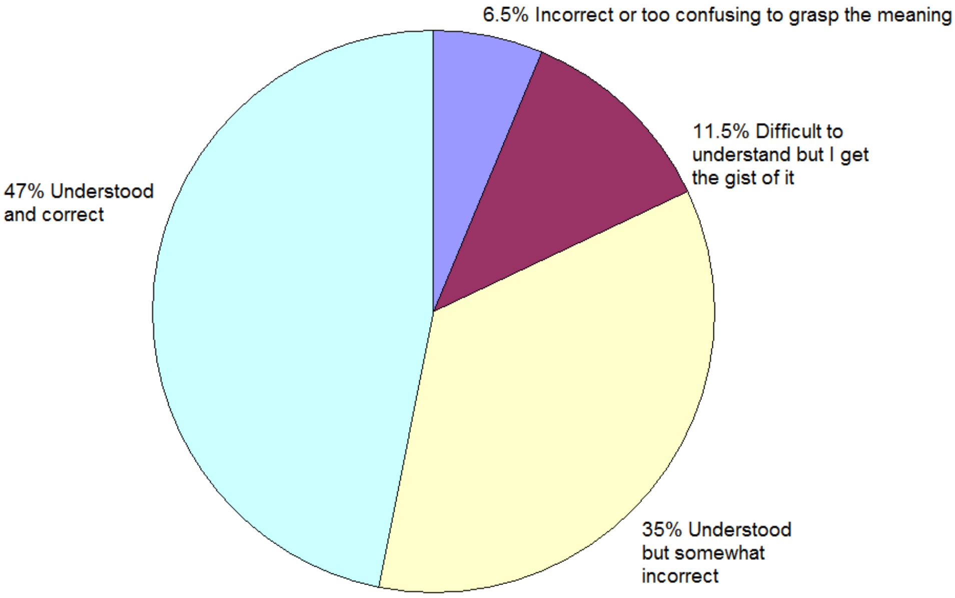


Figure 6.16: Pie-Chart Showing Percentage of Scores for Intelligibility

This shows that the large majority of animations produced from our translations were understandable and mostly correct. This is shown more clearly on the pie-chart in Figure 6.16 where 47% of scores given were in the highest category *understood and correct*. In fact, combining the top two scoring categories, 82% of sentences were considered intelligible by the monitors.

From the pie-chart we can also see that only 6.5%, i.e. 13 of the 200 sentences evaluated, were considered completely incorrect or too confusing to understand. Only a further 13.5% were considered difficult to understand, but even then monitors managed to understand the gist of the animated sentence.

There are, however, some discrepancies between monitor evaluations. From the bar-graph in Figure 6.15 we can see that Monitor 1 is the most critical of the output giving the most number of 1 and 2 category scores, and only 34% of evaluations being in the highest category, unlike the other monitors where almost half of the sentences' scores were in category 4.

In contrast, we can see that Monitor 3 is the least critical and does not give any category 1 scores, with only one sentence receiving a category 2 scoring. Despite this, as mentioned above, all monitors gave consistently better scores to more sentences.

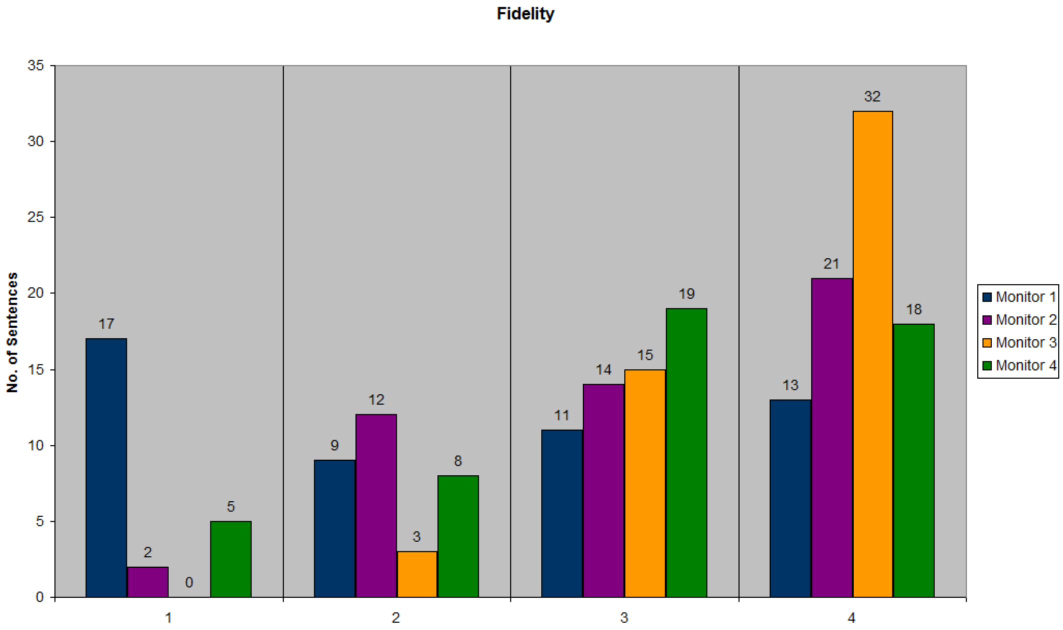


Figure 6.17: Bar-Graph Showing Manual Evaluation Fidelity Scores

6.6.2 Fidelity

In terms of rating the ISL animations as a translation of the English sentence, the Figure 6.17 shows a bar-graph detailing each monitor's fidelity scores, rating the ISL animations as a translation of the English sentences. The *x-axis* illustrates the score given (1–4) and the *y-axis* shows the number of sentences which received that mark. Each monitor is shown using a separate colour.

Assessing the quality of translations, there is increased variation across evaluations compared with the intelligibility scores.

The number of sentences per evaluation category range between:

- 0–17 (6 average): completely incorrect,
- 3–13 (8 average): basic concepts are correct but mostly incorrect or information missing,
- 11–19 (14.75 average): good translation, a few things incorrectly translated or missing,
- 13–32 (21 average): excellent translation.

