

## University of Southampton Research Repository

Copyright © and Moral Rights for this thesis and, where applicable, any accompanying data are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis and the accompanying data cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content of the thesis and accompanying research data (where applicable) must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holder/s.

When referring to this thesis and any accompanying data, full bibliographic details must be given, e.g.

Thesis: Author (Year of Submission) "Full thesis title", University of Southampton, name of the University Faculty or School or Department, PhD Thesis, pagination.

Data: Author (Year) Title. URI [dataset]



University of Southampton

Faculty of Engineering and Physical Sciences  
School of Electronics and Computer Science

A Step Towards Machine Translation Between  
Communication Symbols and Arabic Text

by

Lama Abdullah Alzaben

ORCID: [0000-0001-9521-8093](https://orcid.org/0000-0001-9521-8093)

A thesis for the degree of  
Doctor of Philosophy

November 29, 2021



University of Southampton

Abstract

Faculty of Engineering and Physical Sciences  
School of Electronics and Computer Science

Doctor of Philosophy

A Step Towards Machine Translation Between Communication Symbols and Arabic  
Text

by Lama Abdullah Alzaben

Pictographic symbols may be used as an alternative form of communication by individuals with complex communication needs. These symbols are a collocation of drawings that depict concepts often associated with glosses that express the meaning in spoken language. This research investigates the problem by focusing on one particular language, Modern Standard Arabic, and one set of symbols, the ARASAAC set. The outcomes are generalisable to other symbols sets and spoken languages.

Symbols can be used as part of an electronic speech generating devices. Users may use symbols to express messages or to understand received messages. Thus, translating to and from text is an important task that increases communication with a wider community. Translating text to symbols requires awareness of the exact sense of the textual units that are part of the input message, to determine relevant symbols. Translating from symbol to text, on the other hand, requires in addition to associated glosses, an awareness of the grammar, and word sequence likelihoods, to be able generate fully-formed sentences. Machine translation has often been tackled by using methods that require large amounts of data. This data needs to match the domain and cover the same source and target registers that are expected by a translation system. However, a parallel corpus of pictographic symbols and MSA is currently unavailable. This research addresses this issue by proposing an approach that creates the training data needed by making use of existing multilingual textual resources to resolve ambiguity.

The outcome was a corpus that had been automatically tagged with morphological annotations and pictographic symbols, the approach followed, and an investigation of the data involved and produced. The availability of this symbol tagged corpus is a step towards Arabic symbol/text translation. This has the potential to enhance communication for those requiring Arabic speech output from symbol messaging, and provide a better understanding of the complexities of automated symbol/text translation processes in Arabic.



# Contents

|  |      |
|--|------|
| List of Figures  | ix   |
| List of Tables   | xi   |
| Declaration of Authorship                                | xiii |
| Acknowledgements   | xv   |
| 1 Introduction   | 1    |
| 1.1 Contributions . . . . .                              | 4    |
| 1.2 Thesis Outline . . . . .                             | 7    |
| 2 Literature Review and Background                       | 9    |
| 2.1 Overview . . . . .                                   | 9    |
| 2.2 Introduction . . . . .                               | 9    |
| 2.3 AAC Symbols and Text Processing . . . . .            | 11   |
| 2.3.1 Symbol Sets . . . . .                              | 12   |
| 2.3.2 Managing Associated Glosses . . . . .              | 15   |
| 2.3.3 Generating Symbols . . . . .                       | 17   |
| 2.3.4 Generating Text . . . . .                          | 18   |
| 2.3.5 Evaluation . . . . .                               | 20   |
| 2.4 Arabic as a Target Language . . . . .                | 22   |
| 2.4.1 MSA with Symbols . . . . .                         | 23   |
| 2.5 Related Work . . . . .                               | 24   |
| 2.5.1 Word Sense Disambiguation . . . . .                | 25   |
| 2.5.2 Part of Speech Tagging and Lemmatisation . . . . . | 27   |
| 2.5.3 Language Models and Machine Translation . . . . .  | 30   |
| 2.5.4 Computer Vision . . . . .                          | 34   |
| 2.5.4.1 Colour . . . . .                                 | 34   |
| 2.5.5 Local Descriptors . . . . .                        | 35   |
| 2.5.5.1 Global Descriptor . . . . .                      | 36   |
| 2.6 Data . . . . .                                       | 37   |
| 2.6.1 Corpus . . . . .                                   | 37   |
| 2.6.2 Training Data . . . . .                            | 39   |
| 2.7 Conclusion . . . . .                                 | 41   |
| 2.8 This Thesis . . . . .                                | 42   |
| 3 Understanding Graphical Symbols for Communication      | 43   |

|         |  |     |
|---------|--|-----|
| 3.1     | ARASAAC . . . . .  | 44  |
| 3.1.1   | Graphical Content . . . . .                                      | 46  |
| 3.1.2   | Associated Glosses . . . . .                                     | 46  |
| 3.1.3   | Parts of Speech or Word Classes . . . . .                        | 47  |
| 3.1.4   | Synonymous Glosses . . . . .                                     | 48  |
| 3.1.5   | Gloss Morphology . . . . .                                       | 51  |
| 3.1.6   | Ambiguous Glosses . . . . .                                      | 52  |
| 3.1.7   | Gloss Translations . . . . .                                     | 53  |
| 3.1.8   | Language Coverage . . . . .                                      | 53  |
| 3.2     | Symbol Content Similarity . . . . .                              | 55  |
| 3.3     | Symbol Markers . . . . .   | 55  |
| 3.3.1   | Textual Content . . . . .  | 57  |
| 3.3.2   | Cultural Differences in Visual Representation . . . . .          | 57  |
| 3.4     | Conclusion . . . . .   | 58  |
| 4       | Methodology . . . . .  | 61  |
| 4.1     | Translation . . . . .  | 62  |
| 4.1.1   | Symbol to Text . . . . .   | 62  |
| 4.1.2   | Text to Symbols . . . . .  | 64  |
| 4.2     | The Pictographic Symbol Component . . . . .                      | 67  |
| 4.3     | Corpus . . . . .   | 69  |
| 4.3.1   | Preprocessing . . . . .  | 70  |
| 4.3.2   | Domain Relevance . . . . .                                       | 71  |
| 4.4     | Generating Training Data . . . . .                               | 72  |
| 4.4.1   | Corpus Lemmatization and Vocalisation . . . . .                  | 73  |
| 4.4.1.1 | Out of Context Lemma Disambiguation . . . . .                    | 74  |
| 4.4.1.2 | In Context Morphological Disambiguation . . . . .                | 79  |
| 4.4.1.3 | Adapting a Morphology Analyser . . . . .                         | 80  |
| 4.4.2   | Telegraphic Text and Fully-Formed Text Parallel Corpus . . . . . | 82  |
| 4.4.3   | Symbol Tagging . . . . .   | 83  |
| 4.5     | Visual Content of Symbols . . . . .                              | 86  |
| 4.6     | Evaluation . . . . .   | 88  |
| 4.7     | Summary . . . . .  | 89  |
| 5       | Results . . . . .  | 91  |
| 5.1     | Corpora and Extracted Data . . . . .                             | 91  |
| 5.2     | Text Pre-Processing . . . . .                                    | 93  |
| 5.3     | Symbol to Text Experiments . . . . .                             | 94  |
| 5.4     | Symbols to Text Experiments . . . . .                            | 95  |
| 5.5     | The Visual Content . . . . .                                     | 98  |
| 5.6     | Summary . . . . .  | 101 |
| 6       | Discussion . . . . .   | 103 |
| 6.1     | Corpus . . . . .   | 103 |
| 6.2     | Lemmatisation and Morphological Analysis . . . . .               | 106 |
| 6.3     | Symbol to Text Translation . . . . .                             | 108 |
| 6.4     | Symbol Tagging . . . . .   | 109 |



---

|   |                                       |     |
|---|---------------------------------------|-----|
| 6.5   | Visual Content of Symbols . . . . .   | 112 |
| 6.6   | Text to Symbols Translation . . . . . | 114 |
| 6.7   | Summary . . . . .                     | 115 |
| 7   | Conclusion and Future Work . . . . .  | 117 |
| 7.1   | Conclusions . . . . .                 | 117 |
| 7.2   | Limitations . . . . .                 | 121 |
| 7.3   | Future Work . . . . .                 | 122 |
| 7.4   | Summary . . . . .                     | 123 |
| Appendix A Message List . . . . .                                     |                                       | 125 |
| Appendix B Examples Showing Arabic Morphological Variations . . . . . |                                       | 129 |
| Appendix C SIFT ARASAAC Examples . . . . .                            |                                       | 133 |
| Appendix D HOG ARASAAC Examples . . . . .                             |                                       | 135 |



# List of Figures

|     |   |     |
|-----|---|-----|
| 1.1 | A symbol linked to the word ‘trainer’ found at <a href="https://www.widgit.com/-training/index.htm">https://www.widgit.com/-training/index.htm</a> . . . . .  | 4   |
| 2.1 | Bliss example showing that size changes the meaning of a symbol . . . . .   | 13  |
| 2.2 | Bliss sentence from the Swedish National Agency for Special Needs Education and Schools (SPSM) (2010) <a href="https://www.blissymbolics.org/images/Bliss_English_SPSM_10964.pdf">https://www.blissymbolics.org/images/Bliss_English_SPSM_10964.pdf</a> . . . . . | 13  |
| 2.3 | Makaton sentence example <a href="https://www.makaton.org/shop/examples_computer_use">https://www.makaton.org/shop/examples_computer_use</a> . . . . .  | 14  |
| 2.4 | Example of a symbol board using ARASAAC symbols . . . . .   | 15  |
| 3.1 | The number of ARASAAC symbols developed by year . . . . .   | 45  |
| 4.1 | The process of creating training data for the symbol to text task . . . . .   | 64  |
| 4.2 | The process of automatically tagging the corpus with symbols . . . . .  | 67  |
| 4.3 | Example of directional word alignment for the following sentences: English : - you think it’s a good idea ? Arabic : - هل تظنها فكرة جيدة ؟ . . . . .   | 75  |
| 4.4 | Ranking lemma candidates . . . . .  | 78  |
| 4.5 | Disambiguation process for symbol tagging . . . . .   | 83  |
| 4.6 | Extracting translation equivalent set . . . . .   | 85  |
| 6.1 | An ARSAAC symbol with the gloss ‘clean the glasses’ . . . . .   | 111 |



# List of Tables

|      |   |    |
|------|---|----|
| 2.1  | Bliss-characters versus Bliss-words . . . . .   | 13 |
| 3.1  | Commonly used symbols (ISO 7010, 2019) . . . . .  | 44 |
| 3.2  | A sample of ARASAAC symbols with their associated glosses . . . . .   | 45 |
| 3.3  | Examples of symbols representing advanced concepts . . . . .  | 45 |
| 3.4  | Symbols in different classes: an icon; an index; a symbol; another symbol   | 46 |
| 3.5  | The number of glosses associated with each symbol . . . . .   | 47 |
| 3.6  | Symbols with more than one gloss . . . . .  | 47 |
| 3.7  | A random sample of symbols for each word type . . . . .   | 49 |
| 3.8  | The number of gloss and symbol pairs for each word type . . . . .   | 49 |
| 3.9  | Four different symbols for the concept “learn to swim” . . . . .  | 49 |
| 3.10 | Words having multiple representations (e.g. open) . . . . .   | 50 |
| 3.11 | Some different human referent representations for the concept ‘teacher’ .   | 50 |
| 3.12 | Example of a concept represented only with stick figures . . . . .  | 50 |
| 3.13 | Symbols that are very specific on the left, and that are more general on<br>the right . . . . .   | 51 |
| 3.14 | The handling of plurals . . . . .   | 52 |
| 3.15 | An example of symbols associated with irrelevant glosses . . . . .  | 52 |
| 3.16 | Symbols associated with glosses in multiple languages . . . . .   | 53 |
| 3.17 | Example of symbols associated with irrelevant glosses . . . . .   | 53 |
| 3.18 | Words not covered by ARASAAC glosses . . . . .  | 54 |
| 3.19 | Variation in lexicalised concepts between languages . . . . .   | 55 |
| 3.20 | Visual similarity between relevant concepts . . . . .   | 56 |
| 3.21 | Symbols having a special element or qualifier . . . . .   | 56 |
| 3.22 | Symbols showing inconsistencies . . . . .   | 56 |
| 3.23 | Example of symbols containing text . . . . .  | 57 |
| 3.24 | Symbols containing human figures . . . . .  | 57 |
| 3.25 | Some of the few symbols that show variation in colour . . . . .   | 58 |
| 3.26 | The concept ambulance informed by communities having different medical<br>symbols due to different religious backgrounds . . . . .                    | 58 |
| 4.1  | Language model evaluation results . . . . .   | 72 |
| 4.2  | The number of Arabic surface forms compared with English Arabic forms<br>of a single root and appearing at least five times in the subtitles corpus . | 73 |
| 4.3  | Several vocalised forms have the same normalised forms . . . . .  | 78 |
| 5.1  | Arabic English Corpus Statistics . . . . .  | 92 |
| 5.2  | Arabic Spanish Corpus Statistics . . . . .  | 92 |
| 5.3  | Top Arabic translations of the English word book . . . . .  | 92 |

|              |  |     |
|--------------|--|-----|
| 5.4          | Top English translations of the Arabic word كتاب . . . . .   | 93  |
| 5.5          | Top Spanish translations of the Arabic word كتاب . . . . .   | 93  |
| 5.6          | Top Arabic translations of the Spanish word libro . . . . .  | 93  |
| 5.7          | The number of extracted Pairs with frequency $\geq 50$ and PMI score $\geq 5$ from each alignment . . . . .                                  | 93  |
| 5.8          | Parts of Speech in the Arabic dictionary . . . . .   | 94  |
| 5.9          | The linguistic annotation for an Arabic sentence that was a translation of “do you ever think about the future ?” . . . . .                  | 94  |
| 5.10         | Symbol to text parallel statistics . . . . .   | 94  |
| 5.11         | Fully-formed text and its corresponding telegraphic form . . . . .   | 95  |
| 5.12         | BLEU score for unseen test segment . . . . .   | 95  |
| 5.13         | . . . . .  | 97  |
| 5.14         | . . . . .  | 97  |
| 5.15         | . . . . .  | 97  |
| 5.16         | SIFT example . . . . .   | 99  |
| 5.17         | HOG example . . . . .  | 99  |
| 5.18         | Example showing HOG vs. SIFT . . . . .   | 100 |
| 5.19         | Visually similar symbols Example shows the potential of using visual content in acquiring additional data (c) . . . . .                      | 100 |
| 5.20         | Percentage of symbols with at least one similar symbol . . . . .   | 100 |
| 5.21         | ARASAAC markers and number identified . . . . .  | 101 |
| 6.1          | The null subject problem and agglutination showing the difference between Arabic and English, using the English Penn Treebank tagset . . . . | 109 |
| 6.2          | Top 50 nouns missing from the symbol set in order of frequency . . . . .   | 110 |
| 6.3          | Top 50 verbs missing from the symbol set in order of frequency . . . . .   | 110 |
| 6.4          | All symbols in the set having the substring ‘story’, showing the lack of a symbol for ‘story’ . . . . .                                      | 111 |
| 6.5          | Several symbols for the gloss ‘telephone’, showing no strong similarity . .  | 112 |
| Appendix C.1 | . . . . .  | 133 |
| Appendix C.2 | . . . . .  | 133 |
| Appendix C.3 | . . . . .  | 134 |
| Appendix C.4 | . . . . .  | 134 |
| Appendix C.5 | . . . . .  | 134 |
| Appendix D.1 | . . . . .  | 135 |
| Appendix D.2 | . . . . .  | 135 |
| Appendix D.3 | . . . . .  | 136 |
| Appendix D.4 | . . . . .  | 136 |
| Appendix D.5 | . . . . .  | 136 |

## Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission

Signed:.....

Date:.....





## Acknowledgements

I would like to express my sincere gratitude to Mrs EA Draffan for her patience, guidance and encouragement. I would also like to thank Prof Mike Wald for his support. I am also grateful for the time and support my family have given me throughout this period in my life and to the University of Southampton for their understanding during the difficult times that have arisen recently. I would also like to thank ARASAAC for allowing me to use their symbol set in this thesis and providing access to their application programming interface.



To family and friends



# Chapter 1

## Introduction

The ability to communicate with those around us is an essential aspect of our lives. People communicate for different purposes; to express their needs and desires, to exchange knowledge, or to socialise (Light, 1988), and this is considered a Human Right under the United Nations Convention on the Rights of Persons with Disabilities (UNCRPD) (United Nations, 2006). Forms of augmentative and alternative communication (AAC) may be introduced to address the inability to use or understand spoken language to support daily communication needs. AAC is a range of methods that includes “gestures, manual signs, picture- or symbol-based communication systems, and computer-based speech-generating devices” (Sigafoos, 2010). The aim of this research is to assist automatic translation between graphical symbols used for communication and Arabic text. This is important for individuals with expressive and/or receptive language disabilities, who rely on symbols to communicate with those around them, and who could use AAC speech-generating devices with automatic translation to and from Arabic speech.

Typical face-to-face communication is achieved using both verbal and non-verbal actions such as words, facial expressions, gestures and other forms of body language. Having a common language between two communicating parties is crucial to achieve the goal of communication. A typical infant at 18 months will have a vocabulary of 50 words (Beard, 2018) and university students were found to speak an estimated 16,000 different words per day (Mehl et al., 2007). However, some people face difficulty when expressing or comprehending verbal messages. These may be as a result of a wide range of physical, sensory and/or cognitive disabilities, such as cerebral palsy, stroke, autism and intellectual impairment (Beukelman & Mirenda, 2013).

Graphical symbols have been used to overcome natural language problems and are a communication tool as part of an AAC intervention. These symbols are designed to communicate meaning through drawings or photographs which are independent of the

conventional orthography of spoken languages. An AAC symbol<sup>1</sup> represents a concept, object, occasion, activity or setting and can be printed on cards or embedded in electronic devices. Dada et al. (2013) stated that “graphic symbols form a very important component of most aided AAC systems”. Several symbol developments exist across the world. Differences between symbol sets are mainly seen in the relationship between depiction and referent, the design of the visual aspects of the symbols, the linguistic schema developed to support semantics and syntax, and localisation attributes to fit the country of origin. This research made use of ARASAAC symbol set<sup>2</sup>.

A single graphical symbol usually conveys a single concept in isolation (Light & McNaughton, 2012) which loosely corresponds to a lexeme in spoken languages. They are often individually labelled with their corresponding gloss in the local spoken language, sometimes with further translations into secondary languages. The labels serve many purposes, such as to generate speech in Voice Output Communication Aids (VOCAs) or allowing people who are unfamiliar with symbols to understand the communicated message. A message can combine several symbols, each aligned with its gloss in their base form, which does not necessarily result in a grammatically correct phrase or sentence. In such a case, the set of associated glosses needs to be processed further to ensure a plausible textual output can be spoken through the speech synthesizer. Several researchers have proposed methods for achieving this aim (Chang et al., 1993; Karberis & Kouroupetroglou, 2002; McCoy et al., 1990; Viglas & Kouroupetroglou, 2002; Waller & Jack, 2002; Wiegand & Patel, 2012b).

Symbol use is not limited to expressive communication but can also be used in receptive communication (Stephenson & Linfoot, 1996). Thus, there is a need to translate/augment a textual sentence to/with symbols. Augmenting text with symbols has been shown to increase text comprehension for adults with a learning disability (Jones et al., 2007). Automating the process of augmenting text with relevant communication symbols is important. Technological solutions have been proposed, i.e. automatically determining corresponding graphical symbols needed to convey the meaning of a given sentence (Goldberg et al., 2009; Mihalcea & Leong, 2008; Zhu et al., 2007).

Being able to automatically translate to and from communication symbols is a challenging task, but it would have great impact, by helping people with communication difficulties to better understand and be understood by those around them. Each direction – symbol to text and text to symbol – poses different challenges. Translating from symbols to text is challenging due to the fact that much of the information needed to produce a textual realisation of the input is missing, such as morphological specifications, e.g. verb tense and function words. Likewise, translating

<sup>1</sup><https://www.communicationmatters.org.uk/what-is-aac/types-of-aac/#graphic-symbol-sets>

<sup>2</sup><http://arasaac.org>

text to symbols is challenging because of underlying ambiguities Figure 1.1. Text is ambiguous from a computational perspective because words can have several senses, which can be related (polysemy) or unrelated (homonymy) (Vicente & Falkum, 2017), and context awareness and knowledge are needed to determine the intended meaning. Both text to symbol and symbol to text require the use of natural language processing techniques.

The natural language that is the focus of this research is Modern Standard Arabic (MSA) alongside communication symbols. MSA is the standard form of Arabic used in schools and news media across Arabic countries and coexists with the local Arabic dialect. MSA is spoken in formal settings by a population of around 274 million (Ethnologue, 2021). However, few AAC technologies support Arabic (Alsari et al., 2020). Although considerable progress has been made in MSA language processing (Habash, 2010), some essential tools and resources remain missing or are not robust for practical use. For instance, a resource often used in AAC applications, corresponding to WordNet in English (Fellbaum, 2010), does not have the same good coverage as evident from the statistics (Bond & Foster, 2013). Unlike English, such a resource cannot be used independently and requires other external tools, e.g. a lemmatiser which needs additional linguistic knowledge about the lexicon covered.

Arabic language processing is complex when compared to English. Arabic often has a larger vocabulary than English in any comparable corpora (Alotaiby et al., 2009). This is due to morphology (e.g. adjectives inflect for number and gender) and agglutination (e.g. attaching object pronouns to verbs). Furthermore, Arabic is often written without diacritics which contributes to the amount of surface word ambiguity. Also, the subject of a verb may be dropped when it is a pronoun that can be inferred from the verb's inflected form. Additionally, there is no indication (such as capitalisation) that can be exploited to distinguish proper nouns from other word classes. All these characteristics pose challenges when wishing to link a word to a relevant symbol.

Processing natural languages, in general, is also difficult. To begin with, grammar is generative; there are always novel sentences that can be created which a computer (and the listener!) has never seen before. Then, there is often the issue of ambiguity on many levels, such as lexical, semantic and syntactic levels (Allen, 2003). Despite this, significant progress has been made in the field. The majority of natural language processing (NLP) tasks in the last three decades have been tackled using data processed by statistical or machine learning techniques. The data needs to be large, match the domain and the specific language. However, in the field of disability, genuine data is not available Kane et al., 2017. Finding a corpus that matches the AAC domain is particularly hard in English, let alone other languages with more limited resources. Researchers have often had to resolve the issue by using alternative data resources.

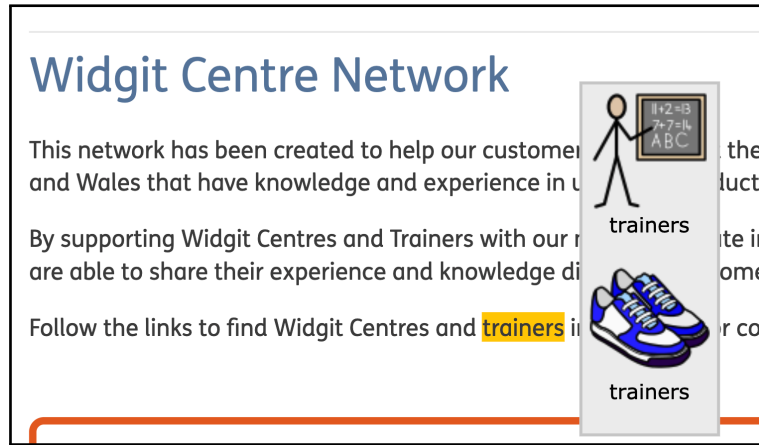


Figure 1.1: A symbol linked to the word ‘trainer’ found at <https://www.widgit.com/-training/index.htm>

A textual corpus, once available, may require further annotation, since raw data is not sufficient for certain tasks. These annotations can be implemented by hiring experts to do them, which is expensive. Therefore, successfully finding better ways of annotating that approximates expert linguistics will have a huge impact on projects with limited resources. Automatic tagging has been applied to word sense disambiguation (Diab et al., 2004; Diab & Resnik, 2002; Ide et al., 2002), and other linguistic tagging (Yarowsky et al., 2001).

This research proposes an approach to automatically annotate a corpus with data needed for text to symbol and symbol to text translation. The resulting annotated corpora will be suitable for data-based methods for developing symbol to text/text to symbol solutions. Previous contributions to the symbol translation task have not focused on the data component. They often tackled the problem using existing tools (such as word sense disambiguation) that are not always available in other languages. Further, these tools may not be well-suited to handling conversational text. Certainly, the quantity and quality of the data plays a critical role in the performance of the tasks when using data-based methods. Thus, it is crucial to focus on the data. Focusing on the data also allows researchers to increase their understanding of the relationship between communication symbols and text from the textual processing point of view.