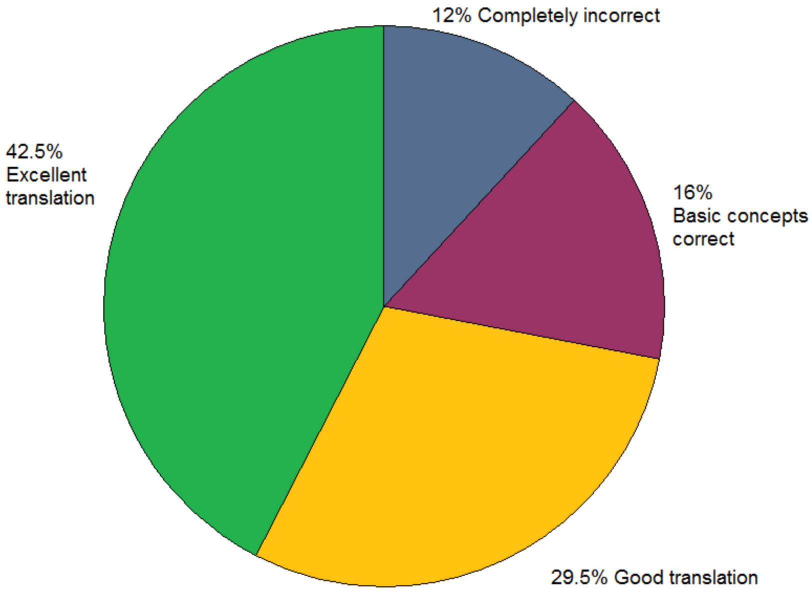


Figure 6.18: Pie-Chart Showing Percentage of Scores for Fidelity

Compared to the intelligibility scores, this shows that the large majority of ISL animations were considered good translations. From the pie-chart in Figure 6.18, we can see that 42.5% of sentences evaluated received the top score. Furthermore, the top-two scoring categories combined make up 72% of the evaluations. This shows that almost three quarters of the sentences evaluated were considered good or excellent translations of the English source sentence. 28% of sentences received a score from category 1 or 2. This shows that not even a third of sentences were considered bad translations, and only 12% were considered completely incorrect.

With the exception of Monitor 1, scores increase from the lowest to the highest category. Monitor 1, on the other hand, gives a category 1 evaluation to the most number of sentences. Similar to the intelligibility scoring, Monitor 3 again gives the most positive scoring with no category 1 evaluations attributed. These discrepancies in scoring could be attributed to the visual factor inherent in SL animation evaluation that is not present in spoken language text evaluation. It is possible that Monitor 1 did not like the animation format and was, therefore, less inclined to give positive scores. Similarly, Monitor 3 may have particularly liked the idea of a signing avatar and may have been more inclined to give positive scores. Given this variance in the users' opinion of fidelity, if we had followed van Slype (1979)'s opinion that only 1–2 evaluators are needed for this type of evaluation, our results would have been less conclusive, particularly if only Monitor 1 and Monitor 3 had been employed. We further address the reasons for this inter-monitor variance in Section 6.6.4 when discussing the questionnaire.

6.6.3 Comparing Intelligibility and Fidelity Scores

Given the small scale of only 4 evaluation categories for the two evaluation criteria, it is possible that the score of one would influence the score of another. We examine this in Figure 6.19.

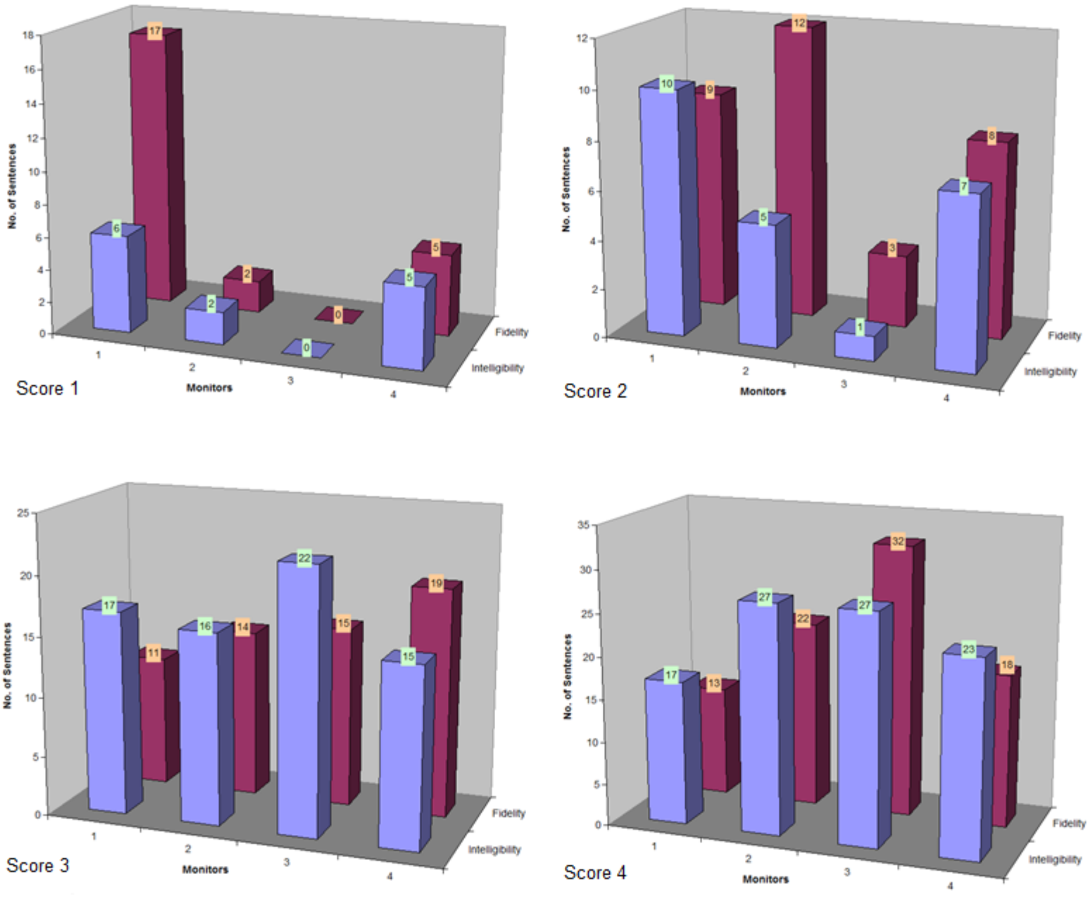


Figure 6.19: Comparison of Intelligibility and Fidelity Scores

With the exception of Monitor 1, we can see that category 1 scores were evenly distributed across sentences for both intelligibility and fidelity. This indicates that if a sentence was not understandable, it was also not a good translation.

Category 2 evaluations show a marked variance across monitors when comparing intelligibility and fidelity. Monitors 2, 3 and 4 gave more scores of 2 for the fidelity evaluations, 7, 2 and 1 respectively. Monitor 1, on the other hand, gave one more score of 2 to intelligibility over fidelity.

Category 3 evaluations were more evenly distributed across intelligibility and fidelity ratings with an average difference of 4.74 sentences between them. Category 4 evaluations show similar even scoring by each monitor, with an average difference between intelligibility and fidelity of 4.75 sentences. With the exception of Monitor 3 for both these categories, intelligibility scored higher than fidelity. This indicates that even if an animation is not quite an accurate translation of the English source sentence, it can still be considered understandable and correct ISL.

Close examination of the individual evaluations showed that with the exception of Monitor 1, all evaluators scored the sentences with either the same rating or a rating with a difference of 1, where 72.5% of the time intelligibility was rated higher than fidelity. Across all monitors, the same score was given for both ratings an average of 65.5% of the time. This demonstrates a high correlation between our animations being a good translation as well as being understandable in their own right.

6.6.4 Questionnaire

We chose to include a questionnaire with our evaluation in order to assess the monitors' opinions of the animations and the avatar in general. We asked the monitors questions about what they thought of the animations, the speed of signing and the naturalness of the mannequin. We queried their opinion on the NMFs added to the mannequin and whether they found it difficult to understand a computer-generated animation. We asked what they thought could be improved and what they liked most and least. We then concluded by asking if they thought technology such as the MT system they had been evaluating would be useful in the Deaf community.

In general, the animations were well received with one monitor noting in particular the details of the fingers and mouth patterns. Three of the four found the speed of signing just right for them, with the third saying it was fine for first time users but a little slow for them. The mannequin was considered to be quite natural for the most part but a little stiff. All monitors admitted that they found our smoothing technique of resuming a neutral position between signs somewhat distracting and that it took away from the naturalness of the animated signing.

Our NMF additions to the mannequin were for the most-part considered to be 'good' or 'ok' but some inaccuracies caused confusion. One monitor reported not seeing any NMFs at all, while another commented on the eyebrow and mouth patterns saying that they added to their understanding. It is interesting to note that it was Monitor 1 who did not see any NMFs and Monitor 3 who appreciated them. These are the monitors noted previously in this section for giving the most negative and most positive evaluation scores, respectively. It is clear from this that the monitor's perception of the avatar is indeed linked to how they evaluate. This was confirmed by all monitors in a further question.

When asked if they found a computer-generated animation difficult to understand as opposed to a video of a human signing, all of the monitors admitted that they had some degree of difficulty. Reasons cited tended to focus on the naturalness of human signing in terms of increased facial expression. One monitor highlights the importance of this kind of natural NMF detail by noting that 70% of what is understood comes from the facial expression and only 30% from the signs themselves.[11] In addition to this, all monitors reported replaying the videos more than once when evaluating, especially the longer sentences. Another possible reason for the difficulty is that Deaf people may not be familiar with SLs articulated through an artifical means such as an animated avatar. As a result, it must be noted that the judgements of the translations could have been biased by this.

The primary errors and problems reported by the monitors were related to the

[11] This is a comment from the monitor.

pause in fluidity caused by the return to a neutral signing position, some confusion about the NMFs or incorrect signs, and one monitor finding that the avatar's arms were overly large.

Three of the four monitors said they would prefer to have information available to them in ISL rather than in English, with the fourth person saying they had no preference over either. All of the monitors agreed that technology that translates English text into ISL has a potential use in the Deaf community, particularly due to the shortage of interpreters and for private matters or issues that only require a few minutes of translation. One monitor felt it would be more useful for hearing people who are trying to learn ISL rather than the actual Deaf community itself. In addition, all monitors reported that should the system they were evaluating be extended and used in Dublin Airport, they would find it more useful than what is currently available, saying that it would provide increased access for Deaf people. Other scenarios suggested for our technology included private meetings such as doctor visits and legal matters, information points in shopping centres and on public transport. In addition, the monitors felt that system showed promise in cases where only a small amount of information is required to be translated.

6.7 Summary

In this chapter, we described ISL animation of 50 translated annotated sentences that were created using Poser software and manually evaluated by native signers.

In the first section, 6.1, we discussed the practical constraints that prohibit real-time production of human video signing, as is the preference of Deaf people. Outlining the next best option – animation of a virtual signing mannequin – we laid out four criteria that this method should meet in order to emulate the most natural, human-like signing, namely realism, consistency, functionality and fluidity.

The MT animation approaches carried out by seven of the MT projects discussed in Section 3 were outlined in Section 6.2. We compared each approach according to the criteria outlined in Section 6.1 and deduced that no previous methodology met all the criteria. Many lacked a human-like avatar, while others lacked functionality of the facial features for NMF inclusion. The addition of a fifth criterion, evaluation, demonstrated that only three of the seven systems employed manual evaluation procedures. The absence of such an assessment precludes any further comparison of animations.

In Section 6.3 we described our own SL generation model. Through an overview of our choice of software, Poser 6, we illustrated the benefits of choosing a significantly developed animation tool over creating our own by discussing the features that facilitate natural signing, e.g., a human-like 3–D model with extensive finger and facial feature manipulation facilities. We discussed our exploitation of the posing features of our chosen mannequin (Robert) in the creation of the 66 individual sign animations hand-created for this project. By referring back to the criteria previously outlined for animation development, we showed that our animation choices meet all the criteria: Robert is one of the most realistic 3–D human-like models; all lighting, camera and body positions remain consistent throughout; manipulation of the eyebrows and lips as well as additional movement ensures the avatar is functional and natural; and interpolation between the frames of each sign helps to smooth the transitions during full utterances.