## 1.1 Contributions

The main interest of this research is symbol to text and text to symbol translation, focusing on the data component and how to automatically annotate a corpus with the required labels. The research concentrates on MSA since it is a language that has not gained much attention from AAC technological intervention. Thus, there is a need to address linguistic challenges that may not be present in highly researched languages



such as English. Challenges such as rich morphology, agglutination, pronoun-dropping (pro-drop), and the problem of missing diacritics, have a significant impact on processing text, communication symbols and the quality of synthesised speech all of which are important for AAC users, families and carers and other people working in this field. Although a novel approach has been suggested with MSA in mind, it uses data-based methods that can be applied to any other language given the availability of the same resources, namely, a multilingual parallel corpus, a rule-based morphological analyser, and an open licensed electronic dictionary. However, it is important to point out that an inferred MSA morphological analysis annotation process is considered a significant contribution to MSA research in the field of AAC.

This research is intended to contribute to technological natural language processing aspects of improving output from symbol to text and text to symbol on AAC speech generating devices, rather than any human computer interaction aspects of AAC technology such as user interface design or access options, which are considered out of the scope of this research.

The following research question is being asked:

Q: Given the lack of a manually tagged corpus, how can one automatically tag an Arabic corpus with relevant communication symbols using a multilingual parallel corpus, that would be suitable for building a system that will translate text to symbols and symbols to text using data-based machine learning techniques ?

For this research, the ARASAAC pictographic symbol set was selected, and an exploration of it was carried out (3). A multilingual parallel corpora of movie and TV subtitles Lison and Tiedemann, 2016 was selected, of which the English portion was examined for its relevancy against an AAC sample of text. This sample was made available by experts Beukelman and Gutmann, 1999) and showed a high relevancy score compared with other corpora. An approach was then adopted that automatically annotated the Arabic side of the corpus with data needed for translating to and from symbols. The approach made use of multilingual parallel corpora (covering Arabic, English and Spanish) and that communication symbols are often distributed with glosses in more than one language. This process provided the knowledge needed for both lexical and semantic disambiguation. The approach also used open dictionaries to provide a set of possible full base form in MSA in addition to making use of their definition as an additional source of knowledge used for disambiguation.

The result is large set of utterances from the selected corpus that has been successfully automatically tagged with the needed information, which includes lemmas and ARASAAC pictographic symbols. A step further was undertaken by inferring morphological attributes that agree with the identified lemma, the surface form and other cross-linguistic contextual data. The symbol annotation was analysed through

symbol and contextual words co-occurrences to demonstrate the reduction in ambiguity that has been achieved due to the use of data collected from the multilingual corpora.

The resulting tagged corpus was used to generate an additional parallel corpus of telegraphic against full form sentences. The process of transforming fully formed text to telegraphic text was described and the resulting corpus simulates a symbol input message when ignoring the pictograph element. The process allows the creation of training data needed to train a symbol to text translator. The resulting telegraphic and full formed sentences parallel corpus was used to train a neural translation system and its performance was measured using Bilingual Evaluation Understudy score (BLEU).

With respect to symbol tagging an additional step was carried out that examined the visual content of symbols. The motivation was to examine their potential in reducing the ambiguity of the glosses that are associated with symbols. Computer vision algorithms were used to identify visually similar symbols that are presumably semantically similar and make use of their associated glosses as knowledge to disambiguate the target gloss and add additional contextual data. Similarity between the visual content of symbols was captured based on overlapping features extracted using scale-invariant feature transform (SIFT) in addition to histogram of oriented gradients (HOG). Examples showing similarity between symbols, which often correlates with similarity between the conveyed meaning, was captured and presented. Further observations on the tagging process were discussed.

A summary of the contributions of this thesis are as follows:

- The content analysis of an open licensed pictographic symbol library, exploring the visual attributes of individual symbol types and their impact on symbol to text translation that can inform best practice.
- An approach to automatically tagging an MSA corpus which exploits a domain-relevant bilingual parallel corpus in addition to available dictionaries.
- Proposing a method to simulate textual input from a symbol message suitable for a symbol to text translation system.
- Re-framing the text to symbol task as a sense disambiguation task, which is tackled using translations as a sense inventory and a bridge to relevant symbols, resulting in a corpus tagged with symbols.
- Demonstrating the potential of computer vision methods in discovering visually similar symbols; such an awareness can improve disambiguation.
- Proposing a method to obtain additional contextual data, based on associated glosses and similar visual content.

## 1.2 Thesis Outline

This introduction has provided an overview of the context of this research, the motivation behind it and its contributions. It first identified the domain and clarified the types of symbol that concern this research. It also highlighted the difficulty of NLP in general, and MSA and symbol with text in particular. It then focused on the importance of data in tackling NLP problems and the need of an annotated corpus that could be expensive to acquire if done manually, and suggest automatic annotation instead.

Chapter 2 is a literature review of related NLP tasks in the context of AAC technology. Some of the characteristics of vocabulary and utterances provided by experts in AAC are presented. AAC research prototypes involving some text processing are reviewed. Contributions involving graphical symbols are explored, and how evaluation is performed. It also mentions a few of the widely known communication graphical symbols. It includes some background information about Arabic and related tasks, including word sense disambiguation (WSD), surface realisation and automatic tagging.

Chapter 3 provides a content analysis of the ARASAAC communication set. The visual content of the symbols and the vocabulary associated with them are explored. The problem of sense ambiguity with respect to the associated glosses are highlighted through examples found in the symbol set.

Chapter 4 describes the problem of converting both text to symbol and symbol to text, and how the required annotations are automatically generated. It compares the selected corpus with other corpora. A process is introduced that makes use of the resulting tagged corpus that created telegraphic and full form parallel text, which was aimed at the symbol to text task. The focus then shifts from associated glosses towards the visual content of the symbols by examining the similarities between the visual content of the symbol set to show their potential in disambiguation.

Chapter 5 The results are reported in four sections. The first shows some statistics about the collected data. The next presents statistics about the resulting textual corpus aimed at the symbol to text task. The corpus was used in machine translation experiments and the results were measured using the Bilingual Evaluation Understudy score (BLEU) score and compared against English. The chapter gives examples showing symbol to context co-occurrence versus word to context from the corpus tagged with symbols, which is intended to be used to build a text to symbol translation mechanism. Finally, the chapter reports results from a symbol set analysis, with examples showing the potential of knowing symbols with similar content to aid the text based analysis.

Chapter 6 is a discussion that includes observations and insights as well as the limitations of selected resources and the approach followed.

Chapter 7 is the conclusion and future work.

## Chapter 2

# Literature Review and Background

### 2.1 Overview

The aim of this chapter is to review the literature that is concerned with translating between text and symbols to support those people with speech, language and communication difficulties. The review begins by clarifying the term Augmentative and Alternative Communication (AAC). It then looks at AAC technologies that involve natural language processing before focusing on the main problem. Next, pictographic symbols are clarified, briefly exploring a few symbol sets and the particular contributions involving symbols within AAC systems and how they have been evaluated. Thereafter, the focus turns to the target language MSA, and its characteristics that affect translation between text and symbols. NLP techniques are reviewed – beyond the AAC literature – that are related to the translation task and a few techniques in computer vision briefly reviewed that can be useful with symbols. The final section is concerned with data and related issues.

### 2.2 Introduction

Face-to-face communication involves the exchange of messages between two individuals. Messages can be encoded in speech, text, gestures and images. However, communication may fail due to “regional, social, or cultural/ethnic variation of a symbol system” (American Speech-Language-Hearing Association [ASHA], 1993), or a communication disorder, which is “an impairment in the ability to receive, send, process, and comprehend concepts or verbal, nonverbal and graphic symbol systems” (ASHA, 1993). People with certain communication disorders use various forms of AAC to help overcome their impairments. Communication Matters (the UK Chapter of ISAAC) (Communication Matters, n.d.) defines AAC as

“the term that describes various methods of communication to get around problems with ordinary speech. AAC includes simple systems such as pictures, gestures and pointing that “add on” to speech. More complex help involves the use of sophisticated computer technology.”

According to data collected in the UK, the top groups that could benefit from AAC are those with Alzheimer’s/dementia, Parkinson’s disease, Autistic spectrum disorder, learning disabilities and stroke (Creer et al., 2016).

As the definition suggests, the various AAC interventions range across many sensory, physical and cognitive disabilities (Beukelman & Mirenda, 2013). Some AAC interventions do not require an external resource such as manual signs, gestures and facial expression, while other interventions depend on some external resources, such as communication books and electronic devices (Moorcroft et al., 2019). The choice of AAC intervention depends on many factors: the user’s fine or gross motor ability, their understanding, their expressive speech and language skills, and other external factors such as support and funding (Goldbart & Marshall, 2004). For example Binger and Light (2006) reported that a group of pre-school children in special education, who had developmental delays and autism, used “different types of AAC system (often, more than one), including gestures (62%), sign language (35%), objects (31%), pictures (63%), and high-tech devices or SGDs (15%)”.

The use of electronic devices which generate speech are one type of AAC intervention. These devices can be referred to as voice output communication aids (VOCA) or speech generating devices (SGD). Communication devices may be designed specifically for the purpose of aiding communication or can be AAC software or apps that are installed on mainstream platforms such as iOS or Android. The technology can be designed to support one or more specific communication tasks such as conversation, transaction or narration, and may target a certain communication medium, such as face-to-face or email exchange (Higginbotham et al., 2007). The technology is aimed at users with atypical physical, cognitive or/and linguistic requirements. Depending on the user’s needs, these devices may be augmented with additional components to allow alternative input methods, such as eye trackers, brain-computer interface, or switches that can be activated using several parts of the body (Higginbotham et al., 2007). The technology often supports several modalities such as text, speech, symbols and manual signs. Software can facilitate communication in many ways which can be further improved by making use of machine learning and artificial intelligence techniques. Since such a technology is often designed for a specific community, this can result in some communities having few or no AAC technological solutions due to its lack of availability or appropriateness. In such cases, such technology may require localisation to support local languages and cultural needs.

Research on AAC technology spans multiple disciplines (Newell & Alm, 1994). Some studies focused more on the user rather than the technology itself, such as research in psychology, special education science, and speech and language therapy. Other fields, such as human computer interaction (HCI), accessibility and especially natural language processing, paid more attention to the technological aspects of the problem.

Natural language processing research is concerned with a wide range of functions that require some form of language knowledge such as word prediction, spelling correction and translation between modalities. Because AAC systems are multidisciplinary, it is important to point out that this research is concerned with the text processing part of an AAC device, and specifically the text and communication symbol interaction to create messages. Thus, the literature reviewed focuses on contributions involving either text processing or pictographic symbols in the context of AAC technology.

A significant number of processes in AAC technology involve natural language processing (NLP) (Newell et al., 1998). More than two decades ago researchers pointed out the lack of NLP research in the field of AAC (Langer & Hickey, 1999). Compared to NLP research into machine translation and dialogue systems, little research has focused on NLP as part of AAC systems.

The existing literature on NLP as part of AAC systems is often task specific. Among these tasks is supporting storytelling and conversational narratives (Black et al., 2010; Black, Waller, Reiter, & Tintarev, 2012; Black et al., 2008; Black et al., 2011; Black, Waller, Turner, et al., 2012; Newell & Alm, 1994; Tintarev et al., 2016; Waller & Black, 2012; Waller et al., 1999; Waller et al., 2013; Waller & Newell, 1997). This type of conversational messaging involves talking about a sequence of events that a person has gone through, such as when responding to: “How was your school/work today?”. Such a task can be supported by providing additional information that helps the user. Other contributions have focused on transactional conversation which have been supported by developing scripts for certain situations, such as booking a flight (Dye, Alm, Arnott, Harper, et al., 1998; Dye, Alm, Arnott, Murray, et al., 1998; Vanderheyden et al., 1996). Research has also proposed ways to support telling jokes and riddles as part of an AAC system (Manurung et al., 2006).

## 2.3 AAC Symbols and Text Processing

AAC research has targeted a language processing task needed as part of an AAC system, rather than a communication task such as sign language recognition, e.g. (Cooper et al., 2011), text to sign language translation (Kahlon & Singh, 2021), allowing phonemes to be the unit of input (Trinh et al., 2012), transforming telegraphic input into fluent output that targets word-based systems (Demasco &

McCoy, 1992; McCoy, 1997; McCoy et al., 1998), and translating between text and graphical communication symbols (2.3), which is the focus of this research.

### 2.3.1 Symbol Sets

Graphical symbols are an important communication tool for individuals with communication disorders. Symbols refer to concepts that may be a representation of a physical object or person, action or idea that may be concrete or abstract, e.g. Mum, beauty or happiness. Most symbol sets cover the semantic aspect of the language without an inherent syntax or morphology system (Light & McNaughton, 2014). A symbol set is expected to cover at least a core vocabulary list of a language, which is usually made up of single words such as pronouns, verbs, nouns, adverbs and adjectives as well as other functional words. Some sets provide morphological markers to the symbols, which may change the morphology and syntax of the corresponding glosses or sentence.

These symbols can be printed on cards or communication boards or embedded in electronic systems (see Figure 2.4). These non-alphabetical symbol sets can be used to replace text and speech but in many cases there is the hope that they will augment spoken communication. A user can point at or select symbols to convey a message to the communication partner. These symbols are mainly used to increase communicative competence (Bondy & Frost, 1994) or as a tool for literacy learning (Edran, 2002). They can be added to AAC systems (Vandeghinste et al., 2018), to a special word processor, or to an authoring tool (Lundälv & Derbring, 2012b).

For the purposes of this research the term ‘pictographic symbols’ will be used as this encompasses line drawings that depict a concept that can be represented in colour or black and white. These symbols have also been referred to as icons (Chang et al., 1992; Chang et al., 1993), or pictographs (Sevens et al., 2015a) and pictorial symbols/representations (Goldberg et al., 2008; Mihalcea & Leong, 2008).

Many symbol sets have their own style such that experts can recognise the source symbol set when presented with only a few symbols. Symbol sets can differ in their interpretability; the meaning of some symbols can be obvious while others may require some learning in advance (McClure & Rush, 2007). In order to understand the differences between some symbol sets and the impact this can have on language generation, the following paragraphs provide a description of three widely used AAC symbol sets.

Blissymbolics, developed by Charles Bliss and published in 1949 (Bliss, 1949), is an international semantographic system that was initially developed to be used by people who do not share a common language. The symbol set was later found useful as a communication tool for children with cerebral palsy. It has a special way in



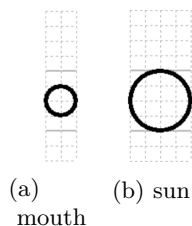


Figure 2.1: Bliss example showing that size changes the meaning of a symbol




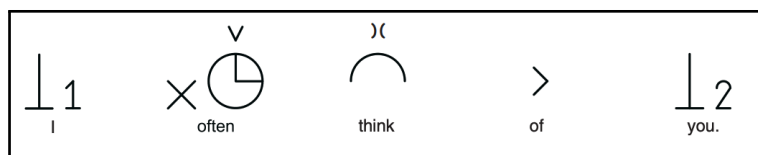
|  |   |   |
|--|---|---|
| <br>mouth | <br>nose | <br>breath,<br>breathing,<br>respiration |
|--|---|---|

Table 2.1: Bliss-characters versus Bliss-words

Figure 2.2: Bliss sentence from the Swedish National Agency for Special Needs Education and Schools (SPSM) (2010) [https://www.blissymbolics.org/images/Bliss\\_English\\_SPSM\\_10964.pdf](https://www.blissymbolics.org/images/Bliss_English_SPSM_10964.pdf)

representing meaning through geometrical shapes. The position as well as the size of the shape has an effect on the meaning, for instance, a small circle represents a mouth but a large circle represents a sun (Figure 2.1). The symbol set is made of Bliss-characters which are the smallest unit of meaning and can be combined into Bliss-words to represent additional concepts, for instance, mouth and nose are Bliss-characters that can be used independently and can also be combined to form a symbol for breath (Figure 2.1). Symbols can be combined to form a sentence (Figure 2.2). The symbols also include optional morphological indicators. In fact, Blissymbolics is not merely a symbol set, but can also be considered a visual language (Archer, 1977). It has been widely used alongside many spoken languages and by those with different kinds of disabilities.

However, due to the geometrical type of representation of the symbols, the set is categorised as having low transparency (Mizuko, 1987). The meaning of its symbols is difficult to recognise compared to other symbol sets (Mirenda & Locke, 1989; Mizuko, 1987) and needs to be learned. Thus, such a symbol set may not be suitable for all potential AAC users, including those with cognitive impairments.

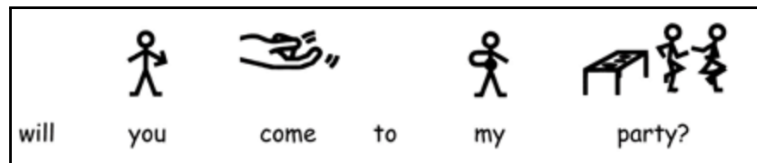


Figure 2.3: Makaton sentence example [https://www.makaton.org/shop/examples\\_computer\\_use](https://www.makaton.org/shop/examples_computer_use)

Makaton graphic symbols (Grove & Walker, 1990) is part of a multimodal communication programme, developed in the United Kingdom, which combines verbal speech with hand signing that follows the same word order. The programme was later augmented with graphical symbols (Figure 2.3). The target population was initially deaf adults with learning difficulties, but was later found to be useful for anyone with communication difficulties. Today Makaton is used in nurseries and schools across the UK and by those developing programmes for children’s national television channels. The majority of the symbols are pictorial; however, some symbols are linked to British Sign Language (Grove & Walker, 1990), which is not a universal language and thus may not be appropriate for use with other communities who may have their own sign language.

ARASAAC is a communication symbol set developed by Sergio Palao and owned by Aragonese Government in Spain. The set is freely available for non-commercial use as it is distributed under Creative Commons License which makes them accessible for AAC users on any VOCA. An initial examination of the symbol set suggests that the majority of the symbols are pictorial and thus may be easier to recognise, similar to other symbol sets that are described as pictorial, such as PCS (Mirenda & Locke, 1989; Mizuko, 1987), and can be classified as a highly transparent set. The symbol set contains over 10 000 symbols distributed with their corresponding glosses in several languages and appears to provide good coverage for frequent English words (3). The ARASAAC symbols have been used by several researchers (Lundälv & Derbring, 2012a; Paolieri & Marful, 2018; Tuset et al., 2010). This study selected this as its main symbol set and is further discussed in Chapter 3.

However, using a symbol set developed for one community may not be acceptable by another community (Huer, 2000). AAC symbol sets developed for a certain community are usually biased towards a certain appearance, e.g. skin/hair colour, clothes, furniture or buildings. The set can also lack concepts that are not universal such as religious precepts, local meals, social practices, rituals and festive events. This matter has been highlighted in research, for example when using Picture Communication Symbols (PCS) with Palestinian students (Patel & Khamis-Dakwar, 2005). One group have worked on developing a symbol set that addresses these cultural differences in their target community (Draffan, Wald, Halabi, Sabia, et al., 2015). Nevertheless, the cultural gap in symbol sets is an issue beyond the scope of this research.

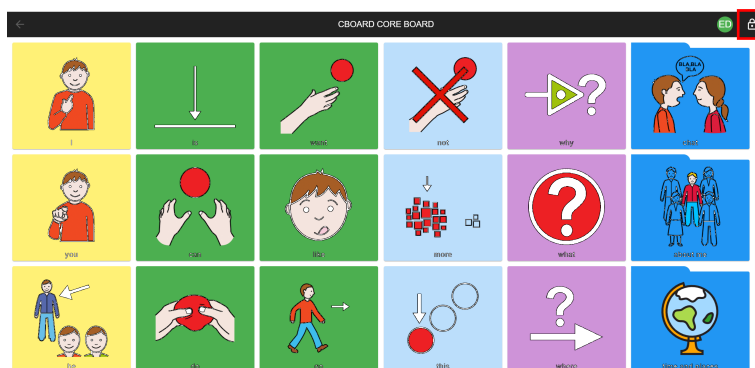


Figure 2.4: Example of a symbol board using ARASAAC symbols

### 2.3.2 Managing Associated Glosses

Symbol sets are often distributed with glosses in multiple languages. For example, Blissymbolics is distributed with glosses in 17 languages. Glosses need to be translated into the local spoken language if not already available. Automatic translation of glosses is to be avoided as it will result in errors due to word ambiguity, since glosses are out of context and often lack metadata that could be useful for disambiguation. It is important to be aware of the actual sense that a certain symbol is designed to convey before translation. For instance, a symbol may be associated with an English word such as ‘lead’ which can not be translated before knowing whether the symbol is ‘a person with a dog on a leash’, ‘a piece of grey metal’ or any other meaning associated with the word ‘lead’ since each meaning might correspond to distinctive words in another language. Few errors have been observed in ARASAAC English glosses (chapter 3). Thus, the translation requires manual effort, which depends on the size of the symbol set.

In order to resolve gloss ambiguity, research has made use of pre-existing ontologies such as WordNet, and linked a gloss to a specific meaning. This was undertaken as part of the development of a joke generator intended for AAC users (Manurung et al., 2006; Ritchie et al., 2007) in which they manually aligned their chosen symbol set with WordNet synsets (a set of words that convey the same meaning along with a definition) so that generated text is supported by relevant graphical symbols. A text to symbols translator system intended for Dutch users (Vandeghinste & Schuurman, 2014; Vandeghinste et al., 2017) was manually linked to more than 5 000 symbols using Cornetto (a database which combines the Dutch WordNet with additional data (Vossen et al., 2008)).

The Concept Coding Framework (CCF) was developed as part of a project that aimed at promoting web accessibility (Lundälv & Derbring, 2012b; Lundälv et al., 2014; Lundälv et al., 2006). CCF was an attempt to solve the problem of word to symbol

mapping as well as the issue of replacing one symbol set with another given the existence of multiple graphical symbol sets. The system attempted to avoid the need for additional effort in order to develop applications that support multiple symbol sets and also multiple natural languages. CCF was composed of a set of concepts taken from WordNet and augmented with additional ones to cover closed class words. WordNet provided CCF with a lexicon and a variety of relationships between words. The goal was to link symbols from different libraries as well as words from different languages to this set of concepts. The idea was to encode words/symbols into concepts then decode them into other desired symbols set/words. However, due to licensing issues, the work eventually only covered Blissymbols and ARASAAC. Ultimately, it was not clear how much of the linking was made manually or had been reviewed, so that symbols were mapped to relevant concepts where the same sense was maintained when exchanging a symbol in one set with a corresponding symbol in a target set. For example, a symbol representing nail in a “fingernail” sense would be translated into a symbol in another symbol set representing the same sense rather than arriving at a “thin pointed piece of metal” sense. Such an ambiguity needs to be carefully resolved for each symbol set. It seemed that this was not achieved (Lundälv et al., 2014), although adding some form of disambiguation was mentioned as a future task in their earlier paper (Lundälv et al., 2006).

A CCF-Symbol Server was implemented that users could download on their local machines or use via the web. Several applications made use of this server: a symbol editor (Symbered), a word processor extension (CCF-SymbolWriter), an AAC application (CCF-SymbolDroid), on screen keyboard (SAW 6), and a symbol text to symbol tool for the Swedish language (Lundälv et al., 2014). Although, the objective was that glosses – in the database – were disambiguated, the ambiguity of the text that needed to be augmented with symbols was not discussed. The next section reviews the literature concerned with this problem.

CCF was provided as an open source framework with documentation to attract AAC developers to use it. However, part of the team who worked on the applications above were also part of the group who worked on CCF development. Unfortunately, there is not a clear distinction, in the CCF web page, between CCF as an ontology and the applications that made use of it (CCF-SymbolWriter and CCF-SymbolDroid). It seems that there are no textual files, such as ones in XML format, that allow others to examine the ontology without installing a server on their local machines. Additionally, there is no documentation that provides instructions on how to extend the coverage by adding additional symbol sets or extending an existing symbol set with newly-developed symbols. Technical issues such as how symbols are represented are not discussed. For instance, is it based on actual file names, so that whenever a symbol set management team decides to change their naming system (perhaps adding a part of speech tag), links to the concept codes will be lost? Unfortunately, according to the

project's website [www.conceptcoding.org](http://www.conceptcoding.org), the CCF has not received any updates since 2014.