Section 6.4 introduced general manual evaluation methodologies, including the trend toward dual ratings of intelligibility and fidelity using a given scale and a number of evaluators. This segued into our own experimental set-up, described in Section 6.5, where we followed the mainstream choices of Pierce et al. (1966) and van Slype (1979) and developed intelligibility and fidelity ratings using even numbered scoring scales with labelled descriptions. We employed 4 native ISL signers to perform the double evaluation using a user-friendly web-based interface. We also outlined our decision to support our evaluations with a survey questionnaire.

The results of our experiments were detailed in Section 6.6 and displayed using parallel bar-graphs and pie-charts. Intelligibility results showed that our animations were successful, with the majority, 82%, scoring in the two most positive ratings, *understood and correct* (47%) and *understood but somewhat incorrect* (35%). Fidelity results showed that our animations were successful translations of the English source sentences with 72% considered to be good–excellent translations. Highlighting the lower than average ratings of Monitor 1 and the higher than average ratings of Monitor 3, we showed the necessity of having multiple monitors for the task in order to more objectively assess the results. Correlations between these marked ratings and the opinion of the monitor of the animations in general was later made upon reviewing the questionnaire at the end of the evaluation.

By examining the correlations between intelligibility and fidelity ratings given by each monitor, we showed that for the most part the same scores were given to each sentence, but as intelligibility ratings were higher, particularly for the two highest ratings, a sentence can be considered a poor translation yet be correct and understandable ISL.

From our discussion of the survey results, we showed that the animation of our ISL translations did meet the four animation criteria. Indicating that our avatar was realistic, monitors stated that they were impressed with the dexterity of the avatars fingers and its ability to move facial features. This also served to demonstrate that the animations were functional, and displayed human-like movements and features. Furthermore, three of the monitors rated the animations a somewhat natural-to-quite natural. That the monitors found the speed of signing to be just right in most cases indicates a fluidity in the signing process.

The questionnaire also served to highlight areas of future development for our animations procedure, namely improvement of NMF capabilities and removing the unnatural and slightly distracting pause in the neutral position between signs. Given that this was our first attempt at the animation process, and allowing for the manipulation features and Python scripting possibilities of our approach, these issues can surely be improved.

Finally, our questionnaire showed that the monitors would prefer to have information available to them in their first, native language and that should our system be extended it would be more useful in an airport situation than what is currently available. This shows that despite issues of fluidity and NMF improvement, our system can be considered successful for the task it was developed to address. With careful development of our complete MT system, we are confident that a solution to the problem of Deaf–hearing communication can be addressed.

Chapter 7

Conclusions

In this thesis we have explored the area of MT for SLs and addressed a comprehensive set of issues relating to a complete bidirectional multilingual data-driven SL MT approach, namely:

- data representation, quantity and quality,
- translation of SLs to spoken language,
- translation of spoken language to SLs,
- SL video recognition,
- SL animation,
- evaluation methodologies.

Our examination of SL linguistics and corpora demonstrated that, for data-driven MT, SL videos annotated with semantic glosses provide the most appropriate data set. In addition, our experiments showed that the closed-domain and simple gloss annotation of our purpose-built ISL ATIS corpus produced better translations and evaluation scores than the open-domain, multi-tiered, detailed gloss annotations of the NGT data. This showed that the quality, as opposed to the quantity of the data, that has the biggest effect on our translations. Further experimentation with the ATIS corpus supported this, as demonstrated by our German–English translations which obtained scores comparable to experiments using data sets hundreds of times larger than the ATIS data.

Exploiting the flexible, modular design of the MaTrEx system in our experiments, we showed that data-driven MT facilitates bidirectional SL translation as well as its extensibility to new language pairs with the provision of a suitable data set. We demonstrated this through experiments in both SL-to-spoken language and spoken language-to-SL for English, German, ISL and DGS. We provided more evidence in favour of the suitability of data-driven methodologies through parallel experiments on the same data compared with the RWTH Aachen SMT system, which achieved comparable automatic evaluation scores.

Our SL-to-spoken language text experiments demonstrated that the statistical MaTrEx system is far superior to a simple EBMT system. Contrary to our initial hypothesis, namely that additional sub-sentential information would improve translation further, we showed that the addition of EBMT-style chunks to the baseline in fact decreased the translation scores for this translation direction. However, our text-to-SL experiments showed that these chunking methodologies improved scores for the reverse direction, leading us to contend that more exploration of EBMT methodologies are worthwhile. In further experiments, we showed that increasing the distortion limit to allow for jumps of up to 10 places facilitates the largely different word order between the SLs and the spoken language texts. Overall, our SL experiments in both directions obtained scores comparable with German–English spoken language scores on the same data set, which showed that data-driven MT is as capable of translating SL/spoken language as it is between spoken language.

The visual nature of SLs calls for the addition of external computational modules to the normal text-based MT system. In order to facilitate a fully functioning bidirectional translation system, these modules have an important contribution. Our experiments with the addition of automated SL video recognition from the RWTH Aachen University research group showed that the language-specific training process for recognition is currently too labour-intensive to produce recognised input of a high enough standard for use with our MT system. On the contrary, our collaboration with the MUSTER research group in University College, Dublin, demonstrated that the addition of a speech synthesis module to MT is a functional and easily executed extension to the broader MT system.

While the inclusion of a speech component is a useful addition, the development of an SL production module is somewhat more critical given that SL annotations will not suffice as MT output. In order to present a complete SL MT system, we included an additional animation module in our SL MT system. We showed that an animated signing avatar is the most practical approach to SL output but that it must be realistic, consistent, functional and fluid in order to be deemed acceptable. Our description of the animation creation process demonstrated that we addressed each of these issues successfully in our animation, unlike previous approaches. However, the neutral-posed pause between the signs generated as part of our animations affected the fluidity of our visual translations and thus the manual evaluations. This was shown in our discussion of the questionnaire, where the monitors felt that their opinion of the avatar and its fluidity affected their judgment of the translations.

Both automatic and manual evaluation metrics were used with varying degrees of success throughout our experiments. We illustrated that both metrics have their place in SL MT, with automatic metrics providing objective and speedy evaluation of MT text and annotation output, and manual evaluation providing a slower, more subjective, evaluation of both the animations and the broader capabilities of the MT system. Comparing the multi-tiered NGT annotated output with the linear ISL annotated output, we showed that automatic metrics are more adept at evaluating the linear output of either English/German text or the ISL annotated format. We also noted that as is normally the case in MT evaluation, translation quality was limited by the presence of only one reference text. Automatic evaluation scores for ATIS SL experiments in both translation directions were compared with the parallel German–English ATIS experiments. The resulting evaluation scores indicated that data-driven MT is capable of achieving comparable translation scores for SL/text translation as can be achieved for spoken language translation. This is supported by the manual evaluations we carried out on the SL animation translations. Using intelligibility and fidelity ratings, we showed that the large majority of animations were considered understandable (82%) and good translations (72%). Through examination of the evaluation scores, we showed that some monitors were biased in their scoring depending on their general opinion of the avatar and animated signing. Finally, the results of our questionnaire showed that Deaf people prefer to receive information in their native language and that should our system be implemented in an airport environment, it is likely to be considerably more useful than what is currently available. We consider such a field study to be outside the scopre of this thesis, but we very much intend to seek the deployment of our system in a pilot study in future work.

Finally, below we revisit our research questions mentioned in Chapter 1:

(RQ1) *Is it possible to employ data-driven MT methodologies to SLs with no consistent written form?*

(RQ2) *Can data-driven MT make information more accessible to the Deaf community by providing it in their first language?*

In this thesis, we have shown that through using glossed annotation versions of SLs, and even on a data set as small as 595 sentences, data-driven MT can indeed be employed. Furthermore, through our sets of bidirectional experiments and both automatic and manual evaluations, we have shown that we can produce good translations in multiple formats. Manual evaluation of our animations, together with a questionnaire, have shown that our system does make information in the airport domain more accessible to the Deaf.

7.1 Future Work

While this thesis has described a comprehensive data-driven approach to SL MT, it is by no means complete and there remain a number of avenues for future work.

As noted in Chapter 5, the EBMT chunking methodologies employed to add further sub-sentential alignments to the SMT phrasal alignments were mostly unsuccessful at improving automatic evaluation scores. We propose that further investigation is needed here to determine appropriate chunking methods for obtaining accurate SL and spoken language alignments. Provided the method produces correctly aligned chunks, it does not matter if the same or different segmentation algorithms are employed.

In Chapter 6 we outlined the importance of NMFs in the production of animated SL, this was supported by comments from our ISL monitors. Future work on the animation avatar would seek to improve the NMF details of the signing mannequin in order to produce a more realistic and natural avatar, but also to ensure correct lexical and grammatical detail. In the current system, the assistance of ISL native signers to oversee the animation creation would help in this respect. A further, and important, improvement to the animations is the development of smoother transitions between signs in sentence production. As commented by the monitors, the mannequin pausing in the neutral signing space in between each sign disrupted the flow of articulation. We intend, in future work, to remedy this disruption by developing methods to smooth between articulations. As previously noted, this is possible through the development of Python scripting. It is possible to create a set of Python templates held in an animation lexicon for compilation before signing. NMF detail and phonological features can also be included in the earlier annotation process so that these could directly feed the development of a Python script to sign the complete sentence without the need for a stored lexicon. The latter method could also help resolve deictic references and other SL linguistic phenomena. Either way, there is a manual component to each method as automatic annotation methodologies are not yet accurate enough.

There are also many possible extensions that could be made to this work described in this thesis. Ideally, the data set would be extended to include more information and help to improve translations with a view to the development of a pilot study as noted previously. Furthermore, there is scope to extend the system to other practical domains so that it could be of use to the Deaf community in places where an interpreter is impractical, e.g. medical or legal environments, as suggested by the ISL monitors. This would involve the development of new data resources in new, more practical domains, something that could also be of use the wider SL linguistics domain.

Our focus to date has been on translation between spoken and sign languages. Given developments in SL recognition and synthesis technology coupled with the ISL and DGS data sets already used in our experiments, there is scope to develop an SL-to-SL MT system. This could greatly assist in inter-communication between Deaf people who use different SLs at International SL conferences and meetings, especially where personal communication outside of the conference can mean there is no available interpreting service.

In conclusion, through experiments in SL MT employing recognition, synthesis, speech and glossing techniques, this thesis has shown the huge potential for data-driven MT to greatly assist communication between Deaf and hearing communities.

Bibliography

Armstrong, S., Groves, D., Flanagan, M., Graham, Y., Mellebeek, B., Morrissey, S., Stroppa, N., and Way, A. (2006). The MaTreX System: Machine Translation Using Examples. In *TC-STAR OpenLab Workshop on Speech Translation*, Trento, Italy. http://www.tc-star.org/openlab2006/day1/Groves_openlab.

Baker, C. and Cokely, D., editors (1980). *American Sign Language: A Teacher's Resource Text on Grammar and Culture.* Silver Spring, MD: T.J. Publishers.

Banerjee, S. and Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-05)*, pages 65–72, Ann Arbor, MI.

Bauer, B., Nießen, S., and Heinz, H. (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*, page [no page numbers], Siena, Italy.

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R., and Roossin, P. (1988). A Statistical Approach to Language Translation. In *Proceedings of the 12th International Conference on Computational Linguistics (COLING-88)*, pages 71–76, Budapest, Hungary.

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R., and

Roossin, P. (1990). A Statistical Approach to Machine Translation. *Computational Linguistics*, **16**:79–85.

Cahill, P. and Carson-Berndsen, J. (2006). The Jess Blizzard Challenge 2006 Entry. In *Proceedings of the Blizzard Challenge 2006 Workshop, Interspeech 2006*, page [no page numbers], Pittsburgh, PA.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2007). (Meta-)Evaluation of Machine Translation. In *Proceedings of 45th the Annual Meeting of the Association for Computational Linguistics (ACL-07) Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic.

Chandioux, J. (1976). MÉTÉO: Un Système Operationnel pour la Traduction Automatique des Bulletins Météreologiques Destinés au Grand Public. *META*, **21**:127–133.

Chandioux, J. (1989). Météo: 100 Million Words Later. In *Proceedings of the American Translators Association Conference 1989: Coming of Age. Learned Information*, pages 449–453, Medford, NJ.

Chomsky, N. (2000). *New Horizons in the Study of Language and Mind*. Cambridge University Press, Cambridge, MA.