### 2.3.3 Generating Symbols

Graphical symbols can be used for expressive as well as receptive communication. Those with a receptive language disorder who use graphical symbols benefit from translating text to graphical symbols. Augmenting text with graphical symbols was found useful to increase text comprehension (Jones et al., 2007). Vandeghinste et al. (2017) devised a system that converts Dutch text to symbols. Instead of simply matching words with symbol file names, they made use of a Dutch lexicon (Cornetto) similar to the English WordNet. The coverage of words was extended by making use of relationships in Cornetto. The input text was pre-processed by tokenisation, lemmatisation, part-of-speech tagging and spelling corrections. The system's approach to handling multiple senses is by choosing the most probable sense. The system also takes advantage of different relationships provided by their lexicon that allows for the detection of synonyms and hypernyms. At an early stage, the system was extended to other languages using WordNet (Sevens et al., 2015b), but the latest version seems to be limited to Dutch text. The addition of a word disambiguator was examined (Sevens et al., 2016). They used an existing Dutch word sense tagger. These senses are used to locate relevant symbols using their sense-aware lexicon. Word sense disambiguation (WSD) has improved the accuracy of the output and further improvements to the word sense tagger is mentioned as future work.

Symbered is an editing tool that aims to find suitable symbols for a given text (Lundälv et al., 2006). The system made use of the CCF. Although the task involves text as the input, authors did not point out the need to perform any natural language processing tasks. The task of lemmatisation and part of speech awareness is important in locating relevant symbols. This issue was apparent since their system was not able to find a relevant symbol for 'books' even though 'book' was among the symbols in the database. They also had ambitions to support multiple languages, without mentioning the difficulties of translating between languages such as the differences in word order and length between the source and target texts.

The CCF was also used to build CCF-SymbolWriter: a word to symbols lookup system integrated into a word processor. The tool shows all matching symbols (matching is based on word form) and allows the user to choose the relevant one. No NLP tools are mentioned and the system appears to handle word inflection by adding them explicitly to the database. Nysnö is another system that also makes use of the CCF. It augments Swedish simple text with symbols. It uses part of speech tagging to narrow down the number of matched symbols. If the lemma is attached to more than one sense, the

system uses external resources to determine the most frequent sense and chooses it as the matching symbol.

Among the commercial systems that provide similar functionality is Symbolate by Boardmaker and SymWriter by Widget. Both work as a text editor where a symbol appears as soon as a word is typed and allows the user to replace the suggested symbol with another one. It is not clear how exactly a symbol is suggested when multiple symbols are found. Widget claims it uses smart symbolisation, since it takes the part of speech tag into account to find suitable symbols.

The PicNet illustrated dictionary translates words to pictures and also uses WordNet (Mihalcea & Leong, 2008). They developed a sense tagger to determine the sense of a word among the multiple senses defined by WordNet. The identified sense was then used to choose a relevant picture. However, their goal was to evaluate how capable the images were of expressing meaning conveyed in simple sentences without any glosses in the target language. They found that the amount of information that can be understood from a sequence of images is comparable to a machine generated translation.

Goldberg et al. (2008) pointed out the importance of finding a layout of pictures that best communicates the meaning of a sentence. They argued that a flat sequence of images may not be the best layout and suggested a specific layout, which groups images and uses a classifier to predict the grouping of images.

### 2.3.4 Generating Text

Symbols can be embedded in electronic devices. These devices are often augmented with speech synthesis which allows a message composed of graphical symbols to be spoken. The spoken message can be simply the sequence of glosses associated with the selected symbols. However, the derived message often lacks function words and morphology, which may not sound natural to the hearer. Research has addressed this issue by proposing methods of translating the telegraphic text to full form text.

AAC users that communicate using graphical symbols, rather than spoken words, are dependent on symbol to text translations linked to speech synthesis, if literacy skills have not been achieved (Pino, 2014). The task of translating symbols to text has been a topic of research since the 1980s (Hunnicut, 1984). The translation is made for a single message or utterance exchanged in a conversation. Most systems that are built to translate symbols to text in AAC are not actually translating from symbols but rather translating from glosses combined with some specifications such as the part of speech. Thus, they do not take the actual graphical representation into account. In this respect generating text from a symbol-based or a word-based AAC system is the

same. Compansion (McCoy et al., 1998) was a system that developed for word- as well as graphical symbol-based AAC systems.

While the challenge when translating from text to symbols is word sense disambiguation (section 2.3.3), the challenge in the other direction is word order and recovering inferred missing tokens and morphological features. Most contributions assume the user will follow the natural language word order. McCoy et al. (1994) compared input composed by an AAC symbol user against a corrected version suggested by a speech therapist. The participating users were adults with cerebral palsy. Several edits were made: inflecting words, adding missing words (usually function words or words that can be inferred by the context), deleting words, replacing words and changing word order.

The Compansion system (Demasco & McCoy, 1992; McCoy, 1997; McCoy et al., 1998) is a project focused on developing a tool for translating telegraphic words to fluent English. The process of generating text was made in three stages. The first stage was to chunk text and tag words with their part of speech. The next stage was to generate a semantic parser. Semantic parsing uses predefined cases that are linked to a specific verb. These cases specify a semantic role (i.e. agent or theme) with respect to a specific verb and a ranked list of possible fillers. These roles were ranked based on their importance. The authors also used WordNet to be aware of the presence of certain properties for a candidate such as ‘animate’ or ‘edible’. The next stage was to generate text given the semantic parsing. They made use of an existing text generator, which was based on functional unification grammar to generate possible fluent text sentences. A system for real use planned to make use of the Compansion prototype. They pointed out that the prototype assumed an English word order, which was not a reasonable assumption when they observed the target population. The problem of unexpected word order was addressed by using an icon prediction system and allowing an icon to be selected only if it offered a valid sequence.

Karberis and Kouroupetroglou (2002) proposed a similar rule-based system which generated fully-formed Greek sentences from a sequence of symbols. The system relied on handcrafted rules to recover missing function words and to ensure a grammatically accurate output. They assumed that the symbols were in an order that matched Greek syntax but might lack function words.

Other proposed systems used statistical methods to generate fluent text, such as an n-gram language model which could be generated using a large corpus (Sevens et al., 2015b; Vandeghinste et al., 2018; Waller & Jack, 2002). To address the lack of morphological cues in the input sequence, possible surface forms are hypothesised. Sevens et al. (2015b) and Vandeghinste et al. (2018) also used a reverse lemmatiser that generate, possible inflected forms for a given word. Other possible synonyms are added to the list of hypotheses, which are extracted from the lexicon linked to their symbol



set. Articles are also suggested. The authors used trained their model with Dutch text to find the most likely sequence among several possible that were formed using the hypothesised tokens. The generation used a beam decoder, and an optimisation algorithm to tune the decoder's parameters to control the search space. They claimed that their proposal was language independent. However, it still depended on the availability of a tool to generate possible inflections of a word for a given language.

Waller and Jack (2002) also used an trigram model to obtain a fluent English sentence given the symbols by predicting missing function words and correct word form. They assumed that the input symbol message followed English syntax, but might lack function words.

SymbolPath is a system that does not assume that users will follow a certain word order (Wiegand, 2013; Wiegand & Patel, 2012b). It is a symbol-to-English text system and, as the name suggests, the input is made by drawing a path over selected symbols. As a result of this selection method, it assumes that selected symbols may not be precise in either content or order. It determines the intended selection based on semantic frames and semantic grams (probably similar to the skip-grams language model (Stolcke et al., 2011)). The sentences generated are then ranked according to the user's drawn path. However, the generated utterance needs to be simple in terms of the number of verbs, actors and modifiers. The fluency of the text was not evaluated.

Reiter et al. (2009) stated that text generation in the context of AAC is easy since users tend to use short sentences and simple structures. For example, they found that children do not use perfect tense, having examined a sample produced by children who had no communication difficulties. They stated that the main challenge was maintaining coherence since they were looking at an utterance, not in isolation, but as part of a personal narrative. However, it was not clear how they were resolved the syntactic role of each token, perhaps by using semantic frames or data-based methods.

Two studies that involved the use of Arabic have been published. Al-Arifi et al. (2013) reported the development of an Arabic AAC system, tackling the problem from a human interaction viewpoint without dealing specifically with linguistic processing. Ding et al. (2015) proposed a means of translating symbols to Arabic and English text. The Arabic sentence generated was not novel but retrieved from an English-Arabic corpus, based on Arabic and English labels.

### 2.3.5 Evaluation

All tools that are developed in the course of research need evaluation. The goal of an AAC system is to increase the user's communicative competence. In this context, some software interventions have been tested with typical speakers, such as the study carried out by Todman et al. (1995). A quantitative evaluation was carried out by hiring a



number of judges to evaluate randomly sampled chunks from unaided conversations, and conversations aided by the developed system. Most recent contributions are evaluated by running small scale qualitative studies with participants from the same target group, e.g. nine children with cerebral palsy (Ritchie et al., 2007), or a case study with one participant with cerebral palsy (Waller et al., 2013).

Unfortunately, functions that are part of an AAC prototype are not usually tested individually but instead the whole system is evaluated for usability. For example, a text generation component as part of an AAC prototype (Black et al., 2010; Black, Waller, Reiter, & Tintarev, 2012; Dempster et al., 2010), supporting personal narratives, did not demonstrate any evaluation of the text generator component independently of the other functions provided. Many AAC systems have a retrieval component that finds prestored messages, so the idea of using prestored messages is usually evaluated without any attention being paid to the retrieval component itself.

The success of symbol to text translations are affected by many factors including the interface design, the quality of the different functions provided, the response time, and the user's capabilities. Todman et al. (1995) suggested evaluating transcripts of conversation to eliminate the negative effects caused by other variables such as the quality of the speech synthesiser. McCoy and Hershberger (1999) pointed out the difficulties of evaluating a component that is made for an AAC system, since the whole system needs to be designed and implemented to incorporate the component. They also stated that a lack of sufficient training may result in negative outcomes.

Among the few studies which did evaluate the generated text was Waller and Jack (2002), who analysed 20 sentences generated from three different sittings. The text generator proposed by Sevens et al. (2015b) was evaluated using 50 messages (975 word and 746 symbol). The quantitative evaluation used metrics developed for evaluating machine translation, such as BLEU (Papineni et al., 2002). They manually translated the test set and compared the machine generated text from different versions of their prototype, and likewise for symbol generation.

The Dutch text to symbol system (Vandeghinste et al., 2017) was evaluated quantitatively using machine translation methods by comparing the symbols that were produced by different versions of their system. The methodology used precision and recall, a commonly used metric in information retrieval and classification that is not affected by the order of tokens. The evaluation was against a manually labelled test set. A similar system, which translates foreign text to pictures, has been used to evaluate a system with 50 sentences made up of an average length of 15 words (Mihalcea & Leong, 2008). The evaluation asked participants to interpret the pictures generated, and compared their interpretation using MT metrics and manually against actual reference text. Their system was not evaluated against other systems but rather they measured the significance of each intervention that was added to the system. It

appears that many obstacles prevent the comparison of different systems, such as the different symbol sets used, the representation of symbols used (local file names, ID or actual pixels), and the lack of a standard symbol text representation.

## 2.4 Arabic as a Target Language

Arabic is a Semitic language spoken by around 274 million people (Ethnologue, 2021). It is also the language of the Holy Quran which makes it a popular language among Muslims around the world. Modern Standard Arabic (MSA) refers to the language used in books and formal media, and is taught in schools. Other forms of Arabic are classical Arabic, which is the language of the Holy Quran and the spoken language in the pre-Islamic era and the first few centuries after Islam (7th to 11th century C.E.) (Alrabiah et al., 2013), and colloquial Arabic, the local informal spoken Arabic which differs slightly from MSA. Dialects vary phonologically, lexically and morphologically between each other and from MSA. The overlap coefficient between vocabularies of MSA and its various local dialects was 37% at its highest (Bouamor et al., 2018). Many Arabic dialects occur within the same geographic area: Bedouin vs urban communities tend to speak different dialects, and differences exist within the same spoken dialects for different age, gender, social class and religious groups (Watson, 2002), which makes it hard to choose one and also to discriminate one from another. Unfortunately, dialects have limited resources and standards, since Arabic linguists focus on classical Arabic, and colloquial Arabic is not taught in formal education. The language addressed in this research is MSA, due to the larger population that uses this variant (throughout the Arab World) and the availability of linguistic resources.

In the Arabic script, Arabic is written from right to left, and has 40 letters. Arabic makes use of diacritics, which are short vowels that determine the exact vocalisation of a word. However, most text is written without diacritics (as an abjad), since good readers can infer them from the context. However, diacritics are necessary when considering text to speech technologies used in AAC devices or for early literacy skills. Arabic orthography lacks any distinguishing feature for proper nouns, such as capitalisation.

Arabic is a rich morphological language (Al-Sughaiyer & Al-Kharashi, 2004). Arabic morphology is derivational and inflectional (Ryding, 2014). Derivational morphology allows words to be derived from roots by means of templates. A root, in Arabic, is not a word in itself, but has an abstract meaning. Arabic verbs in particular are very systematic (Habash, 2010). There are 19 verb forms and are derived from trilateral and quadrilateral roots. Knowing the root and verb form for a given verb often allows the derivation of its verbal noun and participles. This can be useful in expanding a pre-existing lexicon. Inflectional morphology is what makes a word change within the

same part of speech class. Inflection is expressed through a set of features. Nouns and adjectives inflect for gender, number and case. The feminine and plural form can be regular or irregular and needs to be explicit in a lexicon. Verbs inflect for tense, person, gender, voice, number, and mood. Inflection is made by changing the stem form and attaching prefixes or suffixes occasionally with orthographic adjustment. Commenting on Arabic inflection, Ryding (2014) argues: “Compared to English, words in Arabic are highly inflected”.

Arabic is to some degree agglutinative (Buckwalter, 2004; Farghaly & Shaalan, 2009). Conjunctions, clitics and the definitive article, along with the stem (an inflected word with no attached clitics) can all be part of a single token. Attachments can be a sequence of affixes and the order in which they are attached is regulated. As a result, a space-delimited token in Arabic may contain much more information than, for example, in English. Tokenisation is thus important, but identifying a stem’s boundary is ambiguous and requires awareness of the context.

Knowing the Arabic syntax is essential for some NLP tasks. Arabic usually follows the order of verb (V), subject (S) and object (O), but SVO and VOS are also used. Additionally, Arabic is a partially pro-drop language, meaning that subject pronouns may be omitted. Palmer et al. (2008) found that the subject was pro-dropped in 30% of sentences in the Arabic Tree Bank. The pro-drop together with the flexible order makes it difficult to identify the subject of a given verb and as a result parsing a sentence can be a challenge. Furthermore, a verb is not an essential part of a sentence since the verb ‘to be’ does not exist in the present form (Ryding, 2014).

The syntax of the sentence affects the morphological features of the words. Morphological features in general are determined either semantically or contextually (Ryding, 2014). Knowing the syntax is essential for ensuring the contextual morphology of a word, such as the gender and number of an adjective. These contextual features often agree with the main token, i.e. subject or modified noun, but the position of the verb with respect to the subject makes a difference. Also, the “non-human nouns” (Ryding, 2014) or “irrational nouns” (Habash, 2010) are treated differently in terms of agreement, and awareness of such a classification (human nouns vs. non-human nouns) is important.

#### 2.4.1 MSA with Symbols

Some languages are more challenging than others due to the nature of the language or limited resources or both. Koehn (2005) ran an experiment that involved translation of 11 European languages and pointed out that translating into English was easy compared to other languages, while translating into rich morphological languages (e.g. German) was difficult. There was also the issue of agglutination (where complex words

are made up of several morphemes) resulting in an inferior performance in translation for a language like Finnish. MSA is a rich morphological language often written with no diacritics.

In the text to symbol task, analysing the text morphologically is important in order to link tokens to symbols. For English, knowing the part of speech is almost sufficient to determine the lemma, given a dictionary and a set of rules. For Arabic this is far more complicated as, due to morphology and agglutination, it is difficult to know the base stem boundaries without awareness of the context. The attached clitics are also ambiguous, e.g. an attached 'taa' to the end of an Arabic verb in its perfect form may mean 'I', 'you' (masculine) or 'she'. It also can be part of the base word. Furthermore, the pro-drop aspect of Arabic can make it hard to tell whether the subject is explicit or should a pronoun symbol be (re)introduced. Also, the lemma can have multiple senses and the likely sense needs to be identified to choose a matching symbol.

In a symbol to text task, where the source is a set of symbols with their lemmas (as glosses), predicting the syntax and morphology of the output is needed. For instance, there is a need to determine the likely verb inflection, since no indication is given in the input and the context is unlikely to give any hints. For verbs there is a need to determine the subject to ensure that it agrees with the gender and number. Adjectives also need to agree in gender, number and definiteness with the noun they describe. Additionally, function words can be in different forms depending on the gender and number of the main noun, for example 'This is a bag' and 'This is a book'; when said in Arabic 'This' will not be in the same form for both, because bag is feminine and book is masculine.

## 2.5 Related Work

Translating between text and symbols involves several NLP tasks. Sense disambiguation is needed for generating symbols and was a tool used by other researchers (Mihalcea & Leong, 2008; Sevens et al., 2016) (2.3.3). Sense disambiguation requires some preprocessing of text which mainly involves part of speech tagging and lemmatisation (Vandeghinste et al., 2017; Zhong & Ng, 2010). On the other hand, those who tackled generating text from a symbol message have used language models and a morphological generator (Sevens et al., 2015b; Vandeghinste et al., 2018). However, other approaches can be applied, such as those followed in machine translation. This research takes a step further and examines the potential for using the visual content of the symbol set for further disambiguation.

### 2.5.1 Word Sense Disambiguation

A word in isolation (or out of context) may have many meanings and these meanings are often listed in conventional dictionaries under the same head word. However, a word in context often has a clear meaning which a mother-tongue speaker unconsciously identifies. Yet, identifying the intended meaning computationally is a difficult task. The importance of this for text to symbol translation cannot be underestimated, as any glosses with the same form can have very different meanings, resulting in a choice of images.

Researchers have been interested for many years in the problem of automatically disambiguating the meaning of a word. The task of word sense disambiguation is to identify for a given word, the sense relevant to its context (Edmonds & Agirre, 2008). The disambiguation task can be designed to disambiguate all words (content words), or focus on a subset of words. Methods used can be classed as: supervised learning, unsupervised learning, and knowledge-based (Navigli, 2009).

Supervised learning requires a corpus in which words are tagged with a finite set of senses that apply to the word in context. For instance, Miller et al. (1993) tagged part of the Brown Corpus with WordNet senses. A tagged corpus is used to train a classifier using various machine learning algorithms such as Support Vector Machine(SVM) (Zhong & Ng, 2010). Unsupervised learning does not require a tagged corpus nor a pre-existing sense inventory. It can be achieved using vector space models and clustering algorithms (Schütze, 1998).

Knowledge-based methods make use of data associated with a specific sense such as their definition, synonyms or examples from a tagged corpus. Some researchers have augmented these senses with additional data from Wikipedia (Mihalcea, 2007). The data can be used to determine the likely sense based on the overlap between the collected data and the words in the context (Lesk-simplified and Lesk-corpus (Kilgariff & Rosenzweig, 2000)). This approach has been shown to achieve results that are comparable with more sophisticated methods. Agirre et al. (2001) collected relevant data (topical signatures or context vectors) from the web for each sense. The data was collected by forming a query that ensured that the retrieved data matched the specific sense, and not an arbitrary sense covered by the same word form. This was accomplished by including words and phrases that are part of the definition, as well as synonyms with varying degrees of restriction, until a set of documents is found. The topical signatures were tested in a WSD task, focused on a small sample of seven word-types. These were disambiguated by choosing the sense with the highest weighted overlap score. The results were better than random and superior to using only WordNet as a source of data. The performance of this use of topical signatures was close to the most frequent sense (Cuadros & Rigau, 2006), but using topical signatures alone did not outperform supervised methods.

Supervised and knowledge-based systems require a finite set of senses available beforehand. An inventory of senses is often constructed by lexicographers analysing word uses in a corpus, such as WordNet (Fellbaum, 2010), although the senses identified do vary between lexicographers. WordNet senses – although often used – have been criticised by many researchers for their suitability for WSD, such as Kilgariff (2006). These senses are too specific for the computation task (Ide & Wilks, 2006) which makes them hard to discriminate. The accuracy of the classification task is higher for coarse-grained senses compared to fine-grained senses (such as WordNet). For example, IMS (Zhong & Ng, 2010), a WSD classifier, achieved 68% for all words in the fine-grained sense task but scored 82.6% in a coarse-grained sense task. In fact, determining the accurate sense (WordNet senses) can be difficult even for humans. Snyder and Palmer (2004) carried out a manual tagging experiment. The agreement between tagged samples from two annotators were found to be 72.5%. Coarse-grained senses were automatically generated through clustering fine-grained senses (Navigli et al., 2007). The clustering was guided by an attempt to map senses in WordNet to another dictionary and grouped if they linked to the same external sense. It is difficult to decide how far fine-grained senses (such as WordNet) would yield the desired performance.

Linguists distinguish between two forms of ambiguity, namely polysemy and homonymy. Words that share the same form are homonyms if they express unrelated meaning and have distinctive historical source. For instance, coach meaning “bus” and coach meaning “sports instructor” are homonyms (Vicente & Falkum, 2017). On the other hand, a word is polysemous if it expresses more than one meaning that are related (Lyons, 1995). The word “mouth” is polysemous as expressed in “John has his mouth full of food” and “Watch your mouth” (Vicente & Falkum, 2017). Research has shown that there are differences between how the brain processes polysemy and homonymy. Klepousniotou (2002) confirmed that homonymy appears to be slower to process. Computationally, homonymy is easier to disambiguate (Edmonds, 2005). Ide and Wilks (2006) suggested that differences in meaning between homographs (homonyms that share the same written form) are the ideal level of distinction that WSD should aim at. They suggest that these senses can be identified as “senses that psycholinguists see as represented separately in the mental lexicon, are lexicalized cross-linguistically, or are domain-dependent”. Such a distinction seems appropriate for symbols; however, further research may be needed.

Words that are lexicalized cross-linguistically are easy to obtain from bilingual dictionaries or a bilingual parallel corpus. Resnik and Yarowsky (1999) investigated how often a pair of senses of the same word form are translated into distinct word forms (i.e. are lexicalized) in other languages. They found that homographs are often (95%) translated into distinctive words and the percentage decreases as the semantic gap between the pair of senses decrease. Researchers tackling the WSD problem have made

use of a second language, through a parallel corpus, to use translations as sense tags or to automatically tag words in a corpus with their likely senses from a pre-defined sense inventory, to create training data that is needed to develop a WSD classifier (Diab et al., 2004; Diab & Resnik, 2002; Tufis et al., 2005; Zhong & Ng, 2010). This idea may therefore have potential when considering the disambiguation needed for tagging text with symbols, given the existence of parallel corpora in Arabic and English.

Diab et al. (2004) targeted MSA using bilingual data for disambiguation. The approach taken was to tag Arabic based on a parallel English corpus by using English translations to decide on the appropriate sense from the WordNet sense set for an Arabic word in context. Results were encouraging. However, the disambiguation was limited to nouns. Nouns, as part of WordNet, are privileged with relationships that can be exploited in the disambiguation process, such as the ‘is-a’ relationship which is not available for other classes, such as verbs. The approach made use of these relationships by applying Resnik’s algorithm (Resnik & Yarowsky, 1999). Using the same approach with other classes such as verbs may not be as successful as nouns.

A collection of MSA newswire corpora were tagged manually with Arabic WordNet senses (AWN) senses covering 5218 word types to examine the potential of evolutionary algorithms in tackling WSD (Menai, 2014). This resource is valuable, but newswire data does not match the target domain of AAC and is limited by the AWN. AWN (Elkateb et al., 2006; Rodríguez et al., 2008) covers only 48% of the most frequent 5000 senses (Bond & Foster, 2013).

WSD classifiers preprocess the text by part of speech tagging and lemmatization. The POS tags provide useful information for the lexical disambiguation process (Wilks & Stevenson, 1998).

### 2.5.2 Part of Speech Tagging and Lemmatisation

Lemmatisation and part of speech tagging are common preprocessing tasks. Part of speech tagging labels each token with its relevant tag from a closed set. Recent taggers have been implemented using machine learning classifiers and require a tagged corpus. The tagging accuracy part of speech in English has been around 97% (Ma & Hovy, 2016; Manning, 2011). Manning (2011) commented on the tagger’s accuracy “But this seems surprising – anyone who has looked for a while at tagger output knows that while taggers are quite good, they regularly make egregious errors”. However, the tagging process in many other languages may not have yet achieved this accuracy.