Clark Gunsauls, D. (1998). *Goldilocks & The Three Bears: Basic Storybook*. The Center for Sutton Movement Writing, Inc., La Jolla, CA.

Conroy, P. (2006). Signing In & Signing Out: The Education and Employment Experiences of Deaf Adults in Ireland. Research report, Irish Deaf Society, Dublin.

Conway, A. and Veale, T. (1994). A Linguistic Approach to Sign Language Synthesis. In *G. Cockton, S.W. Draper and G.R.S. Weir (Eds.) People and Computers IX: Proceedings of the Human Computer Interface Conference (HCI)*, pages 211–222, Glasgow, Scotland.

Cunningham, P. and Veale, T. (1991). Organizational Issues Arising from the Integration of the Concept Network & Lexicon in a Text Understanding System. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence*, pages 986–991, San Mateo, CA.

Dreuw, P., Rybach, D., Deselaers, T., Zahedi, M., and Ney, H. (2007). Speech Recognition Techniques for a Sign Language Recognition System. In *Proceedings of Interspeech 2006*, pages 2513–2516, Antwerp, Belgium.

Elliott, R., Glauert, J. R. W., Kennaway, J. R., and Marshall, I. (2000). The Development of Language Processing Support for the ViSiCAST Project. In *Proceedings of Assets 00: Fourth International ACM Conference on Assistive Technologies*, pages 101–108, New York, NY.

Emmorey, K., Grabowski, T., McCullough, S., Ponto, L., Hichwa, R., and Damasio, H. (2005). The Neural Correlates of Spatial Language in English and American Sign Language: A PET Study with Hearing Bilinguals. *NeuroImage*, **24**(3):832–840.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Fourie, J. (2006). The Design of a Generic Signing Avatar. Technical report, University of Stellenbosch, South Africa.

Gamma, E., Helm, R., Johnson, R., and Vlissides, J. (1995). *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley.

Gough, N. and Way, A. (2004a). Example-Based Controlled Translation. In *Proceedings of the Ninth EAMT Workshop*, pages 73–81, Valetta, Malta.

Gough, N. and Way, A. (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)*, pages 95–104, Baltimore, MD.

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 *Proceedings of the Second High Desert Student Conference in Linguistics (HDSL2)*, pages 13–30, Albuquerque, NM.

Grieve-Smith, A. B. (2001). SignSynth: A Sign Language Synthesis Application Using Web3D and Perl. In *Revised Papers from the International Gesture Workshop on Gesture and Sign Languages in Human-Computer Interaction*, pages 134–145, London, UK.

Groves, D. and Way, A. (2005a). Hybrid Data-Driven Models of Machine Translation. *Machine Translation: Special Issue on Example-Based Machine Translation*, **19**(3-4):301–322.

Groves, D. and Way, A. (2005b). Hybrid Example-Based SMT: the Best of Both Worlds? In *Proceedings of the Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond, at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-05)*, pages 183–190, Ann Arbor, MI.

Hanke, T. (2004). HamNoSys – Representing Sign Language Data in Language Resources and Language Processing Contexts. In *Workshop on the Representation and Processing of Sign Languages at LREC 04*, pages 1–6, Lisbon, Portugal.

Hassan, H., Ma, Y., and Way, A. (2007). MaTrEx: the DCU Machine Translation System for IWSLT 2007. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 69–75, Trento, Italy.

Hemphill, C., Godfrey, J., and Doddington, G. (1990). The ATIS Spoken Language Systems Pilot Corpus. In *Proceedings of the Workshop on Speech and Natural Language*, pages 96–101, Hidden Valley, PA.

Huenerfauth, M. (2004). Spatial and Planning Models of ASL Classifier Predicates for Machine Translation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-04)*, page [no page numbers], Baltimore, MD.

Huenerfauth, M. (2005). American Sign Language Generation: Multimodal NLG with Multiple Linguistic Channels. In *Proceedings of the ACL Student Research Workshop at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-05)*, pages 37–42, Ann Arbor, MI.

Huenerfauth, M. (2006). *Generating American Sign Language Classifier Predicates for English-to-ASL Machine Translation.* PhD thesis, University of Pennsylvania, Philadelphia, PA.

Hutchins, W. and Somers, H. (1992). *An Introduction to Machine Translation.* Academic Press Limited, London, UK.

Jerde, T. E., Soechting, J. F., and Flanders, M. (2003). Coarticulation in Fluent Fingerspelling. *Journal of Neuroscience*, **23**(6):2383–2393.

Kamp, H. (1981). A Theory of Truth and Semantic Representation. In *J. Groenendijk, Th. Janssen, and M. Stokhof, (Eds.), Formal Methods in the Study of Language*, Mathematisch Centrum Tracts, Amsterdam.

Kanthak, S., Vilar, D., Matusov, E., Zens, R., and Ney, H. (2005). Novel Reordering Approaches in Phrase-Based Statistical Machine Translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-05)*, pages 167–174, Ann Arbor, MI.

Kaplan, R. M. and Bresnan, J., editors (1982). *The Mental Representation of Grammatical Relations*, pages 173–281. The MIT Press, Cambridge, MA.

Karypis, G. (2001). Evaluation of Item-Based Top-N Recommendation Algorithms. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, pages 247–254, Phuket, Thailand.

Kneser, R. and Ney, H. (1995). Improved Backing-Off for n-gram Language Modelling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Volume 1*, pages 181–184, Detroit, MI.

Koehn, P. (2004). Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. Machine Translation: From Real Users to Research. In *Proceedings of the 6th Conference on the Association for Machine Translation in the Americas (AMTA-04)*, pages 115–124, Washington, DC.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: open source toolkit for statistical machine translation. In *Proceedings of Demonstration and Poster Sessions at the 45th Annual Meeting of the Association of Computational Linguistics (ACL-07)*, Prague, Czech Republic.

Koehn, P., Och, F., and Marcu, D. (2003). Statistical Phrase-Based Translation. In *Proceedings of the Combined Human Language Technology Conference Series and the North American Chapter of the Association for Computational Linguistics Conference Series (HLT-NAACL)*, pages 48–54, Edmonton, Canada.

LDC (2005). Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Translations. Revision 1.5.

Le Master, B. (1990). *The Maintenance and Loss of Female and Male Signs in the Dublin Deaf Community*. PhD thesis, University of California, Los Angeles, CA.