Arabic is a rich morphological language resulting in a large part of speech tag set. Various sets have been used for tagging Arabic text (Diab et al., 2004; Habash, 2010). A part of speech tag set may capture not only the basic part of the speech type but also morphological aspects. For instance, the Buckwalter tag set was used in tagging



the Penn Arabic Treebank (PATB) (Maamouri et al., 2004), which covers inflectional morphology such as verb tense and noun number. The ElixirFM tag set was used in tagging the Prague Arabic Dependency Treebank (PADT) (Hajic et al., 2004) which captures the functional morphology, unlike the form-based Buckwalter tag set (Habash, 2010). The consideration of all morphological features leads to a very large tag set, e.g. 400 tags were used in PATB (Habash & Roth, 2009). As a result, some tag variants that are coarse have been used. For instance, the Columbia Arabic Treebank was tagged with a tagset of only six tags, which was later automatically expanded to 44 tags (Habash & Roth, 2009). The choice of tag set depends on the main task; in some cases, a large tag set needs to be reduced to overcome the problem of sparseness. Zeroual et al. (2017) called for a standard tag set and suggested a hierarchical one. A universal tag set has been proposed to solve inconsistencies between languages (Petrov et al., 2012), which can be useful for a multi-lingual symbol set. Taji et al. (2017) investigated mapping the Buckwalter tagset to the universal dependency set.

Lemmatization is another important preprocessing task for many NLP systems. This is “The reduction of the word tokens in a corpus to their lexemes” which are “the form that heads an entry in a dictionary”(Brown & Miller, 2013b). For English, knowing the part of speech tag and having a lexicon is sufficient for determining the lemma. However, for Arabic the case is complicated. This is mainly due to the absence of diacritics. Several attempts to restore diacritics have been proposed. For instance, Belinkov and Glass (2015) proposed a tool that restores diacritics without the need for any external tools. However, depending on the task, morphology analysers might be needed. Morphological analysers can suggest one or more analyses that include diacritics, POS tags and/or the corresponding lemma. A number of morphological analysers have been proposed (Al-Sughaiyer & Al-Kharashi, 2004). The Buckwalter Arabic morphological analyser uses a lexicon and compatibility table to analyse a word (Buckwalter, 2002). The analyser yields various possible analyses for a single word. Each analysis includes the full vocalised form, part of speech, English glosses and the corresponding Arabic lemma. MADAMIRA takes a step further by determining the likely analysis (Habash & Rambow, 2009; Pasha et al., 2014). This is achieved by extracting features from the context to rank different possible analyses of a word. Scoring is based on how close an analysis is to predicted features. The prediction is made using Support Vector Machines (SVM) and n-grams, trained using the annotated newswire corpus. The tool was tested on an unseen split of the same corpus and achieved 85% accuracy and 95% lemma accuracy. MADAMIRA is a commonly used tool. However, it is not clear how MADAMIRA would perform on conversational data due to domain differences.

There is evidence that syntax is different for each domain (Sekine, 1997), which impacts morphology. A group of researchers have created a part of speech tagger for social media text and showed that their tagger achieved a 25% relative error reduction



compared to a general-purpose tagger when tested with twitter data (Gimpel et al., 2011), evidencing the effect of domain variation. This highlights the domain mismatch problem faced by AAC systems in general, and Arabic AAC systems in particular.

Other common pre-processing steps are tokenisation and normalisation. Tokenisation can be simply done by separating out all punctuation, however this is usually not sufficient for natural languages. Tokenising Arabic text is important due to agglutination which impacts Arabic text significantly. The process usually involves separating out clitics in addition to punctuation, numbers, and other symbols. Also, the base stem may require orthographic adjustment to reverse changes that were needed for attaching prefixes and suffixes. Tokenisation is important for many tasks. When considering symbols, tokenisation is needed to map each sub-token to its relevant symbol.

Multiple tokenisation has been used when handling Arabic text (Badr et al., 2008; Habash & Sadat, 2006; Maamouri et al., 2004). They differ in the class of segments that is detached from the base word. Segments may be classified into conjunctions, particles, definite articles, and pronominal clitics. For instance, The PATB (Maamouri et al., 2004) opts to tokenise only clitics that have a different syntactic category from the base word, and so does not tokenise definite articles. Habash and Sadat (2006) defined three tokenisation levels:

$$[CONJ + [PART + [AI + BASE + PRON]]]$$

Schemes vary in how deep they tokenise and whether prefixes and suffixes are further tokenised or not. Habash and Sadat (2006) described several schemes. The first scheme tokenised the conjunctions only. The second tokenised particles as well as conjunctions, while the third also tokenised the definite article and pronominal clitics. Badr et al. (2008) compared translating English to Arabic using a scheme that is similar to the deepest scheme developed by Habash and Hu (2009), which tokenises a word to a similar base, but without tokenising the prefix and suffix. The authors found that keeping the prefix and suffix untokenised yielded the best results. El Kholy and Habash (2010) compared all the schemes mentioned in an English to Arabic phrase-based translation task and found that the PATB scheme achieved the highest BLEU score (Papineni et al., 2002). The scheme was also used as a preprocessing step in a neural-based MT system and improved the BLEU score (Almahairi et al., 2016). The suitability of this scheme with English suggests that such a tokenisation might also be appropriate for symbols.

Given a tokenised text, detokenisation is needed to generate Arabic text. However, this is not a straightforward task. It may not be clear whether a token is part of the previous token, or the following token, or is independent. Attia (2007) pointed out the ambiguity that results from detokenising clitics and suggested inserting a mark to

indicate the direction in which it needs to be attached could solve the problem. El Kholly and Habash (2010) compared six different methods of detokenising. Regardless of the tokenisation scheme deployed, they found that the best method was to use a mapping tool alongside a language model and to adjust the concatenation as a back-up. The mapping tool maps a tokenised sequence to the most probable detokenisation form, drawing on a table learned from a corpus containing tokenised words against observed detokenisation and their probability. Adjustments are basically orthographic corrections made while concatenating segments by following a specified set of rules. Al-Haj and Lavie (2012) made almost the same comparison but suggested that the addition of a language model led to insignificant enhancements in accuracy at greater computational cost.

There are often spelling variations throughout a corpus that need to be standardised. In English this can be upper and lower case variations. In Arabic, these variations usually fall into three classes: variations of Alef (Alef with Hamza above, Hamza below or bare Alef), Yaa (dotted Ya with dotless Ya) and Altaa Almarbuta (sometimes written as a Haa). There is also the problem of diacritics occasionally appearing in the text. These inconsistencies cause many problems in NLP tasks such as sparseness in language models (Heintz, 2014). Dictionary-based morphology analysers will fail to find a corresponding analysis (Buckwalter, 2004) and cause a low recall in information retrieval. Thus, a preprocessing step (normalisation) is often carried out to reduce orthographic variation (Habash & Rambow, 2009). This is typically achieved by removing diacritics and transforming variations of Alef and final Yaa into a single form.

### 2.5.3 Language Models and Machine Translation

Text generation can be broken down into stages: text planning, sentence planning and finally surface (or tactical) realisation (Rambow et al., 2001). Text planning might be important in some applications, such as dialogue systems where an input may require the generation of several utterances. The sentence planning phase decides the abstract syntactic structure. Determining the syntactic structure varies depending on the application. For instance, in a dialogue system this can be guided by the communication goal (Rambow et al., 2001), while in image captioning it can be limited to a specific form such as declarative present-tense syntactic structures (Mitchell et al., 2012). Finally, the surface realiser uses the abstract structure to generate a meaningful sentence by ordering words, adding function words, and generating the relevant morphological form of each word. Surface realisation can be achieved using templates (Mitchell et al., 2012; Yang et al., 2011), but the resulting system may be limited in coverage. It can also be implemented as a rule-based system using grammar formalism (Elhadad & Robin, 1996). However, rule-based systems require detailed input specifications and are unable to handle missing data. Some overcome this issue by

making use of statistical data to find the best generated output (Langkilde & Knight, 1998).

Alternatively, realisation can be approached statistically using language models built using a corpus (Mairesse & Young, 2014; Oh & Rudnicky, 2002). The language model is used alongside a search algorithm to find the best sentence among different possible hypotheses, given the input and possibly some heuristics that limits the space of possible hypotheses. This approach of text generation is used as part of the statistical machine translation framework (Koehn, 2009; Lopez, 2008) where the input is possible translation phrases or words and the output is a sentence that is fluent and adequately covers the source sentence. The same concept is also followed in neural machine translation (Koehn, 2020), but the text generated is based on a source sentence directly, rather than translating segments and then stitching them together. Language model-based generation have also been used in generating text given symbols (Sevens et al., 2015b; Vandeghinste et al., 2018; Waller & Jack, 2002).

Language models assign probabilities to sequences of words and these are an essential component of many NLP tasks such as machine translation (MT), speech recognition, and spelling correction (Rosenfeld, 2000). The model needs a corpus to be built but no annotation is necessary. The model estimates the probability of a word 'w' given a history 'h'  $P(w|h)$ . A history is the preceding words (a special start symbol is used when predicting the first word). The probability of a sentence is approximated by making use of the Markov assumption which proposes that the probability of an event depends on recent history (or the last few words) (Markov, 1954). This was important to overcome memory limitations and to generalise the unseen language that is not captured by the corpus.

The probability of a sentence is approximated by calculating the joint probability of words conditioned by recent history 1-4 words (bigram - five-gram model). The probability is calculated using the maximum likelihood estimation by collecting counts of sequences of n words from a given corpus. A problem arises when an unseen word, which yields a zero-probability, makes the whole sentence have zero probability. Researchers have tackled this problem by giving some probability to unknown words (Chen & Goodman, 1999; Kneser & Ney, 1995). There are various ways of obtaining such a probability, known as probability smoothing, which is beyond the scope of this research. Interpolated Kneser-Ney smoothing is the most widely used method (Brants et al., 2007). To tackle unseen sequences, it combines the n-gram model with all lower-order n-gram models in addition to a weighted uniform distribution that includes the unknown token (Chen & Goodman, 1999). The corpus needs to be large and match the domain.

However, statistical language models are unable to capture long term dependency, in other words they are unaware of words that occurred before the n-1 words which may

be a key in predicting the next word. Another issue is they are not able to generalise by making use of similarity between words beyond observed sequences. This causes issues with highly morphological languages, due their sparse counts as a result of their larger vocabulary.

Researchers have suggested methods to make these models more general, such as a class-based language model (Brown et al., 1992), which is based on word clusters. Another example is the factored language models (Bilmes & Kirchhoff, 2003), which make use of multiple levels of linguistic information per token and is especially helpful for highly morphological languages (Kirchhoff et al., 2006; Novais & Paraboni, 2012).

However, statistical language models have been replaced with neural-based models (Bengio et al., 2003; Mikolov et al., 2011), with many different architectures. Recurrent neural network (RNN) based language models have shown to significantly outperform statistical models. Chelba et al. (2014) showed that RNN language models (Mikolov et al., 2011; Pascanu et al., 2013) outperformed statistical language models even when trained over a one billion word corpus. RNN is a neural network that sequentially reads a variable length of tokens. This is achieved by passing accumulated information from one step to the next (e.g. each step reads one word). Theoretically this can be repeated infinitely, resolving the long dependency issue that was a problem with statistical language models. The network is trained to predict the next word by adjusting weights to minimize the loss.

A probability distribution can be simulated by applying a SoftMax function over the output. However, RNN is a simple sequence model that suffers from memory issues, especially for long sequences. Other RNN variants with improved memory have been used, such as LSTM (Schmidhuber, 1997) and GRU (Cho, van Merriënboer, Gulcehre, et al., 2014). In some situations, bidirectional sequence models are used, which combine two sequence models: one reads a text from left to right and the other from right to left (Schuster & Paliwal, 1997). This allows information about both the left and right context to be available for any word. It also provides a stronger coverage for the whole sentence. Additionally, several networks can be stacked to form what is known as deep learning models, which have shown further improvements (Pascanu et al., 2014). The power of linguistic neural-based models have not yet been exploited in symbol translation AAC.

Language models are essential but they are not sufficient for text generation. For instance, Sevens et al. (2015b) used a process responsible of adding function words and a reverse lemmatizer to generate hypotheses. However, the approach in which machine translation has been addressed seem to offer a better, cleaner solution. Machine translation has been approached using statistical methods (Brown et al., 1988; Koehn et al., 2003) and later using neural networks (Cho, van Merriënboer, Bahdanau, et al., 2014). The statistical approach of several components and many variants have been

proposed. The phrase-based model (Koehn et al., 2003) was widely accepted before being replaced by neural methods. It depends mainly on three components: a language model built from a corpus in the target language, a phrase translation model that is extracted from a parallel corpus, and a decoder. Using a phrase model (sequence of tokens), as opposed to a word model, has resulted in significant improvements. This is not surprising since translating single words in isolation will introduce additional translations into the set of hypotheses that do not match the context and miss less likely translations that do match the context. Translation requires awareness of the context, which makes translating symbol glosses automatically prone to errors.

The Encoder-Decoder approach to machine translation (Cho, van Merriënboer, Bahdanau, et al., 2014), uses a neural network with two sequences, an input sequence that is fully read in order to generate the output sequence. The architecture is composed of two main parts: one that encodes the input into a single vector, and the second receives the encoded input to generate the output. This approach has been successfully used in tasks such as image captioning (Vinyals et al., 2015), question-answering, and translation (Cho, van Merriënboer, Bahdanau, et al., 2014). The model has been improved by adding an attention model (Bahdanau et al., 2015) that allows the decoder to focus on certain segments of the input in addition to the encoded vector.

Vaswani et al. (2017) later discovered that attention is actually sufficient and a new architecture emerged known as the Transformer that achieved higher accuracy. This new architecture is more efficient to train compared with RNN models, especially when the text length is less than the number of dimensions. Transformer, instead of sequentially reading one word at a time, reads all input tokens at once, considering their word positions. It also has multi-headed attention that may be the key to the increased accuracy, by allowing awareness of different features. As a result, transformers became a good alternative for recurrent based models. Progress has been made since 2017 and several similar models have emerged. BERT is a stack of transformers (Devlin et al., 2019). Two variations in size were explored (12 and 24 transformers). The model was trained in two phases: first to understand the language, and then trained (or fine-tuned) on a specific task. In the first phase, they were trained on two tasks, masked language prediction and sentence prediction. Their intuition was to allow the model to understand language before training on a specific task.

Regardless of the model used, the process will sequentially generate the probability distribution over the vocabulary. The final sequence of words can be the most probable token for each step, i.e. greedy decoding, but this has not always produced optimal results (Gu et al., 2017). Alternatively, beam search can be used, which keeps the top  $k$  sequences or hypotheses as it progressively reaches the end token. The choice of  $k$

may have unexpected effects in neural nets. Britz et al. (2017) found that increasing the beam width beyond 10 resulted in a lower accuracy.

There has been lots of progress in data based methods in MT and many technical options, but the essential requirement for them all is data. For the current research, a corpus is needed covering both Arabic and a pictograph symbol set. Such a corpus needs to be large to be useful. Irvine and Callison-Burch (2013) showed that the size of training data correlates with accuracy.

#### 2.5.4 Computer Vision

Symbols are image files that are associated with glosses. Glosses are isolated words with no textual context and as a result remain ambiguous. As mentioned earlier, some symbol sets are provided with glosses in more than one language which can be useful in disambiguation. Yet, the content of the image is a key in gloss disambiguation. As a result, researchers have opted to manually link symbols to entries in a dictionary. However, it remains unknown whether the visual content of pictorial symbols can be useful in gloss disambiguation. For instance, identifying visually similar symbols and retrieving their associated glosses may provide some textual context for disambiguation. Computer vision research yields many methods that can be used to calculate visual similarity between two symbols.

##### 2.5.4.1 Colour

Colour distribution within an image has been used as a feature for image search engines. It is a simple representation based on the distribution of colour over pixels without considering their spatial location. It was among the earliest features used for image retrieval. However, usually the number of colours is very large, which makes comparing the colour distribution of two images challenging. This can be handled using colour quantization, which limits the colour space and avoids sensitivity to insignificant colour variation. Several methods have been used to compare colour histograms and the Euclidean distance algorithm appears to be one of the most used methods. This process involves a bin-by-bin comparison (grouped data with equal width in a histogram). Consequently, with fine-grained colour space, two closely similar colours might exist in two different bins and will be considered different regardless of the closeness of their actual values or similarity to the human eye.

Coarse-grained colour space may combine two colours that are perceptually different in one bin. Instead, a cross-bin comparison can overcome such a problem. Rubner et al. (2000) argued that Earth Mover Distance (EMD) outperforms other cross-bin methods. This avoids the need for a fixed set of colour bins and instead generates a compact

colour distribution of an image by clustering colours in each image independently. EMD was used with Euclidean distance and can be used in image retrieval systems. The method is also able to detect partial matches. Other colour comparison methods can be made using colour moments (Stricker & Orengo, 1995), and colour correlogram (Huang et al., 1997) which considers the spatial correlation of colours.

A specific colour can be represented by the magnitude of three channels: red, green and blue. Such a representation is referred to as RGB. RGB suffers from a perceptual non-uniformity which affects Euclidian distances between colour points Tkalcic and Tasic, 2003. Apart from RGB, various colour spaces have been suggested by experts. Rubner et al. (2000) suggested using LAB colour space, since the spatial distance between points in LAB are designed to better match human perception.

However, identifying similar symbols based on colour alone may not result in semantically similar symbols. For instance, the symbol for 'orange' the one representing the fruit and the other representing the colour can not be distinguished based on colour alone. Therefore, similarity based on shape is also needed.

### 2.5.5 Local Descriptors

Local descriptors aim to identify interesting local areas and generate a vector that compactly describes these sub-areas, so that two perceptually similar areas will correspond to two spatially nearby vectors. Many local key descriptors have been proposed. The most widely used method is SIFT, which generates descriptors that are robust against scale and orientation differences (Lowe, 2004). Wu et al. (2013) showed that SIFT outperformed other key descriptors in scale and rotation. Although neural-based methods have dominated computer vision in recent years, SIFT remains useful in some situations (Zheng et al., 2018).

The algorithm search for interesting areas in the image then encodes a description in a vector. The algorithm searches for potential points by comparing the oriented gradients over multiple Gaussian smoothing degrees. Candidates are points that are local extrema (maximum and minimum values). The search is repeated over multiple scales. Candidates that have low contrast or are near the edges are discarded. The local area (16 by 16) is segmented into a 4 by 4 grid, and 8-bin histogram of directed gradients is generated for each cell resulting in a 128-dimension vector. Main orientation is determined so that other orientations are computed with respect to it. Multiple descriptors may be generated for the same area if more than one significant orientation is found. Directed gradients are weighted, giving lower weights to points that are far from the centre. The resulting vector is modified to make it less sensitive to illumination changes.



A single image will have a number of interesting areas encoded as vectors. Two images are compared by calculating the number of similar descriptors irrespective of their position, rotation or scale. Similarity is determined by calculating the Euclidean distance and finding the nearest match. Lowe (2004) suggested a test to reduce the number of false matches. However, even matched key points will often contain many false matches. A common filtering technique is using Random sample consensus (RANSAC) (Fischler & Bolles, 1981). RANSAC attempts to find the largest subset of matching points that fits a geometric model, and considers other points as outliers. RANSAC may fail when the number of false matches is greater than true matches (Lowe, 2004).

SIFT is a successful method used in many problems such as object recognition and image stitching. It appears to have some potential for identifying similar symbols. However, SIFT will fail if no sophisticated area is found in a symbol. In some cases, it may be useful compare the whole symbol, although it has some limitations, against another rather than local areas.

#### 2.5.5.1 Global Descriptor

Histogram of Oriented Gradients (HOG) (Dalal & Triggs, 2005) is a descriptor similar to SIFT described above, but is derived from the distribution of the directions of the gradients. It is considered a global descriptor generated by segmenting the area covered into cells and a window slides over a block of cells with some overlap. Each covered block will produce a normalized histogram of nine bins. The histograms are concatenated so that a single vector represents the whole area. In contrast to SIFT, no normalisation is used for the orientations. HOG was shown to be successful in a pedestrian detection task (Dalal & Triggs, 2005). It was used by sliding a window over an image and a HOG descriptor was generated for each covered area. HOG is rotation invariant. However, it was not an issue in the given task since people appear in an upright position. To overcome object scale variation, the process was repeated over different fixed-size batches extracted from multiple scales of the image (Dalal et al., 2006). The resulting vector was fed into an SVM classifier.

SIFT (2.5.5) and HOG (2.5.5.1) both seem promising and they do not require any training data. They can be used to find visually similar symbols for any given symbol that may be expressing a semantically similar concept. Such knowledge can be exploited in disambiguating associated glosses. Other computer vision techniques make use of machine learning techniques and require a large set of training data. However, traditional methods were appealing when examining communication pictographic symbols due to the significant difference between these symbols and training images used to develop deep learning object recognition classifiers. Also, symbols in the same



set may have some common visual elements between semantically-related symbols that can be easy to identify using traditional methods.

An examination of the content of the ARASAAC symbol set will be carried out and described in the next chapter (chapter 3). The following section focuses on textual data.

## 2.6 Data

### 2.6.1 Corpus

Developing NLP tasks requires the availability of data that matches the domain, whether for training, for testing, or to understand the domain. AAC researchers often use a core vocabulary list (Balandin & Iacono, 1999; Banajee et al., 2003; Marvin et al., 1994). These lists are words that are frequent, used in many contexts, and cover the basic needs for communication Beukelman and Mirenda (2013) and Fallon et al. (2001). The word lists are largely made up of pronouns, verbs, conjunctions, interjections, prepositions, adverbs, and adjectives, rather than nouns (Renvall et al., 2013; Witkowski & Baker, 2012).

Another list of words is common among young learners, gathered by parents and caregivers, such as The American English The MacArthur-Bates Communicative Development Inventories (MB-CDIs) (Dale & Fenson, 1996; Fenson et al., 1993). Such a list includes more nouns compared to lists of core vocabularies and includes animal sounds and sound effects (Laubscher & Light, 2020). However, neither lists are ‘one size fits all’, but can be tailored to the individual, their situation, and the task in hand. In English, the core vocabulary have been published for different age groups, for example toddlers, pre-school, school age, young adults, adults, and older adults. Multiple factors play a role in selecting a suitable vocabulary for a user, but very little has been discussed about the differences in core vocabularies for different languages. Baker et al. (2000) claimed that AAC core vocabularies are similar across a series of European languages. AAC devices often cover these lists of core vocabularies (Cannon & Edmond, 2009). Although these lists are important and may be sufficient for symbol sets developers, they are not enough for an AAC system or application.

To be able to for example build language models, there is a need for sentences rather than isolated words. ‘Small talk’ from typical speakers has been collected in order to understand the importance and role it has in a conversation (King et al., 1995). Small talk refers to utterances that do not carry content, but have certain functions in conversation, such as “Politeness markers” (e.g. thanks, you’re welcome, please). The average length of phrases has been found to be 2.5 words, but the majority are made up of one to two words. These phrases have been classified into a set of categories such

as “Repetition Requests or Personal/Social Questions”. King et al. (1995) found that these phrases are different for each age group. They found that they make up 26% to 39% of a conversation for adults, depending on their age group. Ball et al. (1999) carried out a similar experiment with children and found “generic talk” or small talk was 48%. These findings highlight the importance of having a conversational corpus when developing an AAC system. However, these collected lists are often short and insufficient for use as training data.

Finding a corpus for a particular language which is suitable for an AAC setting is a challenge. For instance, the Brown corpus was found to provide irrelevant statistics that did not match the needs of the target group ‘children using AAC,’ while building a joke generator (Manurung et al., 2006). In an attempt to solve this problem, Vertanen and Kristensson (2011) created a corpus of messages by asking participants, through a crowdsourcing platform, to imagine being able to communicate through AAC. To ensure quality, participants had to contribute a message, but also had to review someone else’s contribution. The resulting set of messages were then used to collect similar ones from different social media platforms. The model for collecting the corpus was said to “significantly outperform models trained on the commonly used data sources of telephone transcripts and newswire text”. Black, Waller, Turner, et al. (2012) used a small corpus while developing a system that aimed to support conversational personal storytelling targeting AAC children. The corpus was collected from typical children, who wrote their personal narratives, and was used for the text generation component to capture the writing style. Vandeghinste et al. (2017) used a corpus of Dutch messages developed for a platform designed for users with cognitive disability, which was used to test and tune text to symbols and symbol to text systems as part of an AAC Dutch system. The corpus of over 69,000 messages was available, with an average length of 7.7 words. However, the majority of these messages could be considered as noise, as only a small number of messages were actually used. Precisely 186 messages were used as a development set and 50 messages were used for evaluation.

However, some pre-existing corpora may be suitable for AAC. Wiegand and Patel (2012a) used a corpus of personal blogs in a search for informal text. The corpus was used to build a prediction system to provide “automatic message expansion to generate syntactically correct messages” in English, although, once again, the output did not always produce the accuracy required. Nevertheless, a conversational corpus is better match to the task of face-to-face communication. For instance, the Spoken British National Corpus 2014 (Love et al., 2017) appears to be a good candidate since it is a collection of transcribed spontaneous conversations that occurred in an informal context, spoken by participants covering diverse demographic categories such as age, gender and socio-economic status. Mitchell and Sproat (2012) used a corpus of transcribed dialogues for an American show to build a response prediction system targeting AAC users. They justified their choice by pointing out that it was a large

corpus, composed of conversations occurring between pairs of individuals who were familiar with each other.