Leeson, L. (2001). *Aspects of Verbal Valency in Irish Sign Language*. PhD thesis, University of Dublin, Trinity College.

Leeson, L. (2003). *M. Cronin and C. Ó Cuilleanáin(Eds.), Languages of Ireland*, chapter 8, pages 148–164. Four Courts Press: Dublin, Ireland.

Leeson, L., Saeed, J., Macduff, A., Byrne-Dunne, D., and Leonard, C. (2006). Moving Heads and Moving Hands: Developing a Digital Corpus of Irish Sign Language. In *Proceedings of Information Technology and Telecommunications Conference 2006*, page [no page numbers], Carlow, Ireland.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady*, **10**:707–710.

Liddell, S. (1980). *American Sign Language Syntax*. Mouton: The Hague.

Liddell, S. and Johnson, R. E. (1989). American Sign Language: The Phonological Base. *Sign Language Studies*, **64**:195–277.

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

Marshall, I. and Sáfár, E. (2003). A Prototype Text to British Sign Language (BSL) Translation System. In *Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics (ACL-03) Conference*, pages 113–116, Sapporo, Japan.

Matthews, P. A. (1996). *The Irish Deaf Community, Survey Report, History of Education, Language and Culture*. The Linguistics Institute of Ireland, Dublin, Ireland.

Matusov, E., Zens, R., Vilar, D., Mauser, A., Popovic, M., and Ney, H. (2006). The RWTH Machine Translation System. In *Proceedings of the TC-STAR Workshop on Speech-to-Speech Translation*, pages 31–36, Barcelona, Spain.

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

Morrissey, S. and Way, A. (2006). Lost in Translation: the Problems of Using Mainstream MT Evaluation Metrics for Sign Language Translation. In *Proceedings of the 5th SALTMIL Workshop on Minority Languages at LREC 2006*, pages 91–98, Genoa, Italy.

Morrissey, S. and Way, A. (2007). Joining Hands: Developing a Sign Language Machine Translation System with and for the Deaf Community. In *Proceedings of the Conference and Workshop on Assistive Technologies for People with Vision & Hearing Impairments (CVHI)*, page [no page numbers], Granada, Spain.

Morrissey, S., Way, A., Cahill, P., and Carson-Berndsen, J. (2007a). A Complete Irish Sign Language to Speech Translation System. In *IBM CASCON Dublin Symposium*, page [no page numbers], Dublin, Ireland.

Morrissey, S., Way, A., Stein, D., Bungeroth, J., and Ney, H. (2007b). Combining Data-Driven MT Systems for Improved Sign Language Translation. In *Proceedings of Machine Translation Summit XI*, pages 329–336, Copenhagen, Denmark.

Naqvi, S. (2007). End-User Involvement in Assistive Technology Design for the Deaf Are Artificial Forms of Sign Language Meeting the Needs of the Target Audience? In *Proceedings of the Conference and Workshop on Assistive Technologies for People with Vision & Hearing Impairments (CVHI)*, page [no page numbers], Granada, Spain.

Neidle, C. (2002). SignStream^{TM}Annotation: Conventions used for the American Sign Language Linguistic Research Project. http://www.bu.edu/asllrp/asllrpr11.pdf.

Neidle, C., Sclaroff, S., and Athitsos, V. (2001). SignStream^{TM}: A Tool for Linguistic and Computer Vision Research on Visual-Gestural Language Data. *Behavior Research Methods, Instruments, and Computers*, **33**(3):311–320.

Newkirk, D. (1986). Outline of a Proposed Orthography of American Sign Language. http://web.archive.org/web/20011031221204/members.home.net/dnewkirk/signfont/orthog.

Nonhebel, A., Crasborn, O., and van der Kooij, E. (2004). *Sign language transcription conventions for the ECHO Project*. University of Nijmegen, version 9 edition. http://www.let.kun.nl/sign-lang/echo/docs/transcr_conv.pdf.

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

Och, F. (2003). Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computation Linguistics (ACL-03)*, pages 160–167, Sapporo, Japan.

Och, F. J. and Ney, H. (2000). Improved Statistical Alignment Models. In *Proceedings of the 38th Annual Meeting of the Association for Computation Linguistics (ACL-00)*, pages 440–447, Hong Kong, China.

Och, F. J. and Ney, H. (2002). Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computation Linguistics (ACL-02)*, pages 295–302, Philadelphia, PA.

O'Grady, W., Dobrovolsky, M., and Katamba, F., editors (1999). *Contemporary Linguistics: An Introduction*. London: Addison Wesley Longman Ltd.

Owczarzak, K. (2008). *A Novel Dependency-Based Evaluation Metric for Machine Translation*. PhD thesis, Dublin City University, Dublin, Ireland.

Owczarzak, K., van Genabith, J., and Way, A. (2007). Labelled Dependencies in Machine Translation Evaluation. In *Proceedings of the 2nd Workshop on Sta-*

tistical Machine Translation at the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07), pages 104–111, Prague, Czech Republic.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (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)*, pages 311–318, Philadelphia, PA.

Paul, M. (2006). Overview of the IWSLT 2006 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 1–15, Kyoto, Japan.

Peckham, J. (1993). A new generation of spoken dialogue systems: Results and lessons from the SUNDIAL project. In *Proceedings of Eurospeech '93*, pages 33–40, Berlin, Germany.

Pierce, J., Carroll, J., Hamp, E., Hays, D., Hockett, C., Oettinger, A., and Perlis, A. (1966). Language and Machines: Computers in Translation and Linguistics. Technical Report: Automatic Language Processing Committee, National Academy of Sciences, National Research Council, Washington, DC.

Pinker, S. (1994). *The Language Instinct: How the Mind Creates Language*. Penguin, London, UK.

Pollard, C. and Sag, I. (1994). *Head Phrase Structure Grammar*. University of Chicago Press, Chicago, IL.

Rabiner, L. R. and Juang, B. H. (1989). An Introduction to Hidden Markov Models. *The IEEE ASSP Magazine*, **4**(1):4–16.

Sáfár, E. and Marshall, I. (2002). The Architecture of an English-Text-to-Sign-Languages Translation System. In *Proceedings of the2nd International Conference on Recent Advances in Natural Language Processing (RANLP-02)*, pages 223–228, Tzigov Chark, Bulgaria.

San Segundo, R., Montero, J. M., Macias-Guarasa, J., Córdoba, R., Ferreiros, J., and Pardo, J. M. (2007). Proposing a Speech to Gesture Translation Architecture for Spanish Deaf People. *Visual Language and Computing.*

Shieber, S. (1994). Restricting the Weak-Generative Capability of Synchronous Tree Adjoining Grammars. *Computational Intelligence*, **10**(4):[no page numbers].

Sleator, D. and Temperley, D. (1991). Parsing English with a Link Grammar. Carnegie Mellon University Computer Science Technical Report: CMU-CS-91-196, Carnegie Mellon University, Pittsburgh, PA.

Smith, C., Lentz, E., and Mikos, K., editors (1988). *Signing Naturally, Teacher's Curriculum Guide, Level 1.* Berkeley, CA: DawnSignPress.

Stein, D., Bungeroth, J., and Ney, H. (2006). The Architecture of an English Text-to-Sign Languages Translation System. In *Proceedings of the 11th Annual Conference of the European Association for Machine Translation (EAMT, '06)*, pages 169–177, Oslo, Norway.

Stein, D., Dreuw, P., Ney, H., Morrissey, S., and Way, A. (2007). Hand in Hand: Automatic Sign Language to English Translation. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07)*, pages 214–220, Skövde, Sweden.

Stokoe, W. C. (1960). *Sign language structure: an outline of the visual communication system of the American deaf.* Studies in Linguistics, Occasional Paper, 2nd printing 1993: Linstok Press, Burtonsville, MD.

Stokoe, W. C. (1972). *Semiotics and Human Sign Languages.* Mouton & Co. N.V., Publishers, The Hague, The Netherlands.

Stroppa, N. and Way, A. (2006). MaTrEx: DCU Machine Translation System for IWSLT 2006. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 31–36, Kyoto, Japan.

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

Traxler, C. (2000). The Stanford Achievement Test, 9th Edition: National Norming and Performance Standards for Deaf and Hard–of–Hearing Students. *Journal of Deaf Studies and Deaf Education*, **5**(4):337–348.

Trujillo, A. (1999). *Translation Engines: Techniques for Machine Translation.* Springer-Verlag: Series on Applied Computing, London, UK.

Turian, J. P., Shen, L., and Melamed, I. (2003). Evaluation of Machine Translation and its Evaluation. In *Proceedings of the Machine Translation Summit IX*, page [no page numbers], New Orleans, LA.

van Slype, G. (1979). Critical Methods for Evaluating the Quality of Machine Translation. Technical Report: BR-19142, European Commission Directorate General Scientific and Technical Information and Information Management, Bureau Marcel van Dijk.

van Zijl, L. and Combrink, A. (2006). The South African Sign Language Machine Translation Project: Issues on Non-manual Sign Generation. In *Proceedings of South African Institute for Computer Scientists and Information Technologists Conference (SAICSIT 06)*, pages 127–134, Cape Winelands, South Africa.

van Zijl, L. and Fourie, J. (2007). Design and Development of a Generic Signing Avatar. In *Proceedings of Graphics and Visualization in Engineering*, pages 95–100, Clearwater, FL.

van Zijl, L. and Olivrin, G. (2008). South African Sign Language Assistive Translation. In *Proceedings of IASTED International Conference on Assistive Technologies*, page [no page numbers], Baltimore, MD.

Veale, T., Conway, A., and Collins, B. (1998). The Challenges of Cross-Modal

Translation: English to Sign Language Translation in the Zardoz System. *Machine Translation*, **13**(1):81–106.

Veale, T. and Cunningham, P. (1992). Competitive Hypothesis Resolution in TWIG: A Blackboard-Driven Text-Understanding System. In *Proceedings of the 10th European Conference on Artificial Intelligence*, pages 561–563, Chichester, UK.

Veale, T. and Way, A. (1997). GAIJIN: A Bootstrapping Approach to Example-Based Machine Translation. In *Proceedings of 2nd International Conference on Recent Advances in Natural Language Processing*, pages 239–244, Tzigov Chark, Bulgaria.

Viterbi, A. (1967). Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. **13**(2):260–269.

Vogel, S., Zhang, Y., Huang, F., Tribble, A., Venugopal, A., Zhao, B., and Palmer, M. (2003). The CMU Statistical Machine Translation System. In *Proceedings of the MT Summit IX*, pages 402–409, New Orleans, LA.

Volterra, V., Laudanna, A., Corazza, S., Radutsky, E., and Natale, F. (1984). Italian sign language: The order of elements in the declarative sentence. In *F. Loncke, P. Boyes-Braem, and Y. Lebrun, (eds.): Recent Research on European Sign Language*, pages 19–48, Lisse: Swets and Zeitlinger.

Ward, W. (1991). Understanding spontaneous speech: The phoenix system. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 365–367, Toronto, Canada.

Way, A. and Gough, N. (2003). Developing and Validating an EBMT System using the World Wide Web. *Computational Linguistics*, **29**(3):421–457.

Way, A. and Gough, N. (2005). Comparing Example-Based and Statistical Machine Translation. *Natural Language Engineering*, **11**(3):295–309.

Weaver, W. (1949). *Recent Contributions to the Mathematical Theory of Communication. In Shannon, C.E. and Weaver, W., (Eds.), The Mathematical Theory of Communication*, pages 94–117. The University of Illinois Press, Urbana, IL.

White, J., O'Connell, T., and O'Mara, F. (1994). The ARPA MT Evaluation Methodologies: Evolution, Lessons and Further Approaches. In *Proceedings of the First Conference of the Association for Machine Translation of the Americas (AMTA-94)*, pages 193–205, Colombia, MD.

Wu, C.-H., Su, H.-Y., Chiu, Y.-H., and Lin, C.-H. (2007). Transfer-Based Statistical Translation of Taiwanese Sign Language Using PCFG. *ACM Transactions on Asian Language Information Processing (TALIP)*, **6**(1):[no page numbers].

Zhao, L., Kipper, K., Schuler, W., Vogler, C., Badler, N., and Palmer, M. (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)*, pages 293–300, Cuernavaca, Mexico.