Waller et al. (2005) emphasised the importance of having an AAC user involved when designing an AAC system. It is also important to have a corpus of genuine AAC messages when designing AAC components that require textual data. Although a general conversational corpus might be useful for developing an AAC system, a genuine AAC corpus is needed to test a developed system or for fine tuning a system, and to be aware of the user’s needs. For instance, the collected messages written by individuals with special needs contained spelling errors (Vandeghinste et al., 2017), which developers may forget about when developing AAC systems, but such a system needs to be robust against these errors. However, a genuine AAC corpus does not exist and needs attention in the near future.

Other languages are often behind English in terms of publicly available NLP resources. It is not surprising it is difficult to find a relevant corpus for Arabic. This is especially true given the large variety of spoken dialects that are often used in informal sittings. Thus, collecting a corpus of textual messages from social media, as in Vertanen and Kristensson (2011), may not result in a consistent set of messages in a single dialect. Corpora constructed from newswire are often used in Arabic NLP, such as the data collected from multiple news agencies which were used in developing the PADT (Hajic et al., 2004) and PATB Maamouri et al. (2004). However, these were not relevant for use in AAC, since they do not match the target domain. They lack the important “small talk” for establishing and maintaining a conversation. Fortunately, a corpus of movie subtitles exists covering Arabic, as well as many other languages (Lison & Tiedemann, 2016). This might be the best currently available resource for conversational MSA Arabic. It is similar to research done in English (Mitchell & Sproat, 2012).

Having a relevant corpus is an essential step toward building an AAC system that makes use of the latest technology. However, a raw corpus may not be sufficient. As mentioned in section 2.5, several annotation is needed to tag a corpus with symbols. These annotations are normally carried out manually which can be expensive. Alternatively, an automatic approach using some source of knowledge have been undertaken.

### 2.6.2 Training Data

The lack of training data is a common problem in NLP research and researchers have looked for ways to address this problem. For instance, in machine translation, researchers have attempted to address the absence of a parallel corpus by using a third, pivot, language. This approach requires the availability of two bilingual parallel

corpora, both covering the pivot language and each covering one of the pair of languages of interest. Habash and Hu (2009) explored the potential of using English as a pivot language while translating between Arabic and Chinese. The approach worked well and also outperformed direct translation. The idea of using an additional language to create training data has been done in other tasks as well.

The problem of insufficient data has always been a concern in word sense disambiguation, referred to as the “knowledge acquisition bottleneck” (Gale et al., 1992; Navigli, 2009). Developing a tagged corpus requires hiring experts and takes time to complete, which is expensive. However, these expenses have been avoided by automatically tagging a corpus with senses using another aligned language, to create training data that is needed to develop a WSD classifier (Diab et al., 2004; Diab & Resnik, 2002; Tufis et al., 2005; Zhong & Ng, 2010). Brown et al. (1991) published one of the earliest WSD experiments that made use of a parallel corpus for sense disambiguation. Church and Gale (1991) highlighted the availability of a parallel corpus as a valuable resource in WSD. Diab et al. (2004) and Diab and Resnik (2002) exploited a parallel text to create training data that can be used for both languages. They collected a set of all target words that had been aligned to a single word-form in the source throughout the corpus. Words in a set are disambiguated by choosing the semantically nearest sense with respect to other words in the same set. This knowledge can be then used to tag words in both languages. Testing the English, showed that this approach was better than most of the other, unsupervised, methods. However, evaluations focused on nouns.

Ide et al. (2002) used aligned corpora covering multiple European languages and aligned sense inventories covering the same set of languages. Using word alignment, a pair of aligned words are assigned to the corresponding aligned senses when possible or the nearest pair of senses if no aligned senses are found. If multiple senses are found, the most frequent sense is chosen. The remaining uncovered words are disambiguated using the common sense in their cluster. They concluded that automatic tagging was found to be as good as manual tagging. English and Chinese parallel text was also used to augment available training data, with additional data covering the most frequent ambiguous words to train the SVM classifier, which has yielded encouraging results (Zhong & Ng, 2010).

Parallel corpora were shown useful in projecting other linguistic information from one language to another. Yarowsky et al. (2001) proposed a system that projects NLP analyses, such as part of speech tags, from one language to another. The results were successful. For example, they achieved a 99% lemmatization accuracy for French using an English parallel text as a source of analyses. Rogati et al. (2003) used a parallel text to train an Arabic stemmer and showed high agreement with a state of the art Arabic stemmer. These findings suggest that parallel text, once available, can be a

valuable resource for automatic annotation and especially for generating a corpus annotated with symbols.

## 2.7 Conclusion

Contributions in the field of AAC do not need to be focused on a full system or a specific communication task, but can also address specific linguistic tasks. Clearly, for AAC to be useful for an individual, it needs to support the spoken language of their community. There is lots of potential in recent NLP approaches that AAC researchers have not yet explored. However, the main issue is that recent methods require a large body of data, which may be hard to find, collect or create.

This review has shown that an AAC software in general, and symbol AAC systems in particular, developed for one language cannot be effortlessly used with another language. The amount of work depends on several factors, such as the sophistication of the target language processing, as well as the availability of linguistic resources. Furthermore, data and tools need to be well suited to the domain. Tools such as lemmatisation, part of speech tagging, and sense disambiguation are needed in addition to relevant corpora.

The task of translating between text and symbols have been approached in several ways and the task may seem reversible, but that is often not the case. Translating into symbols requires sense disambiguation, while translating into text does not. However, word sense disambiguation is not straightforward and is a research field in itself. Pre-existing sense-taggers in other languages have been used to determine the relevant symbol. Unfortunately, a reliable sense tagger is not available for MSA. Translating into text, on the other hand, has been approached mainly using n-gram language models. However, machine translation appear to be more suited to this problem by undertaking the addition of missing words and reordering words. The task still requires a relevant corpus and morphological analyser to synthesise a parallel corpus.

A corpus of genuine AAC messages is hard to find, but a corpus that is conversational seems a good alternative. However, research in MSA has often made use of newswire data which does not match the target domain. This is problematic since the corpus requires multiple levels of annotation and manual annotation can be expensive. However, automatic annotation is appealing and has been used with other NLP tasks. It would allow the generation of a tagged corpus for any spoken language and symbol set, given the availability of a parallel corpus that covered the target language as well as any other language that is well supported with NLP tools.

MSA is a rich morphological language often written with no diacritics. This increases the ambiguity of words, which can be a challenge for the morphological analysis

required. Morphological analysis is important since it is a key in identifying the lemma, the part of speech tag, and diacritics. Diacritics are needed to ensure plausible synthesised speech. Identifying the lemma and part of speech reduces the ambiguity significantly. Also, word disambiguation is essential for reducing any further ambiguity to determine relevant symbol matches.

Finally, solutions to this problem span many research areas within Natural Language Processing (NLP), such as machine translation (MT), Text Generation, Word sense disambiguation (WSD), Part of speech (POS) tagging, and lemmatisation. Additionally, since symbols are images, a few computer vision techniques have to be considered.

## 2.8 This Thesis

The problem of translating between AAC pictographic symbols and MSA text is the main interest of this research. The literature described in this chapter addressing this problem (2.3) has made use of pre-existing tools, such as word sense disambiguation tools, that may not be available in other languages or, if available, do not cover the vocabulary that is covered by the symbol set. In order to overcome the lack of tools, the approach undertaken by this research is to annotate a corpus with the required data. The annotations will be generated automatically by making use of available translations in other languages to overcome text ambiguities. The resulting tagged corpus can be used with tools available as part of a machine learning library such as Pytorch (Paszke et al., 2019).

Furthermore, the visual content of symbols that are part of the symbol set were explored using SIFT and HOG to examine their potential in providing cues that can be used to disambiguate the meaning of the associated gloss. The outcome of the review shows that disambiguation can be used to further improve the tagging process. The methodology by which the automatic tagging was carried out, and the visual content, will be explored in chapter 4, after the chosen symbol set has been discussed.

## Chapter 3

# Understanding Graphical Symbols for Communication

Various graphical symbol sets have been designed for communication with different purposes. One of the most commonly-used symbols globally are the health and safety symbols published by the International Organisation for Standardisation (ISO) (ISO\_7010). These symbols have a concrete meaning, such as a fire hazard warning or a no photography allowed policy (see Table 3.1). However, the type of symbols that concerns this study are pictographic symbols designed specifically for human communication to be used by people who struggle to use spoken or signed languages. This chapter discusses some of the characteristics of these symbols as well as their associated glosses. This study focuses on the ARASSAC symbol set, as representative of other pictographic symbol sets, which was selected due to its size, open licence and the wide range of language translations it has undergone over the years; it is also well-maintained. Analysis of the ARASSAC symbol set is needed before using it as part of any computational task, such as the provision of a corpus of annotated symbols or a symbol recommendation system. The analysis could also be useful for AAC professionals seeking further understanding on some aspects of the symbol set.

The term ‘symbol’ is defined by the Oxford English Dictionary as “A written character or mark used to represent something; a letter, figure, or sign conventionally standing for some object, process, etc.” A graphical symbol represents a concept through strongly outlined, black and white or coloured drawings. A concept as defined by The Cambridge Dictionary of Linguistics (Brown & Miller, 2013a) as “A mental representation constructed from information about the surrounding world received and processed by human beings”. For AAC purposes, the concepts that most graphical symbol sets aim to cover are a person’s needs for daily communication in many contexts.





|   |   |   |   |
|---|---|---|---|
|  |  |  |  |
| Emergency exit  | Fire extinguisher   | General warning sign  | No photography  |

Table 3.1: Commonly used symbols (ISO 7010, 2019)

### 3.1 ARASAAC

The ARASAAC symbol set was created in Spain as part of a project funded by Aragonese Government and is made available under a Creative Commons license. The total number of image files collected from ARASAAC API (each with at least one gloss in English or Spanish) was found to be 12,662 symbols <sup>1</sup>. The creation date associated with the symbols indicates that ARASAAC developers appear to have been actively extending their symbol set since 2007 (Figure 3.1), which suggests that the symbols are up-to-date and that the current gaps will be covered in the near future.

In terms of its depiction style, researchers have compared a few symbol sets to examine their translucency “the degree to which individuals perceive a relationship between a symbol and its referent when the referent is known” (Lloyd & Blischak, 1992), and transparency “the degree to which the meaning of a symbol can be readily guessed in the absence of the referent” (Lloyd & Blischak, 1992). Variations in translucency and transparency were found between symbol sets (Bloomberg et al., 1990; Mirenda & Locke, 1989; Mizuko, 1987). However, no empirical studies have compared ARASAAC symbols to other symbol sets. However, the depiction of tangible or physical objects highly resembled their referents and the set is pictographically similar to Picture Communication Symbols (PCS), which was found more translucent (Bloomberg et al., 1990) and transparent (Mirenda & Locke, 1989) than some other sets.

The open licence, the size of the symbol set, the continuous development, and the depiction style, have all motivated the decision to choose ARASAAC throughout this research. Table 3.2 shows a sample of ARASAAC symbols with their associated glosses (multiple glosses are separated with a comma). In this chapter the term symbol is used to refer to an instance of ARASAAC symbols unless explicitly stated otherwise.

Few studies have made use of ARASAAC. It appears that one of the earliest papers that made use of the symbol set was as a “pictogram-based instant messaging service that is intended to bridge the social and digital gap of people with cognitive impairments” (Tuset et al., 2010). ARASAAC was also one of the symbol sets used in the Concept Coding Framework (CCF) which was developed to map between several

<sup>1</sup>Last updated 11 June 2020



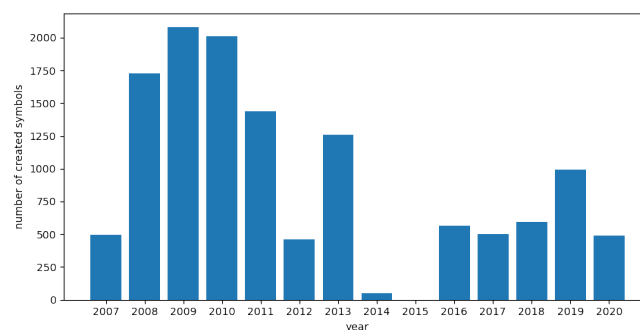


Figure 3.1: The number of ARASAAC symbols developed by year



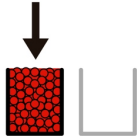



|  |  |  |
|--|--|--|
|   |   |   |
| put  | goodies, sweets, candies   | full   |
|  |  |  |
| 74, seventy-four   | Austria  | What do you have?  |

Table 3.2: A sample of ARASAAC symbols with their associated glosses


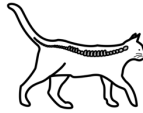

|   |   |  |
|---|---|--|
|  |  |  |
| trilobite   | vertebrate  | Murray River   |

Table 3.3: Examples of symbols representing advanced concepts

symbol sets and lexicons covering a few spoken languages (Lundälv & Derbring, 2012a). It was additionally used as a baseline while developing a localised Arabic symbol set for a Qatari Assistive Technology Centre (Draffan, Wald, Halabi, Kadous, et al., 2015). However, there do not appear to be any publications that specifically detail aspects of the visual and linguistic attributes of this graphical symbol set that have an impact on symbol to text and text to symbol translations. The only discussion found in this area was related to the ‘reliability and validity’ of ARASAAC symbols, based on responses from 219 students from the University of Jaén in Spain who were not AAC users (Paolieri & Marful, 2018).




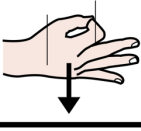
|   |   |   |  |
|---|---|---|--|
|  |  |  |  |
| apple   | fever   | past  | To be  |

Table 3.4: Symbols in different classes: an icon; an index; a symbol; another symbol

### 3.1.1 Graphical Content

ARASAAC symbols are two-dimensional simple drawings which may be coloured, in addition to a black and white version. Peirce’s theory of signs (The Cambridge Companion to Peirce, 2004, p. 8) suggests that signs can be classified into icons, indices and symbols, based on their type of relationship with the object they refer to. Icons are signs that resemble their related object. Indices are signs that are related indirectly to their object by depicting a cause or an outcome. Symbols are related by either a rule or habit.

An examination of ARASAAC symbols suggests that the set covers all three classes, i.e. icons, symbols and indices (Table 3.4). The symbol of an apple is an icon since it resembles its physical object. An symbol of a person sweating in bed with a thermometer in her mouth is designed to represent the concept ‘fever’ or ‘temperature’ depending on the language, and is an index. An image designed to represent the concept ‘past’, which is depicted using an arrow pointing anti-clockwise can be considered a symbol sign. A depiction of a hand sign to convey the concept ‘to be’ is another example of a symbol sign.

However, the use of the term “symbol” in the theory of signs may conflict with its use in the AAC domain where it is used to refer to a graphical representation that has been created/used for communication, regardless of the type of relationship with its referent. This issue was also pointed out by Lloyd and Blischak (1992). However, the use of the word symbol throughout this research will follow AAC community usage. Such a distinction between the depiction type of these symbols may have an impact on processes that make use of the visual content. For instance, images that contain icons may benefit from some pre-trained computer vision models, but images with symbols may not since their content differs from mainstream images.

### 3.1.2 Associated Glosses

A symbol set is often not a bare set of image files but rather supplemented with glosses. The glosses are words in some spoken language that express the same meaning as the pictographic symbol. These glosses are expected to cover the core spoken

| number of associated glosses | Percentage of symbol set |
|------------------------------|--------------------------|
| 1                            | 66.50                    |
| 2                            | 23.99                    |
| 3                            | 6.51                     |
| 4                            | 1.98                     |
| 5                            | 0.75                     |
| 6                            | 0.16                     |
| 7                            | 0.07                     |
| 8                            | 0.02                     |
| 12                           | 0.01                     |

Table 3.5: The number of glosses associated with each symbol




|   |   |   |
|---|---|---|
| (a)<br><br>'school table', 'school desk' | (b)<br><br>'forbidden', 'forbid' | (c)<br><br>'hands to your shoulders', 'physical education' |
|---|---|---|

Table 3.6: Symbols with more than one gloss

language lexicon. Examining the English glosses for ARASAAC symbols, it was found that most associated glosses were one word long, but some glosses ranged from two to nine words. The number of glosses associated with each symbol also varies. The majority of symbols (66.5%) in the set at the time of the analysis were associated with one gloss. 24% of the symbols were associated with two glosses. The remaining 9.5% had more than two glosses (Table 3.5). An additional gloss may have different purposes. It may be a synonym as in Table 3.6 (a). It may be a morphological variation in another part of speech class such as 'forbid' and 'forbidden' (b). It may act as a topic or category such as the gloss 'physical education' shown in (c).

Glosses are important for people who are not familiar with a specific set since symbols are not always recognisable (Mirenda & Locke, 1989; Mizuko, 1987). Glosses are also essential for some computational tasks such as symbol retrieval and translating between symbol and text. However, glosses alone are not sufficient for translation due to word ambiguity. Also, glosses need to be translated into the local language to be useful.

### 3.1.3 Parts of Speech or Word Classes

Associated glosses fall into various part of speech classes. Many glosses are common nouns which include for instance colours, fruits, vegetables and body parts. The set also includes verbs and actions which are usually depicted using an actor, i.e. human

and a common object such as a person opening a door, to illustrate the concept ‘open’. Among the glosses are descriptors, such as symbols for describing size, order and feelings. Also, participles which act as adjectives are included, e.g. ironed. Pronouns are also covered and are illustrated using human figures with attention to their gender. Prepositions are also in the set and are illustrated using abstract shapes and arrows. Also, plain characters and numbers are included. Finally, utterances such as i.e. “what do you have?” are associated with few symbols.

The majority of the glosses provided are assigned a class indicating its linguistic category. There are five classes. One class covers proper nouns and pronouns (class 1). Common nouns and verbs are each assigned their own class (class 2 and class 3 respectively). Adjectives and adverb are combined into one class (class 4). the last class is assigned to phrases (class 5). An additional class (class 6) is assigned to all remaining glosses that either do not fall in any of the five classes such as letters, numbers and prepositions or to glosses that have been entered with no explicit class (e.g. second is labelled with class 6 rather than class 4). A sample of symbols for each word class is shown in Table 3.7. A distribution of symbols between word classes is shown in Table 3.8. Common nouns make up the majority of the symbol set, which contradicts the idea of focusing on core vocabulary lists (Renvall et al., 2013; Witkowski & Baker, 2012). Cannon and Edmond (2009) suggested that this might be because they are easier to represent. However, it is important to have a good coverage of nouns since they are essential for expressing and understanding various concepts.

The classification provided is coarse compared to English part of speech classes. Nevertheless, it allows discrimination between certain part of speech classes such as noun vs. adjectives. Such a classification is useful for gloss disambiguation (Wilks & Stevenson, 1998). For instance, knowing that a symbol associated with the gloss ‘orange’ is assigned the modifiers class will be evidence that the symbol refers to the colour rather than the fruit. However, the classification should be considered with caution as a few errors have been observed that may have an impact on text to symbol translation.

#### 3.1.4 Synonymous Glosses

Although the ARASAAC set contains a large number of symbols, this does not necessarily imply an equal number of concepts. Many concepts are represented in more than one symbol, offering an alternative illustration some with only slight variations. For example, the concept ‘learn to swim’ was illustrated through multiple symbols with slight variations (Table 3.9). An additional example is the verb ‘open’, which was illustrated in several symbols depicting the action with multiple objects.









|   |   |   |   |   |  |
|---|---|---|---|---|--|
| 1 | <br>Little red riding hood | <br>Eritrea          | <br>Augustus        | <br>Pilar's Virgin | <br>La Coruña         |
| 2 | <br>stamp                  | <br>job              | <br>throne          | <br>tin opener     | <br>exit              |
| 3 | <br>to reject              | <br>to make a racket | <br>to roll up      | <br>to recycle     | <br>handcuff          |
| 4 | <br>curious                | <br>independent      | <br>no              | <br>salty          | <br>theirs            |
| 5 | <br>I have finished        | <br>i want           | <br>I want that one | <br>what is it?    | <br>have you seen it? |
| 6 | <br>a quarter to five    | <b>g</b><br>g   | <br>8:30          | <b>99</b><br>99   | <b>67</b><br>67  |

Table 3.7: A random sample of symbols for each word type

| word type | number of glosses |
|-----------|-------------------|
| 1         | 772               |
| 2         | 11937             |
| 3         | 3442              |
| 4         | 1057              |
| 5         | 273               |
| 6         | 695               |

Table 3.8: The number of gloss and symbol pairs for each word type

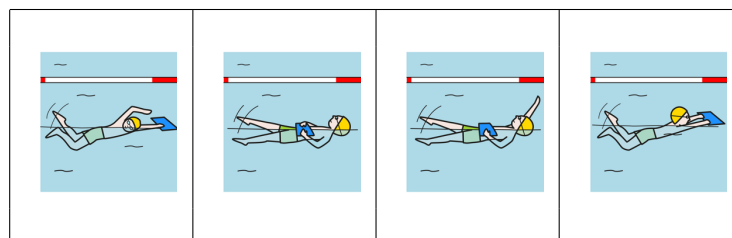


Table 3.9: Four different symbols for the concept “learn to swim”

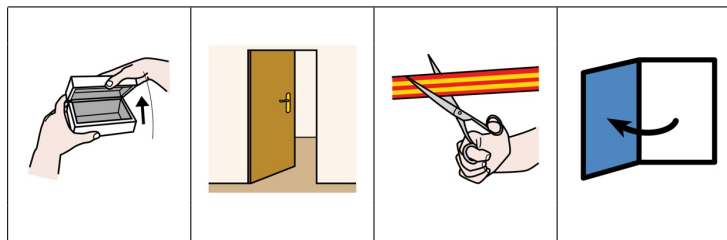


Table 3.10: Words having multiple representations (e.g. open)

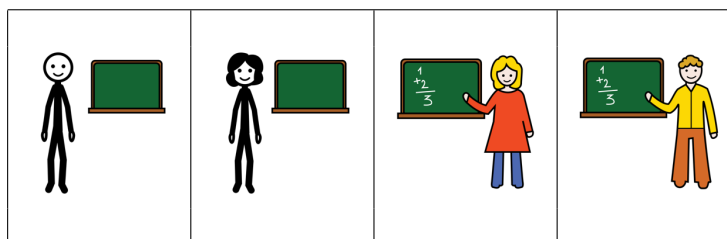


Table 3.11: Some different human referent representations for the concept 'teacher'

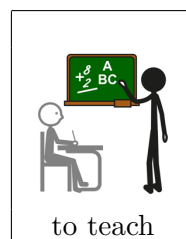


Table 3.12: Example of a concept represented only with stick figures

The different ways of depicting a person is yet another reason for multiple representations of the same concept. Two representations for humans appear to be considered by ARASAAC, a stick figure made of black lines and another more realistic representation. Each of the two representations depicts male or female, adult or child figures. This more often results in the same concept being depicted several times to cover these various representations. Examples that illustrate this property are symbols representing the concept 'teacher' (a subset is shown in Table 3.11). However, this approach is not often maintained throughout the symbol set, as for instance, the concept 'teach' which is depicted only using the stickman figure (Table 3.12). Having many alternative symbols of the same concepts can cause issues for symbol tagging, especially when a single symbol is preferred or required due to display size. Thus, a mechanism that ranks several related symbol may be required.

The same gloss may be associated with more than one symbol but with an additional modifier or object. For example 3.13 shows a symbol representing a shoe shop versus a more general symbol representing a shop and another example showing a symbol of eat versus eat dinner. This observation has several implications. For a text to symbol task, this suggests that multiple words such as 'eat dinner' can be expressed either through one symbol or two symbols one for 'eat' and another representing 'dinner'. These

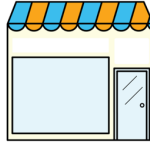


|   |  |
|---|--|
|  |  |
| Shoe Shop   | Shop   |
|  |  |
| Eat dinner  | eat  |

Table 3.13: Symbols that are very specific on the left, and that are more general on the right

multi-word compositional glosses may also be problematic when linking symbols to a lexicon since they will have no correspondence (see 2.3.2). As result, they may be accidentally left out. In situations where a limited number of symbols is preferred, the symbols with multi-word compositional glosses can be excluded.

### 3.1.5 Gloss Morphology

The majority of glosses are in their base form. However, two morphological inflections were observed. Some nouns and adjectives were inflected for gender, which was seen in the Spanish glosses since English do not inflect for gender. The gloss's gender will match the human figure in the corresponding symbol. As mentioned earlier this increases the number of symbols per concept (section 3.1.4).

Plural forms of the glosses have also been included. The associated symbol is often the same as the symbol associated with the corresponding singular form but with a additional qualifier (+S) placed on the top right corner of the symbol (Table 3.14(a)). Adding a plural marker appears to be a common practice among other symbol sets, e.g. the Widgit symbol set (Pampoulou & Detheridge, 2007). However, this increases the the number of image files with the same base content, and symbols with such a marker are not readily distinguishable. Therefore, the addition of the marker should be a dynamic process to avoid redundancy. Some exceptions have been seen in which the symbol associated with a gloss in its plural form actually illustrates more than one object (Table 3.14(b)).

Verbs are only in their base form and no morphological variations are captured. However, English verbs are often preceded by 'to', which may need to be removed to avoid issues in some tasks such as the symbol look up function.





|     |  |   |
|-----|--|---|
| (a) | <br>books | <br>book |
| (b) | <br>boys  | <br>boy  |

Table 3.14: The handling of plurals


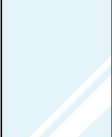


| (a)  |   | (b)  |   |
|--|---|--|---|
| <br>glass | <br>crystal, glass | <br>operation | <br>operation, addition |

Table 3.15: An example of symbols associated with irrelevant glosses

### 3.1.6 Ambiguous Glosses

Although the majority of symbols are associated with a single gloss, it is often ambiguous, having more than one possible meaning. There is often no contextual data provided with each symbol that can help in disambiguating the associated gloss. Table 3.15 shows examples of symbols sharing a gloss but with different senses; (a) two symbols sharing the gloss ‘glass’, while (b) shows two symbols sharing the gloss ‘operation’. In some cases, only one sense is covered in symbol set among many senses a word may have. For instance, ‘second’ a unit of time is not part of the symbol set, while ‘second’ as an adjective is covered (Table 3.16). Thus a gloss appearing only once in the set does not mean it is monosemous – having only one meaning.

Gloss ambiguity is a major issue that concerns symbol tagging and needs to be resolved in advance of use. Ambiguity can be resolved by manually linking symbols to an ontology (2.3.2) or, for instance, adding additional glosses and semantic tags as part of the development process. For example, the symbol for ‘operation’ in Table 3.15 has an additional gloss ‘addition’ that helps to disambiguate its sense, while the other symbol with a single gloss ‘operation’ remains ambiguous.



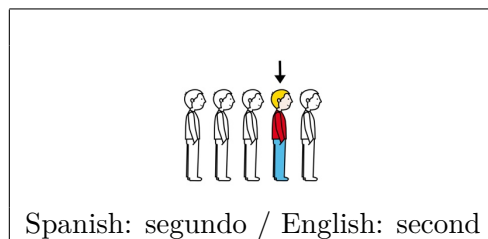


Table 3.16: Symbols associated with glosses in multiple languages

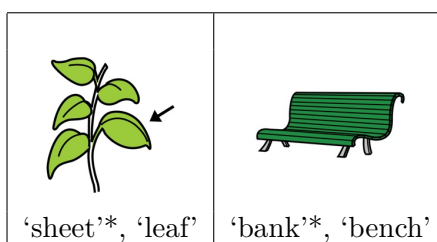


Table 3.17: Example of symbols associated with irrelevant glosses

### 3.1.7 Gloss Translations

Translations of glosses in multiple languages might be included in some symbol sets. In ARASAAC, associated glosses are mainly in Spanish, but also supplemented with translations into other languages (with various degrees of coverage). English glosses have been provided for nearly the full set. Unfortunately, a few symbols have an inaccurate English gloss that is possibly the result of automatically translating an ambiguous Spanish gloss. Two examples illustrate this issue, see Table 3.17. The first example shows a leaf symbol with the gloss ‘sheet’, which may refer to a leaf in a book, meaning ‘page’, but not to the leaves of a tree or plant as suggested by the symbol. The second example shows a bench with the gloss ‘bank’, which could be a translation inaccuracy due to the ambiguity of the Spanish word ‘banco’ that can be translated to the English words ‘bank’ or ‘bench’, depending on the intended meaning. The correct translations have been added but earlier incorrect translations remain in the database. This suggests the importance of manual translation. The availability of translations can be a valuable resource for disambiguation. However, the result is subject to the accuracy of the translations provided.

### 3.1.8 Language Coverage

Symbols for some rather advanced topics have been seen in the symbol set such as a symbol for “vertebrate” (Table 3.3). This suggests that the set might be designed to cope with concepts that would be included in an educational setting as well as general conversation.

|  |
|--|
| daddy - TV - pen - gentle - penny - off - asleep - stroller<br>- tonight - firetruck - not - peekaboo - later - night night<br>- airplane - owie/boo boo - thank you - mommy - pants -<br>people - hurry - raisin - sick - lunch - away - motorcycle -<br>shovel - cheerios - picture - patty cake - hi - naughty - play<br>pen - carrots - bump - kitty - cracker - reffridgerator - back-<br>yard - grandma - babysitter - pool - potty - shh/shush/hush<br>- sleepy - child - bunny - down - shorts - yucky - diaper -<br>puppy - bye/byebye - crib - cookie - out - fine - trash - couch<br>- pajamas - careful - home - don't - all gone - wanna/want<br>to |
|--|

Table 3.18: Words not covered by ARASAAC glosses

The number of unique glosses in ARASAAC is 10,712. However, it is difficult to assess the coverage. The American English early vocabulary list, called the MacArthur-Bates Communicative Development Inventories (MB-CDIs) (Dale & Fenson, 1996; Fenson et al., 1993)<sup>2</sup> was used to examine the coverage. The list contains 375 items, after excluding the sound effects, animal sounds, people names specific to a child, and removing redundancy. The majority of the list was covered in ARASAAC except for the 65 items shown in Table 3.18. Some of the concepts covered by the list are actually part of ARASAAC symbols but with another gloss such as dad vs. daddy, ill vs. sick. Also, ARASAAC English glosses appear to be biased towards British English; for instance, it has ‘aeroplane’ but not ‘airplane’. This may suggest the need for expanding the number of glosses per symbol by adding all possible spelling variations and synonyms. ARASAAC is already large, but a few core concepts remain absent and it is hoped that they get covered in the near future.

Additionally, lexicalized concepts vary across languages. Fellbaum and Vossen (2012) pointed out the issue of lexical gaps that exist in natural languages, and that languages differ in the concepts they lexicalise. This has caused problems when the English WordNet was used as an index to combine other wordnets developed for other languages. The author suggested that such a problem needed to be addressed by constructing an ontology of concepts that is independent of any natural language. Therefore, one needs to be aware of this fact when examining glosses that are associated with ARASAAC symbols since its first language is Spanish. For instance, the Spanish word ‘calzar’ as shown in Table 3.19 shows a concept that is lexicalised in Spanish, but not in English. These examples suggest that the lexicon covered by ARASAAC may be biased towards Spanish and could have gaps when used with other languages.

<sup>2</sup>[https://www.uh.edu/class/psychology/dcbn/research/cognitive-development/\\_docs/mcdigestures.pdf](https://www.uh.edu/class/psychology/dcbn/research/cognitive-development/_docs/mcdigestures.pdf)


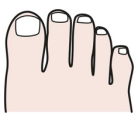
|   |  |
|---|--|
|  <p>En: to put one's shoes on<br/>Es: calzar</p> |  <p>En: fingers*<br/>Es: dedos</p> |
|---|--|

Table 3.19: Variation in lexicalised concepts between languages

## 3.2 Symbol Content Similarity

Symbols with similar graphical content have been noticed for symbols that convey semantically-related concepts (Table 3.20) such as antonymic relationships between adjectives or concepts related to the same topic. For example, symbols for ‘green’ and ‘orange’ both referring to colours are similarly drawn and both are consistent in shape, size and position. Additionally, ‘first’ and ‘second’ symbols are very similar having only slight variation; the position of the arrow and colour, while shape and position of remaining parts are the same. The symbols for ‘happy’ and ‘sad’ although the expressions are different, the shape, position and colour of the face are uniform. This is a useful property that could be exploited using computer vision techniques to find similar symbols automatically when considering symbol to text ambiguities. Knowing similar symbols and accessing their associated glosses may provide contextual information that can be exploited for disambiguation.

## 3.3 Symbol Markers

A few symbols have special elements positioned on the right top corner. These are additional markers or qualifiers added to a symbol to denote a change or additional meaning (as was described by the Widgit schema). Three markers were observed in ARASAAC: a mark indicating pain, a health/medicine mark (red cross), and a mark for plural glosses (+s), Table 3.21. Symbols with a specific marker can be identified using computer vision techniques. For instance, automatically identifying medical symbols by finding all symbols with a cross mark. Unfortunately, this was not found to be a reliable indicator since many medical concepts did not have the cross mark. An example is illustrated in Table 3.22 where one symbol is associated with a verb and the other with a noun, but both express the concept ‘surgical operation’. Two inconsistencies can be seen in this example. First, the presence of the red cross element in one symbol, while it is absent in the other. Second, is the difference in colour. This inconsistency makes it difficult to semantically identify these two symbols as similar, which may pose challenges to identifying similar symbols automatically.




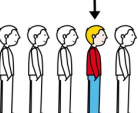


|   |   |
|---|---|
|  |  |
| Dark green  | Orange  |
|  |  |
| First   | Second  |
|  |  |
| Happy   | Sad   |

Table 3.20: Visual similarity between relevant concepts


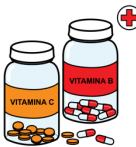

|   |   |  |
|---|---|--|
|  |  |  |
| headache  | vitamins  | wheelchairs  |

Table 3.21: Symbols having a special element or qualifier



|   |   |
|---|---|
|  |  |
| operate   | operation   |

Table 3.22: Symbols showing inconsistencies




|   |   |  |
|---|---|--|
|  |  |  |
| Saturday  | September   | Finishing line   |

Table 3.23: Example of symbols containing text





|   |   |  |   |
|---|---|--|---|
|  |  |  |  |
| race or running race  | play  | argue  | candidates  |

Table 3.24: Symbols containing human figures

3.3.1 Textual Content

Although symbols are mainly drawings, some contain textual elements embedded within the depiction. These textual elements are expressed in the developers’ local language (i.e. Spanish), and no modified representation is made for other languages. This poses a challenge when using the symbol set with other languages. Thus, translation is needed not only to the associated glosses but also the content of images. Table 3.23 shows a number of symbols containing Spanish text. However, foreign text is not the only challenge when using the symbol set with other communities; social settings, religion, and other cultural aspects need to be considered.

3.3.2 Cultural Differences in Visual Representation

It can be hard for a single symbol set to adequately address all cultural variations. The bias towards a certain community is noticeable in symbols through many aspects such as the skin tone, hair colour, and the type of clothing. In the majority of symbols, the human figure has a light skin tone and blonde or brown hair (Table 3.24). A limited number of symbols (mainly a depiction of a single person) show variations in colour (Table 3.25). Certainly, cultural differences are not limited to human figures. For example, the concept ‘ambulance’, which has several forms in the symbol set, shows variation and sensitivity of the graphical content of some concepts (Table 3.26)





|   |   |   |  |
|---|---|---|--|
|  |  |  |  |
| boy   | boy   | girl  | girl   |

Table 3.25: Some of the few symbols that show variation in colour




|   |   |  |
|---|---|--|
|  |  |  |
| Ambulance   | Ambulance   | Ambulance  |

Table 3.26: The concept ambulance informed by communities having different medical symbols due to different religious backgrounds

### 3.4 Conclusion

The ARASAAC symbol set was reviewed and a number of observations reported. The set is composed of thousands of symbols that are coloured drawings on a white background depicting concrete and abstract concepts. These drawings can be classified as icons, indices, or symbols, depending on the relationship between the drawing and the referent. Each symbol is associated with a gloss and few symbols are associated with more than one gloss.

Awareness of the gloss is a key to identify the intended meaning. However, these glosses alone are not sufficient to computationally determine the exact sense. Other information, such as word classes and translations, is available as part of the symbol set and can be useful for automatic gloss disambiguation. Additionally, similarity between related symbols was observed, which can be useful for disambiguation.

There can be multiple representations of the same concept. This may provide an alternative representation that can be more relevant to the user or the context. However, it also creates an issue for some tasks, such as tagging text with symbols when only one symbol needs to be selected. Additionally, there is no way to computationally identify whether two symbols that share the same gloss, also share the same meaning. In some cases, the visual similarity can be an indication that they express the same meaning. However, absence of similarity between two symbols does not imply that they express different meanings.

Symbol sets may be biased towards a certain community. However, ARASAAC appears to be addressing this matter by adding several representations of some

---

concepts, but these are limited so far. The set appears to provide sufficient coverage to satisfy this research. Others have aimed to fill in this gap by contributing culturally-sensitive symbols (Draffan, Wald, Halabi, Kadous, et al., [2015](#)), which can be used alongside existing symbol sets.

Knowing these characteristics is important background for the methodology, which is the focus of the next chapter.





## Chapter 4

# Methodology

This research is interested in the task of tagging text with pictographic communication symbols, and translating symbols to text from a language processing point of view. Text tagging and machine translation are often approached using data based methods, so the focus of this research is to address the absence of training data needed for the development of a system that translates between communication symbols and text in general, and Modern Standard Arabic (MSA) text in particular. The methodology proposed in this chapter aims to answer the research question “Given the lack of a manually tagged corpus, how can one automatically tag an Arabic corpus with relevant communication symbols using a multilingual parallel corpus, that would be suitable for building a system that translates text to symbols and symbols to text using data based machine learning techniques?”. The approach followed avoids the need for manual tagging by making use of a multi-lingual parallel corpus to determine the likely tags, mainly symbols and lemmas, which are essential for the translation task.

The chapter begins by investigating the task of translating symbols to text and text to symbols and what training data for each direction requires (section 4.1). It then discusses the process of preparing the selected symbol set before carrying out the tagging process (section 4.2). Next, it will describe the selected corpus, the preprocessing steps that were carried out, an experiment that examines the relevancy of that corpus, and a justification of such a choice (section 4.3). The chapter then moves on to detail the approach followed in creating training data that involves tagging with lemmas and symbols, as well as adding diacritics (section 4.4). Next, the visual content of pictographic symbols is considered and the potential of knowing similar symbols for gloss disambiguation is discussed, and then context awareness to further improve tagging accuracy, and the description of an experiment that gave insight into this matter (section 4.5). The chapter ends by describing the evaluation of experiments carried out (section 4.6).

## 4.1 Translation

A user may select a sequence of symbols and the task is to generate the corresponding full textual message that conveys or approximates the same meaning. A user may also receive a textual message and needs corresponding pictographic symbols to understand it. Translating to and from symbols can be approached as a machine translation (MT) problem.

MT is often handled using data based methods requiring a parallel corpus covering the source and target languages. Thus, the corpus required by this study is MSA text and AAC pictographic communication symbols. However, such a resource does not exist at present and there only appear to be parallel corpora for pairs of natural languages such as MSA and English. Nevertheless, a parallel corpus could be created by manually translating an MSA corpus to pictographic symbols. However, the corpus needs to be large to reach acceptable performance. For instance, the SemCOR corpus, which has been manually labelled with WordNet senses, was found to be too small to make use of when considering statistical methods (Miller et al., 1993). Manually producing a large corpus is expensive in terms of time and effort, as well as having the limitation of only working with a specific symbol set.

Instead, the method chosen is an automatic tagging process, where the semantic ambiguity of words in context is addressed. This ambiguity needs to be resolved or minimised in order to automatically determine relevant symbols in addition to other linguistic data for a given text. However, training data needed for translating from text to symbols is not necessarily the same as the data needed for translating symbols to text, as discussed in the following section.

### 4.1.1 Symbol to Text

As mentioned earlier, translating from symbols to text can be approached as an MT problem. Ideally, training data should be made up of a large number of pairs, a sequence of pictographic symbols and their corresponding fully-formed text. In practice, the task does not require awareness of the pictographic symbols and, instead, the associated glosses might be sufficient. Ignoring the visual element is beneficial since it overcomes the lack of a symbol corpus. The resulting system will not be limited to the current symbol set, and additional symbols can be added easily. Therefore, the symbol to text problem was approached independently of any symbol set by using glosses instead of symbols as the input to the translator component. This approach has been followed by other researchers as well, but they did not highlight the impact of ignoring the pictographic representations of symbols (Sevens et al., 2015b; Vandeghinste et al., 2018; Waller & Jack, 2002).

As a result, the task of symbol to text becomes a translation not from pictographic symbols to text but rather from associated glosses to fully-formed text. The goal at this point is to simulate such an input. As discussed in chapter 2.5.5, the majority of glosses are single words in lemma form. Thus, the input message is likely to lack morphological clues, as restricted by the symbol set or due to the user's limited literacy skills. Furthermore, pictographic symbols may be associated with more than one gloss, and the relevant gloss needs to be determined given the context. There may also be missing function words that are necessary for generating fluent text, and the order of symbols – and in turn associated symbols – may not be consistent with the target language (Sutton et al., 2000). Also, the input symbol message tends to be shorter than similar face-to-face spoken utterances. Therefore, although each symbol may be labelled with its corresponding traditional orthography, the sequence of labels is far from a fluent sentence that can be spoken by a speech synthesizer, especially for morphologically rich languages.

Characteristics of the symbol input message are summarised as follows.

1. The input are words in their lemma form with no morphological markers.
2. A symbol may be associated with more than one lemma.
3. Function words may not be part of the symbol message.
4. Word order may not be as expected in the target language.
5. Messages are expected to be short, covering common vocabulary.

The simulated input needs to have these same characteristics as the symbol input. The first dictates that lemmatisation is a key. The second property can be hard to achieve since symbol sets vary in how glosses sharing the same pictographic symbol are related. For example ARASAAC has co-occurring glosses that are often synonyms, or morphologically- and semantically-related forms. The third characteristic can be approximated by occasionally performing local swaps in the training data. The fourth aspect can be met by first designating a list of function words, which contribute to the syntax rather than the meaning, then removing them from the input. The fifth one is met by carefully choosing a relevant corpus (section 4.3) which is conversational and whose majority of lines are short. It must also have content that spans various genres, which ensures that a common vocabulary is covered. Such input can be referred to as ‘telegraphic’ text. Telegraphic is a term used in psychology to describe speech “consisting of essential content words but lacking function words, esp. as seen in early language acquisition or in a mental disorder”(OED). Thus, the task becomes translating from telegraphic text into fully-formed text, which needs to have diacritics in place to ensure accurate speech synthesis.

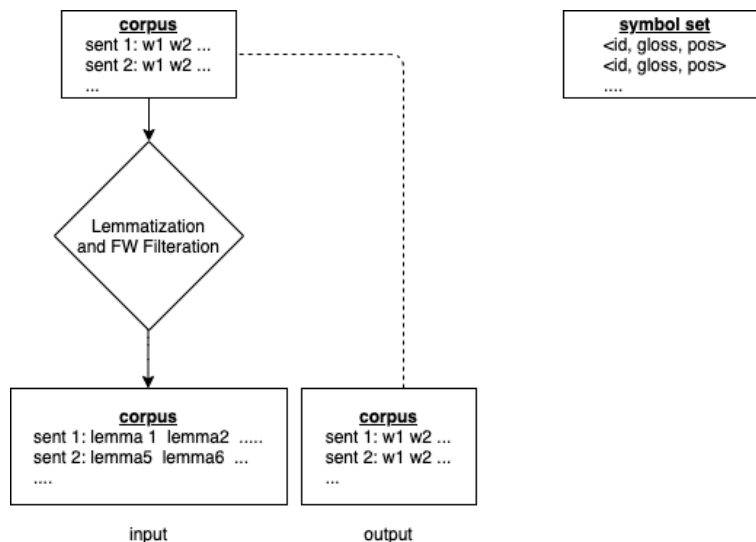


Figure 4.1: The process of creating training data for the symbol to text task

Transforming a relevant monolingual corpus, such as the Arabic subtitle corpus, can enable training data for this task to be created. However, determining the lemma, which is the core task in the transformation, can be a challenge for any Arabic text in general, and Arabic subtitles in particular. Details of the approach followed in creating the training data are given in section 4.4. Once the telegraphic and fully-formed parallel corpus is created, a MT framework can be used for the translation. Sequence to sequence tasks have been the traditional approach using statistical models such as phrase-based MT. Statistical methods have more recently been replaced by neural-based methods, which have shown significant improvements over statistical methods. This research uses a basic LSTM model, as well as the traditional statistical method, to gain some insights into the difficulty of such a task by measuring its performance by calculating the BLEU score (Papineni et al., 2002). No attempt has been made to determine the best possible hyperparameter or the best neural-based architecture.

#### 4.1.2 Text to Symbols

Unlike symbol to text, glosses are not adequate for the text to symbol task. This is due to the fact that knowing the gloss is not sufficient to determine the relevant pictographic symbol, since a gloss may be associated with more than one symbol expressing unrelated meanings due to word ambiguity. This issue can be addressed by either creating a symbol corpus by manually tagging a text corpus with symbols, or by using or developing a method that disambiguates associated glosses as well as words in the source text. The former solution is expensive to acquire and the resulting resource will be only useful for tasks that use exactly the same symbol set. The latter can be carried out computationally and is not limited to a specific symbol set. Researchers

have not created training data for such a task but rather made use of existing word sense disambiguation tools to figure out the relevant symbol (section 2.3.3).

Before discussing possible solutions to the problem of word ambiguity, it is better to have an idea of how significant is ambiguity in a spoken language vocabulary, such as English. This can be estimated by counting how many lexical items in a dictionary have more than one meaning. WordNet can be used to make such an estimate. The percentage of ambiguous words (i.e. having more than one sense in WordNet) was found to be 21.7% of the vocabulary covered. However, the majority of unambiguous words tend to be domain specific such as ‘paediatrics’ and other rarely used words. Frequently used words have more senses (Edmonds, 2005). Thus, since communication symbol sets cover the more frequent words, it is expected that the percentage of ambiguous words is higher than a more comprehensive English spoken vocabulary. The set of glosses covered by the current ARASAAC collection was counted to be 9177. However, only 6093 of these have a potential match in WordNet 3.0. In the overlapping vocabulary, the percentage of ambiguous words is 63.2% . This shows that ambiguity is a serious problem for a text to symbol task and needs to be tackled. This percentage can be considered an overestimation, since senses provided by WordNet are fine grained.

Wilks and Stevenson (1998) examined the potential of using part of speech (POS) tags as the basis of disambiguation. They found that a coarse-grained sense inventory was sufficient for disambiguating the majority of cases. However, for WordNet, the number of ambiguous overlapping words was reduced to 55.3% when counting lexical items that have more than one meaning in the same part of speech class. Thus, with symbol glosses, using POS tags may not be sufficient, since ambiguity will occur even within the same POS class. For example, the word ‘spring’ (as a noun) was a gloss seen in the selected symbol set with three different pictures/meanings, namely ‘season’, ‘metal coil’, and ‘ground water’. Therefore, awareness of the context is essential for resolving ambiguity.

Automatic text disambiguation (section 2.5.1) involving pictographics was carried out by first tagging symbol glosses with a pre-existing sense tag inventory (that was independent of any symbol set), and then a WSD tool was used to tag text with the same sense tags used with the glosses. However, for MSA this is difficult since the Arabic WordNet (Elkateb et al., 2006; Rodríguez et al., 2008) does not offer good coverage (48% of core concepts are covered (Bond & Foster, 2013)). Also, there is no robust WSD tool available for Arabic. Also notice that the English WordNet has not been updated since 2011<sup>1</sup> and is lacking some recent concepts such as ‘tablet’ (meaning a general purpose portable computer). Additionally, conversational text is often short and might not provide enough context needed for reliable disambiguation using pre-existing tools. The data used to train WSD classifiers may have significant

<sup>1</sup><https://wordnet.princeton.edu/news-0>

differences from conversational text. Lastly, the existing WSD sense inventory, such as WordNet, only covers content words which may be problematic because some of the glosses associated with symbols are closed class words.

Instead of depending on a sense inventory produced by lexicographers, senses can be identified through unsupervised methods. Such a solution is appealing, especially as Arabic does not have a sense inventory with good coverage. However, discovered senses will not have meaningful tags, so there is a need for human intervention to understand the meaning of an identified sense and decide relevant symbols which is a time consuming task. Additionally, senses that do not have a clear topical characteristic are difficult to identify using unsupervised methods (Schütze, 1998). Therefore, such an approach was not be considered in this study.

Alternatively, disambiguation needed for symbol tagging can be based on multilingual information. Using other languages for disambiguation is the method that was felt to be truly advantageous, since it does not rely on an external sense inventory. It also does not require any tagging for the symbol set given the availability of multilingual glosses, which is especially useful when several symbol sets need to be covered. It only requires the availability of a parallel corpus for a pair of languages that are also covered by symbol glosses or two parallel corpora covering three languages covering three languages, two of which are covered in the associated gloss and the third is the target language.

The use of translations in another language is especially beneficial when the context does not provide sufficient information for determining the correct sense when creating training data for text to symbol task. For example “I can see a bat” may be a discussion about a flying mammal or a cricket bat. Vague contexts have been found to be a source of error for human taggers (Palmer et al., 2007) but when looking at translations into other languages, the lack of clarity of meaning can often be resolved. However, some words are cross-linguistically ambiguous, such as ‘king’ and ‘operation’ in Spanish, Arabic and English where they have many shared meanings. A method to overcome these occasional happenings could be addressed by manually augmenting those few symbols with contextual data relevant to their meaning, and choosing the best match based on overlap between the symbol’s contextual data and the target word’s contextual data. Such a method is discussed but not carried out as part of this research. The detailed approach of tagging data on the basis of bilingual data is covered in section 4.4.3. The resulting corpus can be a useful baseline and has been used as the basis for observations made in this research.

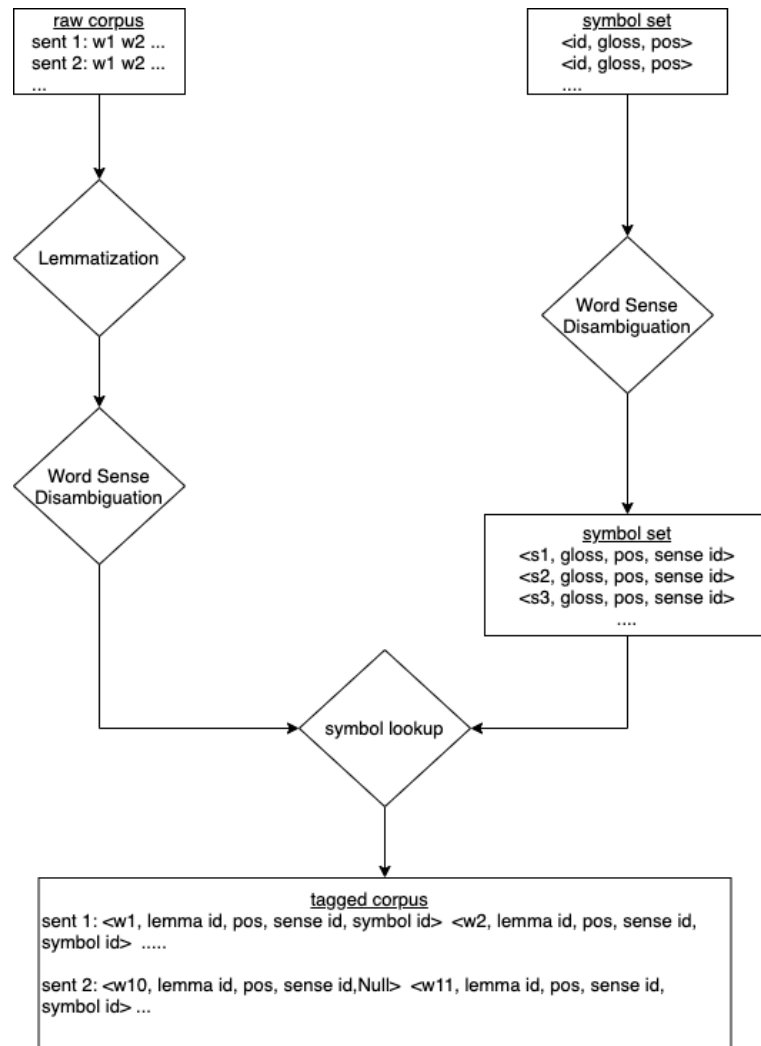


Figure 4.2: The process of automatically tagging the corpus with symbols

## 4.2 The Pictographic Symbol Component

A communication pictographic symbol set, namely, ARASAAC was selected as a concrete example of pictographic communication symbols for this research. Symbols are pictures designed to convey a concept and are often associated with a label or gloss that expresses the same meaning using traditional orthography. The selected set was reviewed and examined by looking at the content and glosses (chapter 3), which arguably can be generalised to other pictographic communication symbol sets. The downloaded set contained 12 000 image files. Symbols were provided with glosses and their corresponding word classes. The symbols obtained did not include any semantic data, such as their usage domain or the semantic class of the depicted concept.

Once the symbol set was obtained, it was checked for redundancy, i.e. same image with a different file name. This was achieved by computing the hash code for each image, with the MD5 hashing algorithm (Rivest, 1992) and using the resulting hash code as

an identifier. Such a hashing algorithm is sensitive to the slightest change (e.g. one pixel difference), and thus it can identify identical image files. By comparing hash codes, only five images in the set were found to be redundant, and were discarded from all subsequent tasks; their associated glosses were merged with the kept image.

The majority of ARASAAC glosses are provided in Spanish and English. The first step when adopting a pre-existing symbol set is to translate associated glosses to the chosen target language (e.g. Arabic) – if not already available – given some other source language. There are many ways to translate associated glosses but not all are of the same trustworthiness. Assuming that a set of translated glosses in addition to gold standard glosses in the same language are available, measuring how far a resulting set of items (or translated glosses) is from the ideal set of items (or gold standard glosses) is usually made by computing recall and precision, and is the standard performance metric in information retrieval and classification problems.

Recall computes the number of items in the ideal set that are included in the resulting set as a proportion of the total number of items in the ideal set, while precision computes the number of ideal items included as a proportion of the total number of items in the resulting set (Cleverdon, 1967). For instance, an automatic translation of associated glosses using a dictionary, either an open-source conventional dictionary or a dictionary extracted from a large parallel corpus, will tag symbols with probably all possible translations. This approach will result in a high recall score when retrieving symbols based on their associated glosses. However, the precision score may be low due to source ambiguity. This introduced false translations of the meaning conveyed by the symbol, which resulted in unrelated symbols being included in the retrieved set.

Precision is extremely important in a text to symbol system because tagging words with unrelated symbols will be confusing to the user and may defeat the purpose of symbol tagging. As a result, manual translation would be the best choice to ensure sense agreement between the associated gloss and the target word in the text.

However, due to cost constraints, the approach followed in this research used an automatic translation approach but minimised ambiguity by including bilingual data tagged into Arabic (detailed in section 4.4.3). This method allows the entire symbol set to be translated, apart from rare glosses not found in alignments extracted from the selected corpus.

Some symbols include linguistic markers, e.g. ‘plus s’ marker for plural forms (chapter 3). However, the use of these markers results in unnecessary repetition in the symbol set. Yet these markers are usually consistent across symbols in terms of size, colour and position. As a result, template matching, a computer vision technique (Goshtasby et al., 1984), was used to identify symbols that had these markers. It searches an image to determine the location of a smaller query image (in this case the marker). It



assumes that the query image exists and searches for the best match. However, it can be difficult to determine the non-existence of a given sub-image.

There are many methods that measure the similarity between the query image and a certain area in the main image. Among these methods is normalised cross correlation coefficient and an efficient implementation has been proposed (Briechle & Hanebeck, 2001; Lewis, 1995). In its basic form, the similarity is computed from single pixel intensities by comparing two two-dimensional matrices. It is suitable when an exact match is expected with a variation in lighting, which seems suitable for this task. The process made use of implementations provided by the Open Source Computer Vision Library (OpenCV)<sup>2</sup>. The existence of the ‘+s’ plural marker, given a query image containing the marker, was determined in two phases. First, the best match was determined using normalised cross correlation and accepted if a score exceeded an empirically specified threshold. Next, a match was only accepted if the spatial distance between the expected location and the discovered location was below an empirically set threshold. The approach was useful in identifying symbols with the plural marker. Identified symbols were discarded for two reasons. First, apart from the marker, their visual content was exactly the same as their corresponding singular symbol and thus could be automatically regenerated. Secondly, the non-lemma forms cause problems when translating associated glosses into other languages using conventional dictionaries since these forms often do not exist as an independent entry, but rather part of an entry headed with the singular form (lemma).

### 4.3 Corpus

The availability of large amounts of training data is a prerequisite for using machine learning methods. A large relevant Arabic corpus was therefore needed. However, a genuine AAC corpus is not available, even for English (Kane et al., 2017) and researchers often use alternatives (section 2.6.1). A corpus that is conversational and is not domain specific was considered to be a good alternative. Forchini (2012) studied a corpus of movie subtitles and concluded that they could be an appropriate representation of face-to-face conversations. Fortunately, such a resource has been made available by Lison and Tiedemann (2016). Therefore, using subtitles for AAC appears to be the best available solution. Lison and Tiedemann (2016) collected data for many languages, and created a parallel corpus for many language pairs. The content was provided in various formats and subtitles could be in separate files, or merged into a single file representing one side of a language pair <sup>3</sup>.

The Arabic and English corpus is among the available pairs of languages covered. A parallel corpus is a valuable resource for disambiguation. In this case English was

<sup>2</sup><https://opencv.org/>

<sup>3</sup><https://opus.nlpl.eu/OpenSubtitles-v2018.php>

chosen to create a parallel corpus with Arabic due to the wide availability of NLP tools that can be used, as well as the number of symbol set lexicons associated with this language (section 4.4). The Spanish and Arabic parallel corpus was also used for the purpose of gloss disambiguation when tagging the corpus with symbols (section 4.4.3).

The form of the chosen corpus is a collection of files, each representing an individual unit either a movie or an episode of a TV show. Lines in each file are listed in their chronological order. The Arabic files downloaded, which were aligned to English, numbered 40,977. The quality of these files varied in terms of the fluency, alignment precision, and translation quality. Many appear to have been produced by people, some seemed to have been generated by machine, while others were partially translated and mixed with English text. A few of the files were not in MSA, but rather in a local dialect. However, the majority of files seemed to be acceptable and were merged into a pair of files. The Spanish Arabic parallel corpus used was the merged form, which is a single large file for each language. The size of these parallel corpora was slightly smaller than the English Arabic corpus.

#### 4.3.1 Preprocessing

The corpora needed pre-processing before making use of them and two tasks were undertaken. Initially, a small number of files were filtered out based on scores using a n-gram language model and an alignment model trained on the same data. However, later this step was skipped because the majority appeared to be linguistically acceptable and the few outlying files did not affect the knowledge extracted.

The second task looked at the contents of the corpus in general. The corpus was already aligned at the sentence level and the English part was already tokenised.<sup>4</sup> The Arabic side was pre-processed by removing all diacritics which occasionally appeared as well as the Arabic Tatweel character (or Kashida) which is normally used for formatting text or for word emphasis in Arabic script.

Next, tokenisation of the Arabic text needed to be carried out as it was not tokenised. Arabic text does not use hyphenated words and nothing similar to apostrophes. A simple tokenisation was employed which inserted a space between any Arabic letter and a non-Arabic letter. However, Arabic text may have instances of missing spaces between two words, and this issue was handled later as part of the disambiguation process.

Normalisation was also needed since Arabic text writers may alternate between several forms of certain letters and not follow the expected written form of a word. These

---

<sup>4</sup>However, the tokenisation approach did not match that of the POS tagger used in this research, e.g. the abbreviated negative particle – n't. This was handled by pre-processing the English side with a script that detokenized the few forms involving an apostrophe.

spelling variants were handled by converting all Alef variations into bare Alef. Other researchers take a step further and conflate final Yaa' and Alif Maqsoura (Maamouri et al., 2004), and final Haa' and Taa' Marboutahn. However, these alterations are believed to be less helpful compared to Alef, especially in a large corpus, and thus was avoided as such a step would unnecessarily increase word ambiguity.

A number of lines contained foreign text and not the expected language. Lines were only kept if they contained at least three concatenated alphabetical characters (of the same language) to filter out noise. In addition, the English side was tagged with a POS tagger (Toutanova et al., 2003). The decision to tag the English side was to reduce the space of possible meanings a word may have (Wilks & Stevenson, 1998), and to be able to lemmatise English words correctly.

The Spanish side of the Arabic parallel corpus was tokenised by simply adding a space between a word character and a non-word character. Also, some lines were left out which contained characters that did not match the expected language.

#### 4.3.2 Domain Relevance

An experiment was carried out to gain confidence in the relevance of the selected corpus to the target domain. The experiment examined one language (English), since results can arguably be generalised to other translations of the same content, such as Arabic. The idea was to estimate how close the selected corpus is to a sample that represented the domain of interest. The computation was repeated with a few other corpora to compare the results. A list of generic messages or small talk (section 2.6.1) suggested by (Beukelman & Gutmann, 1999) for users with amyotrophic lateral sclerosis was used as a domain sample (Appendix A). The list contained sentences allowing a more discriminative comparison, as opposed to single words (e.g. a basic core vocabulary). The final tokenised sample contained 85 sentences and a total of 440 tokens.

The estimation of the closeness of each corpus and the selected sample was made by calculating the perplexity score. Perplexity of a corpus with respect to a test sample of text is estimated using a language model derived from the same corpus. Perplexity has been used in many tasks. For instance, it can be used to compare various language models given the same training data (Chelba et al., 2014). It has also been used to identify domain specific data (Lin et al., 1997), and to determine the best corpus for a given task (Lembersky et al., 2012).

In order to determine the perplexity score, a language model for each of the examined corpora needs to be built. Five corpora were used in this experiment: a corpus of subtitles that have Arabic translations, the whole subtitle corpus, a corpus of United Nations Documents (Eisele & Chen, 2010), the Brown Corpus (Francis & Kucera,

| Training data   | Tokens     | Perplexity |
|---|------------|------------|
| Subtitles (extracted from the English Arabic alignment) | 3234796386 | 26.28061   |
| Subtitles (English)                                     | 256061975  | 25.06272   |
| UN corpus   | 455031530  | 269.2836   |
| Brown Corpus  | 1119633    | 263.0967   |
| Reuters-21578   | 1724355    | 550.9957   |

Table 4.1: Language model evaluation results

1979), and a corpus of 10,788K newswire articles from the Reuters-21578 collection (Lewis, 1997).

This resulted in five language models being built and perplexity calculated for the Beukelman and Gutmann list using the SRILM toolkit (Stolcke, 2002; Stolcke et al., 2011). All language models were trigram models with interpolated modified Kneser-Ney smoothing. Perplexity results are shown in Table 4.1. Both subtitle corpora achieved a significantly lower perplexity compared to the others. This suggests that subtitles are much closer to the AAC domain than other corpora. The scores also show how sensitive language models are to domain variation (Rosenfeld, 2000). The size of the corpus was also a factor that affected perplexity. The complete English corpora had a slightly lower perplexity compared to a subset of the same corpora.

## 4.4 Generating Training Data

Training for both symbol to text as well as text to symbol tasks requires some form of disambiguation. The former requires disambiguation to determine the lemma as well as the diacritics, while the latter needs disambiguation to determine the pictographic symbol in addition to the lemma.

The first main task is tagging the text with lemmas and subsequently other linguistic data (section 4.4.1). This task bases disambiguation on the Arabic/English parallel corpus. The outcome is used to drive a subsequent parallel corpus of telegraphic and fully-formed text aimed at the symbol to text task (section 4.4.2). The second main task is to identify the relevant symbol for a given pair of English and Arabic words that have been aligned, as well as translating associated glosses (section 4.4.3). For disambiguation, this task makes use of information extracted from the two bilingual corpora.

The following is the list of steps carried out to create training data, with a brief description.

1. Align words in the English/Arabic Parallel corpus using MGIZA++
2. Tag the English side with POS.

| English Verb | No of surface forms | Equivalent Arabic verb | No of surface forms |
|--------------|---------------------|------------------------|---------------------|
| say          | 5                   | قال                    | 347                 |
| go           | 5                   | ذهب                    | 163                 |
| read         | 3                   | قرأ                    | 247                 |

Table 4.2: The number of Arabic surface forms compared with English Arabic forms of a single root and appearing at least five times in the subtitles corpus

3. Extract word to word translation equivalent model (English words have their POS tags).
4. Deduce a ranked list of lemmas for each English word form.
5. Tag the Arabic side with the top applicable lemma, given its equivalent English word, and subsequently work out other morphological analyses and tag each word with its morphology.
6. Align words in the Spanish/Arabic Parallel corpus using MGIZA++
7. Extract word to word translation equivalent model.
8. Tag the Arabic side with symbols based on extracted translation equivalent models and associated glosses.

#### 4.4.1 Corpus Lemmatization and Vocalisation

Defining the base word form or lemmatization is considered a crucial part of the methodology in order to overcome the Arabic language's rich morphology and agglutination. Arabic has a larger vocabulary compared to an English translation equivalent vocabulary. For example, the number of forms corresponding to verbs, such as 'say', 'go' and 'read' when lemmatized are small in comparison to the number of Arabic forms, as shown in Table 4.2

Lemmatization is needed to avoid the need to tag symbols with all possible surface forms a word may have, which results in large lists that are hard to review and maintain. For example, in supervised WSD tasks, senses are assigned to lemma forms and WSD taggers rely on the external POS taggers and lemmatisers (e.g. (Zhong & Ng, 2010)). Similarly, symbols are also tagged with lemmas<sup>5</sup> and as a result require tokens to be lemmatised. Extracting the lemma as well as associated pronouns and awareness about a null subject is an essential step to determine the relevant symbol, or to simulate symbol input message, to create relevant data for machine learning. Arabic lemmatization is challenging due to the high degree of ambiguity, so the method has depended on English equivalent translation for this task.

<sup>5</sup>Although some plural glosses have been seen in the ARASAAC symbol

Awareness of the context is necessary to determine the relevant lemma whenever more than one lemma is applicable to a given surface form. This disambiguation process is similar to WSD but with lemmas (or morphological analysis) rather than senses (or meanings). MADAMIRA (Pasha et al., 2014) is an Arabic morphological tool that performs disambiguation to determine the likely morphological analysis, including lemma identification. The tool was trained on the Penn Arabic Treebank (1, 2 and 3) (Maamouri et al., 2004), which is an annotated corpus that is mainly a collection of Arabic newswire articles. Such an external tool was not used due domain mismatch and the availability of corresponding translation equivalents from English words which can be used to resolve ambiguity (especially for open class words) in conversational text.

The morphological analysis disambiguation approach chosen was based on word to word translation equivalents extracted from word alignments. The disambiguation process also made use of bilingual dictionaries and a rule-based morphological analyser. The disambiguation occurred in two stages. The first stage took Arabic/English word pairs out of context, based on the extracted word alignment and bilingual dictionaries. The outcome of this stage was a ranked list of lemma hypotheses for a given English word. The second stage used the outcome of the first stage to determine the likely analysis (when there is more than one analysis per lemma) to figure out whether the verb is in first, second or third person. This is achieved by making use of local contexts in both languages, i.e. the words surrounding the word that needs to be disambiguated.

#### 4.4.1.1 Out of Context Lemma Disambiguation

The disambiguation process depends mainly on knowing the English equivalent word as well as other Arabic word forms aligned to that same English word. Word correspondence between English and Arabic in some corpus is essential for disambiguation. Thus, alignment between words in each pair of sentences is needed. Tools are available for determining word alignments given a raw parallel corpus.

GIZA++ is a toolkit designed to generate word-based translation models from a parallel corpus. Word-based translation models are an essential part of the statistical machine translation (SMT) framework first proposed by (Brown et al., 1993). GIZA++ is the most widely-used toolkit to extract word alignments and implements IBM models (Brown et al., 1993) with additional improvements (Och & Ney, 2000). In this research, MGIZA++ (Gao & Vogel, 2008) was used, which is a multi-thread implementation of GIZA++, made to speed up the extraction (Al-Onaizan et al., 1999; Och & Ney, 2000, 2003). It generates several files, but only word alignments produced by IBM model 3 were used (Brown et al., 1993). By default, the aligner ignores sentences with source to target length ratio above a threshold (which was set at 9). The English side chosen for the aligner was the untagged tokenised text, to reduce sparsity.

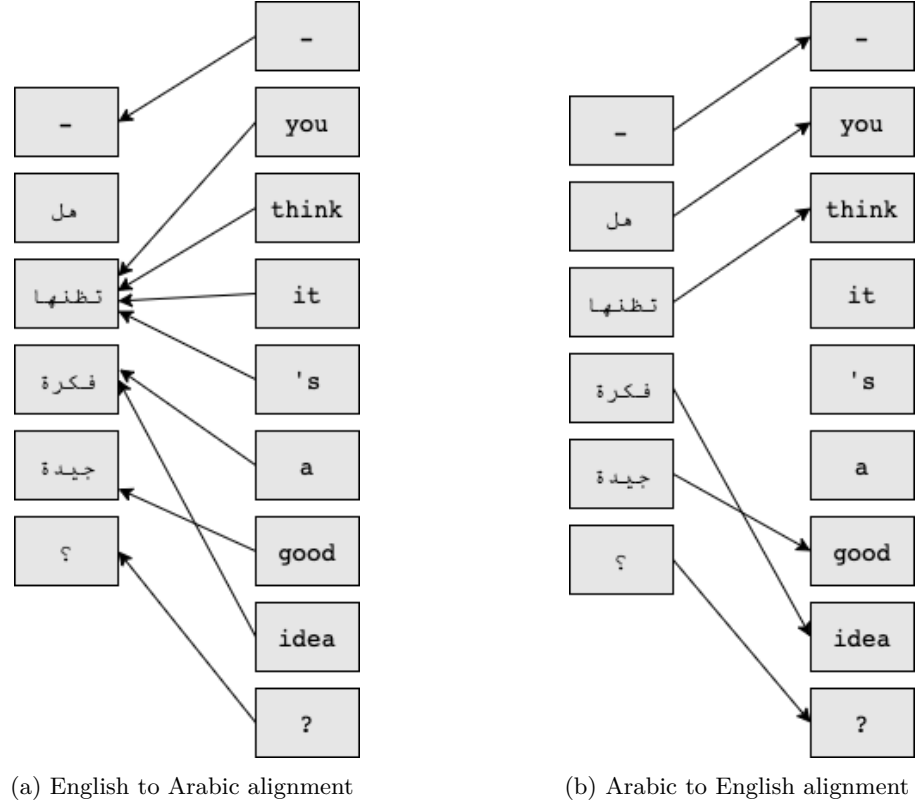


Figure 4.3: Example of directional word alignment for the following sentences:

English : - you think it's a good idea ?

Arabic : - هل تظنها فكرة جيدة ؟

IBM models extract one-to-many alignments: each word in language  $x$  is aligned to at most one word in language  $y$ . Consequently, a word in  $y$  might be aligned to no words, one word, or many words in  $x$ . Thus, the model generates directional alignments. The parallel corpus was aligned in both directions. Figure 4.3 demonstrates the two-directional alignments for a pair of sentences extracted from the corpus. A common practice in SMT is to extract alignments in both directions and perform symmetrisation between them (Koehn et al., 2003; Och & Ney, 2003; Och et al., 1999). The intersection of alignments is one symmetrisation approach. Another common approach is ‘diag-and’ which starts with the intersection alignments and expands (Koehn et al., 2003). The choice of alignment, whether directional or any of the different types of symmetrisation, depends on the task. For instance, the intersection of directional alignment will leave many of the function words unaligned, which can be a desirable property depending on the task.

From these word alignments, an English to Arabic word-based translation model can be extracted. Such a model provides translations for English surface words. Translations are Arabic surface forms, which are scored to reflect how often a translation occurred throughout the corpus. Translations will normally contain irrelevant words, but these will have low scores if the corpus is large enough and is provided with good translation. The English side was tagged with POS and

lemmatised from their POS to conflate English inflections, and merge their corresponding Arabic lists in the word-based translation model extracted. This merge will result in more Arabic forms per list, which helps in finding the likely lemma. The POS tag will provide some form of disambiguation on the English side.

The goal at this point is to determine the Arabic vocalised lemma for each pair of Arabic and English words that have been aligned to each other, better than by chance. A single Arabic surface form can have several possible lemmas because stem boundaries are not obvious and diacritics are missing, so there can be many possibilities. But, the actual lemma needs to be determined. The set of lemma hypotheses are generated by a rule-based analyser (section 4.4.1.3) with an embedded dictionary.

One way of identifying the actual lemma is to use additional data associated with each hypothesised lemma, provided by the dictionary embedded with the rule-based analyser. Assuming each lemma is associated with an English definition, the ambiguity can be resolved by comparing the corresponding English word (in the word-based translation model) and English glosses associated with the lemma hypothesis, and finding a match between them. However, a match does not need to be exact but flexible, by considering possible synonyms from a dictionary and similar words based on vector-spaced models (Turney & Pantel, 2010).

However, the English textual comparison is not always sufficient in determining the likely lemma when no match is found or more than one match is found. For example, for the ambiguous word (مدرسة), two lemmas are possible, i.e. (مَدْرَسَة) and (مُدرِّسَة) the former meaning school while the latter meaning teacher (feminine form) which are semantically related concepts (education). As a result, the actual lemma may still not be clear since they might all share the same word, such as school or education. Further evidence can be obtained by examining the list of other Arabic surface forms that have been aligned to the same English word. Such a list might include other surface forms that can result from one lemma through inflectional morphology, but is not produced by the other. For example, the plural form (مدارس) is only applicable to the first while the masculine form (مدرس) and its plural forms (مدرسون - مدرسين) is only applicable to the second. Thus, counting the number of surface forms per hypothesised lemma is used to resolve ambiguity.

Furthermore, when ranking lemma hypotheses for verbs, other forms that are related to the hypothesised verb through derivational morphology, such as verbal nouns and active participle (often aligned to the same English verb), are good indicators to the likely lemma, especially when two verbs share the same normalised inflectional forms and are only distinguished by their diacritics which are absent, e.g. ذَهَبَ and ذَهَّبَ. Therefore, ranking lemma verbs benefits from counting possible inflectional forms as well as nouns that can be generated by knowing the verb's template.



In some cases it is hard to determine the exact lemma when several nominal forms are generated from the same verb (i.e. verbal noun, active participle or passive participle) and share the same normalised form. For example, eat (the verb), eating (the verbal noun), eater (active participle), and the causative form of eat (Table 4.3) all have the same normalised form. Using lemma counts will prioritise verbs over other POS classes due to the large number of verb inflections. Using verb template counts helps in ruling out the causative form. However, choosing between two nouns sharing the same normalised form that belongs to the same verb template in a list, is a challenge and hard to address; fortunately this is not a typical case. This ambiguity was dealt with by favouring verbal nouns when aligned to an English verb or a noun ending with -ing, and passive participle when the English word is a passive participle, and an active participle otherwise.

There are other clues that can be used to rank lemmas. For instance, both the English word and its corresponding lemma should both be proper nouns or neither should be. Lemmas that are verbs are also unlikely if no other inflections are aligned to the same English word. Another strong clue is the probability of the bare lemma as a translation for the corresponding English word, since bare lemmas are often more probable than lemmas with attachments, which is strong evidence for all POS classes apart from verbs in which other tenses (e.g. imperfect form) might be more probable. Further evidence can be obtained by examining the number of observed surface forms to that expected. It is important to calculate a ratio rather than a count, since the space of surface forms per lemma varies greatly (e.g. verbs compared to proper nouns).

If the ambiguity is still not resolved, the English stem may provide some evidence by counting the number of surface forms that share the same template and root as the hypothesised lemma that has been aligned to the same English word stem. Additionally, English definitions associated with hypothesised lemmas may give a clue. For instance, when an English token (excluding function words) exists with definitions associated with two or more lemmas that do not share the same root, it is unlikely to be a coincidence but rather might be evidence that such a token is common because it is related to the actual meaning, and subsequently associated lemmas should be favoured. If everything fails, a final resolution is to select the most probable vocalised word according to some external resource.

Overall, lemma hypotheses for a specific English word are ranked by score to determine the likely lemma among several possible hypotheses. The order of these scores matters. The final ordered list of scores is shown in 4.4. Using this rank, all words that have been aligned to that English word will be tagged with the highest applicable lemma.

So far, it has been assumed that all alignments will be considered. However, there may be pairs that are incorrect due to issues with the alignment. This can be due to the translation process, or to words that appear infrequently that are not semantically

| Full form | Description                                     | Template form | Normalised form |
|-----------|---|---------------|-----------------|
| أَكَلَ    | verb (eat) <sup>6</sup>                         | Form I        | أكل             |
| أَكْلٌ    | verbal noun of أَكَلَ                           | Form I        |                 |
| أَكِلٌ    | active participle of أَكَلَ                     | Form I        |                 |
| أَكَّلَ   | verb <sup>7</sup> (causative of أَكَلَ or feed) | Form II       |                 |

Table 4.3: Several vocalised forms have the same normalised forms

For a given English word  $e$ , correspond Arabic words  $A$  and a set of applicable lemmas  $L$ :

Rank  $L$   $l_i$  based on a sequence of scores computed for each  $l_i$  as follows:

1. A boolean score, indicating whether a match exists with the gloss associated with  $l_i$  and  $e$
2. A boolean score, indicating whether an intersection exists between tokens in the gloss associated with  $l_i$  and the set of words that are semantically related to  $e$
3. A boolean score, indicating the possibility of the POS of  $l_i$  given the POS of  $e$ , always 1 unless:
  - a.  $l_i$  is a proper noun but  $e$  is not
  - b.  $l_i$  is a verb but main inflective forms are missing in  $A$
4. The probability of the  $l_i$  (unvocalised) given the  $e$ , based on statistics collected from the corpus
5. The number of surface forms in  $A$  that can result from  $l_i$  or the root and template form that applies to  $l_i$
6. The percentage ration of observed surface forms for  $l_i$  to the expected surface forms for  $l_i$
7. The probability of  $l_i$  (vocalised) in some external corpus.

Figure 4.4: Ranking lemma candidates

related. Thus, only aligned pairs that significantly co-occur are considered in the first phase, as it is important to avoid those lemmas that are not accurate and will impact the final results. A threshold can be set for such pairs, such as ignoring those with small counts. However, a threshold based on frequency is not a good idea since words are not uniformly distributed in a corpus but rather in a distribution that closely obeys Zipf’s law (Manning & Schutze, 1999). However, word frequencies are complicated and Zipf law is an approximation of the distribution. There are many other models that are claimed to fit the distribution, but none is exact (Piantadosi\_2014\_Zi). Thus, a frequency threshold appropriate for high frequency words will not be appropriate for other less frequent words. The translation probability – the probability of an Arabic word  $a$  given the English word  $e$  – can be another choice. Word translation probability is a conditional probability of  $x$  given  $y$ , which is calculated using maximum likelihood estimation as follows (Koehn, 2009).

$$t(a|e) = \frac{\text{count}(e, a)}{\text{count}(e)}$$

Such a conditional probability will not be affected by how frequent a translation occurs throughout the corpus, but only how frequent it appears as a translation for that source word. As a result, a frequent word that appears everywhere in the target language can have a probability above the threshold, even though it is not related to the source word and has been aligned by accident.

Instead, reliable pairs can be identified using positive pointwise mutual information (PMI) scores. PMI is widely used as an indication of association between words, e.g. (Bullinaria & Levy, 2007; Church & Hanks, 1989; Levy et al., 2015). PMI is often criticised as it favours rare events (Church & Gale, 1991). However, it is a good indicator of independence (Manning and Schutze (1999), p. 182), when close to zero. For a specific word pair  $x$  and  $y$  the pointwise mutual information is calculated as follows.

$$I(x; y) = \log_2 \frac{P(x, y)}{P(x)P(y)}$$

Pairs with few counts will not produce reliable PMI scores. Church and Hanks (1989) only considered pairs with a frequency greater than 5. The same threshold is used in this research.

#### 4.4.1.2 In Context Morphological Disambiguation

At this point the likely lemma has been determined. However, identifying the lemma is sometimes not sufficient due to ambiguous attachments or several inflections only distinguishable through diacritics, which results in multiple analyses for each lemma. For instance, “my book” and “two books” have the same surface form and the same lemma, but one has a possessive pronoun while the other has a dual suffix (in a certain

case). Another example is verb forms such as “I went”, “you went” and “she went” that all appear with the same surface form. English can provide evidence to resolve these cases. The first example can be resolved by knowing whether the corresponding English is plural or singular or whether it was preceded by a possessive. The second example can be resolved by knowing whether the verb is in first, second or third person by awareness of the subject pronoun type or its absence as a key to identifying the likely solution.

In this way, simple heuristics can be applied to address the problem, exploiting clues in the English side to resolve ambiguity on the Arabic side. However, some cases remain a challenge and can be impossible to resolve by relying completely on the English side. For instance, the second person pronoun – you – when not attached to a verb either as an independent word (not as a clitic) or as prepositional phrase “for you”, remains ambiguous with respect to the gender (singular feminine vs singular masculine) since they both have the same surface form, but no distinctive feature on the English side. The Arabic context may provide some clues. For instance, the gender of verbs in the second person or the gender of nouns or adjectives within a narrow window may give hints, but are not guaranteed to be reliable because the process does not know whether the other noun is syntactically related. In some cases the context may be neutral and both gender forms are plausible options.

#### 4.4.1.3 Adapting a Morphology Analyser

A morphology analyser was needed in order to find possible lemmas for an Arabic surface form. Buckwalter’s Arabic Morphological Analyzer (BAMA) is an Arabic rule-based tool with a lexicon. It produces all possible analyses that are allowed by its lexicon and rules. It was used for example, in manually annotating the Penn Arabic Treebank (Maamouri et al., 2004). It was also used in MADAMIRA (Pasha et al., 2014) to provide hypotheses which are then selected using a machine learning technique.

BAMA is composed of three main dictionaries. A prefix dictionary containing all possible sequences of elements that can form a prefix (including a null prefix), such as a conjunction followed by the definite article. Similarly, a suffix dictionary includes all allowable combinations of sequences including the null suffix, such as a feminine plural suffix, followed by a masculine singular possessive pronoun. The largest of the dictionaries is the stem dictionary which contains open class words: noun, proper noun, verb, adjective and adverb. BAMA also includes rules that govern the concatenation of prefixes, suffixes and stem classes to ensure a plausible linkages, which prevents concatenating a definite article with a verb, for example. Further, the analyser generates all possible hypotheses as suggested by the dictionaries and rules for each token independently, without awareness of the context, and provides no ranking. The analyser also assumes the input is provided without diacritics. The importance of

these components for the chosen methodology should not be underestimated, but they were insufficient for the text to symbol task, and additions were required to expand the stem dictionary.

In order to expand the BAMA stem dictionary, its structure needs to be clarified. The dictionary includes several stem forms for a given lemma, to cover the various forms a stem may transform into. These additional forms result from a change in the verb's aspect (perfective – imperfective – command form), a singular versus an irregular plural form, and orthographical adjustments needed for prefix or suffix concatenation. Each entry in the dictionary includes several fields: the stem, English glosses, POS, the lemma, and a special tag that indicates allowable attachments. The dictionary can be expanded by preserving the same format and thus avoids the limitations of the present open-licensed stem dictionary.

As part of the lemmatization process, the BAMA stem dictionary was extended using data collected from Arabic entries in the English Wiktionary as well as AWN. The target structure, as described earlier, dictates that other inflections need to be added as well. However, dictionaries often list the lemma form but not necessarily other forms. In Arabic, inflections often follow a templatic system which can be exploited to produce other forms, especially verbs. In some cases, the vocalised verb can be sufficient to identify the corresponding template, but it needs to be explicit in other cases. For instance, the exact verb template needs to be specified for triliteral verbs to correctly produce the imperfect form of the stem. Inferring inflections for other verb forms using templates are often straightforward (assimilation and vowels need to be handled). A few irregular command verb forms need to be explicitly mentioned in the dictionary. For nouns and adjectives, information about whether the word is inflected for gender and number needs to be added, in addition to providing irregular plural forms whenever applicable.

Fortunately, Wiktionary specifies the verb template as part of each dictionary entry. It also explicitly specifies other inflectional forms for nouns and adjectives such as plural and feminine forms. However, linguistic information is absent from AWN, apart from irregular plural forms. For the purpose of expanding the stem dictionary for the morphological analyser, a novel script was designed to generate the list of possible orthographic variations using a templatic system for verbs. The script also extracts inflectional forms such as plurals. Extracted forms are stemmed and tagged before being added to the dictionary. Further, orthographic adjustment needed is carried out before prefix or suffix attachment, and is also added to the dictionary.

Issues were encountered when lexical entries from multiple sources were combined. Some slight inconsistencies occurred arising from orthographic forms of the base form, which made a pair of lemmas non-identical, even though they refer to the same lexeme. This inconsistency can incorrectly increase the number of entries. For example,

Buckwalter uses “Alef Wasla” while Wiktionary uses “bare Alef”. Also, a few short vowels may have been omitted, especially those that can be easily determined, such as Sukun, which signals the lack of any short vowel sound, or Fathah that precedes a long vowel Alef. Another inconsistency occurs whenever a single letter has two diacritics, e.g. a Sheddah combined with another diacritic, such as Fatah. Their stored order is inconsistent (whether the Sheddah comes first or second), although this does not make any difference to how they are displayed, but it does make a difference when performing text comparison. Another discrepancy is the position of Tanween, a common spelling error, which should come before the final Alef (or Ta Marbuta) rather than the final character, but both cases were observed. For some word forms, especially those borrowed from other languages, there might be an alternative vocalisation (same form/same meaning, but slightly different diacritics), but there is no easy way to infer the likely vocalisation since most available corpora come without diacritics.

Further differences between the dictionaries were in the choice of lemma (the form that heads an entry in a dictionary). For instance, for common nouns ending with a Yeh, Wiktionary uses the indefinite form, where the final Yeh is dropped and the final diacritic is replaced with Tanween Kasratan. This is different from Buckwalter where the lemma appears in its full form (the final Yeh is kept). As a result, these inconsistencies need to be resolved to avoid redundancy, and were handled when extending the stem dictionary. A process of partial normalisation was carried out to reduce such redundancies. Alternative vocalisations were detected from their definition overlap, and one form was randomly selected as there was no statistical way of knowing which one was used more frequently in corpora having no diacritics.

A large number of proper nouns that exist in the corpus are unlikely to be part of any dictionary. A subset of these proper nouns consist of a transliteration of their English equivalents, and may be written in various ways due to lack of a standard transliteration approach and the different authors’ opinions as to how they should be written. Proper nouns do not inflect, but might be attached to prefixes. The difficulty is knowing whether the first letter (or letters) are part of the proper noun or are a prefix. However, a proper noun with no attachments can be identified if it appears a number of times, which will very likely include instances with no attachments. An algorithm was created to identify the base form for Arabic tokens that are aligned to English proper nouns. Thus, proper nouns can be handled without being included beforehand in the dictionary. A limitation of this approach is that proper nouns will not have any diacritics (no vocalized form), but they will be properly tokenised.

#### 4.4.2 Telegraphic Text and Fully-Formed Text Parallel Corpus

At this point, the corpus has been tagged with lemmas and full vocalisation is available from the morphological analysis. As discussed in section 4.1.1, the task of symbol to

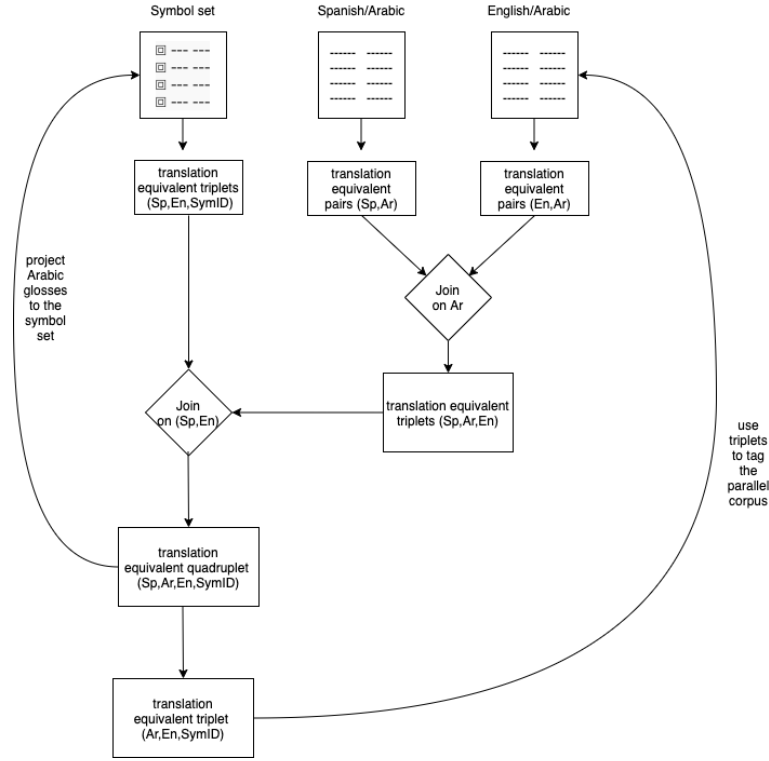


Figure 4.5: Disambiguation process for symbol tagging

text translations does not require actual symbol tags as part of the training data. Thus, the tagging so far is sufficient for creating training data for the task. The creation is straightforward by simulating an input symbol message, which was described in section 4.1.1. The input is a sequence of lemmas without function words. Function words were identified from the POS tags. Only lines shorter than 11 words are extracted. Finally, the target text was produced as fully vocalised sentences. At this stage, no swaps were made to simulate the users' behaviour, as suggested in section 4.1.1, which was left for future work.

#### 4.4.3 Symbol Tagging

The corpus has been tagged with lemmas, which is a necessary task before both text to symbol, as well as symbol to text. For the purpose of text to symbol translation, the MSA corpus needs to be also tagged with relevant symbols from the selected symbol set. As mentioned earlier (4.1.2), the main challenge when tagging a word in context with a symbol is the potential mismatch between the sense expressed by the context and the sense expressed by the symbol, even if they share the same word form.

A few symbols of ARASAAC were tagged with non-vocalised lemmas which were not a reliable translation. Therefore, those glosses were not used and the symbol set was considered without Arabic glosses. Finding relevant symbols by depending solely on the English corresponding word would potentially result in unrelated symbols due to

gloss ambiguity. This ambiguity can be minimised by taking into account another language. Figure 4.5 illustrates the process. An additional language (Spanish) was introduced to disambiguate the English text. Two parallel corpora were used, the main English/Arabic and an additional Spanish/Arabic corpus (Lison & Tiedemann, 2016). This resulted in two sets of translation equivalent pairs, which are extracted after word alignments have been determined.

Each set was filtered, keeping only pairs with an acceptable association score. The two sets were then joined, based on having the same Arabic token, resulting in a set of triplets of English, Arabic and Spanish translation equivalent words. This set needed filtering to avoid the introduction of rare Arabic translations. Frequency is a key to eliminating noise and the threshold was decided empirically, by considering the translation equivalent pairs only where 100 or more alignments occurred, to avoid rare translations as well as alignment errors. The threshold of 100 may seem high but the corpus is large and the mean frequency per English token is 475.3. The frequency of common words, such as ‘book’, is over 33 000, while a less frequent yet common token, ‘pencil’, has a frequency of 1 308. Thus 100 seems a reasonable choice. The pair also need to be significantly associated to avoid the introduction of function words. Thus pairs with a PMI score less than 5 were ignored. This allows several Arabic translations to be added, as well as covering a lemma (or several lemmas) that is likely to have the same meaning due to intersection and the high threshold. However, for infrequent pairs, the single most frequent translation was chosen, to limit potential errors due to its small count.

Another set of triplets is extracted from the symbol set. Each item in this set is made up of an English gloss, a Spanish gloss, and a symbol unique id. These two sets of triplets (corpus set and the symbol set) are further joined, based on English and Spanish tokens, resulting in a set of quadruplets. It is important to pay attention to the form of glosses and perform necessary pre-processing before searching for a match. For instance, verbs are often preceded with ‘to’. Thus, removing the preceding ‘a’, ‘the’, and ‘to’ was done to increase recall. Also, some compounds are separated by a hyphen or space such as ‘make-up’, so dropping the separator as well as keeping the original gloss increases the recall.

The overall process of obtaining translation equivalent tokens is summarised in Figure 4.6. Using the resulting set (T5 in Figure 4.6) links Arabic and Symbols using both English and Spanish as pivot languages. It is a knowledge base for symbol tagging as well as for generating Arabic glosses for each symbol.

Although a link has been established between Arabic and symbols, Arabic tokens in the corpus remain ambiguous. As a result the tagging process does not tag the Arabic token in isolation but also examines its English correspondence to minimise ambiguity. The pair of Arabic and English tokens are used to find a match in the quadruplets set.



1. From En Ar aligned corpus extract all En word type  $x$  and Ar word type  $y$  pairs such that  $\text{PMI}(x,y) > 0$  and  $\text{count}(x,y) > \theta \rightarrow T1$
1. From Es Ar aligned corpus extract all Es word type  $x$  and Ar word type  $y$  pairs such that  $\text{PMI}(x,y) > 0$  and  $\text{count}(x,y) > \theta \rightarrow T2$
2. From the symbol set extract all co-occurring En gloss, Sp gloss and symbol unique ID  $\rightarrow T3$
3. Pairs in  $T1$  and  $T2$  are joined to form a new triplet if the Arabic token in  $T1$  is equivalent to the Arabic token in  $T2 \rightarrow T4$
4.  $T3$  and  $T4$  are joined to form a new quadruplet if the English and Spanish tokens in  $T3$  are equivalent to English and Spanish tokens in  $T4 \rightarrow T5$

Figure 4.6: Extracting translation equivalent set

Matching can be based on their lemmas if matching based on their surface form failed. Further improvements were achieved, when a pair of Arabic and English words had multiple instances with different Spanish glosses, by ranking matching symbols based on the translation probability of Arabic word given the Spanish gloss. After tagging is completed, the Arabic side can be used independently of any other language. The process was further improved by taking into account the part of speech of the associated gloss and the word in context (from the corpus).

A pair of glosses and their POS class may be associated with more than one symbol, arising from either multiple representation of the same meaning or two different meanings with glosses that are cross-linguistically ambiguous, e.g. *operación* in Spanish and *operation* in English. The former case is handled by tagging a word with all possible matching symbols. The latter is complex and requires awareness of the specific meaning for the word in context, and the associated gloss, and considers a match only if both share the same meaning. However, it is hard to determine computationally which is the case. Symbols with cross-linguistically ambiguous English and Spanish pairs need special attention.

Ambiguity of concern is not systematic polysemy, such as *school* the building vs. *school* as educational institution, but rather unrelated senses. Symbols with cross-linguistically ambiguous English and Spanish pairs are often domain specific, such as ‘operation’ in mathematics vs. the medical domain, and ‘king’ the chess piece. Therefore tagging domain specific symbols with their domain might minimise potential errors by only tagging words with domain specific symbol if there is an overlap between context and domain keywords. However, this assumes that once a pair of English and Spanish glosses is covered in a symbol set, all possible meanings will be present in the symbol set, which is not the case. For instance, the sense of operation

and its Spanish equivalent *operación* used in the military domain is not covered by the symbol set. These domain specific meanings are often tagged with their domain in conventional dictionaries. However, this was not carried out as part of this research but left for future work. Furthermore, one should not assume that all meanings are covered in the symbol set.

The assumption so far has been that glosses are single words, which may not be always the case. A gloss can be a multi-word expression (MLE) and they may occur in one language, but not the other (such as the English gloss, e.g. ‘toes’ and the Spanish equivalent ‘dedos del pie’). This matter needs to be addressed before the tagging process and during the extraction of translation equivalents. When joining the corpus-based triplets with the symbol-based triplets (Step 4 in 4.6), it is necessary to allow partial matches between glosses and words (at least one common word that is not a function word). While later tagging text, the retrieval of candidates is based on words, but MWE glosses can be among the retrieved set of candidates and the corresponding symbol is used only if the full gloss exists in the text (English side) by examining adjacent words. However, the coverage of such a symbol with respect to the Arabic side is still hard to determine. This issue can be addressed by including all Arabic tokens that are mutually aligned to any English token that is part of that phrase (including unaligned internal tokens), as long as there is no intervening token that is aligned to an English token beyond the phrase boundary. Some multi-word glosses may not always be contiguous as other words may intervene, e.g. verb-particle constructions (Sag et al., 2002), which pose many challenges such as how to align symbols when the coverage of one symbol is interrupted by another symbol. Ensuring the sequence of words is related and refers to one concept remains an open issue.

Communication symbols are often associated with glosses in several languages, as is the case with ARASAAC and the availability of a parallel corpora for extracting translation equivalent pairs. This process can be used with any language given the availability of the same resources. However, a remaining potential source of error is when a set of senses share the same word form in both Spanish and English, but the Arabic token applies to only a subset of those shared senses. The manual tagging of domain specific symbols was suggested, but visual content may give some indication. Therefore, the potential of the visual content in disambiguation was explored, as discussed below.

## 4.5 Visual Content of Symbols

The tagging approach so far did not make use of the visual content of the symbols used. The final step of this study was to explore the potential of the visual aspect of symbols. The goal was to gain insight into their contribution to the tagging method, as

opposed to using only textual glosses. The motivation behind such an exploration is to overcome the scarce descriptive data provided with the symbols, and the need for additional metadata alongside the available associated glosses. This may avoid manual tagging since the symbol set is large.

Awareness of the graphical content of symbols can be useful to provide additional clues about the likely meaning the pictographs are designed to convey. There is a need for a computational method to make sense of visual content. Computer vision literature has proposed many algorithms useful for this task. The methods selected for this research are designed to measure similarity between symbols. It has been assumed that symbols with similar content will represent semantically-related concepts. Moreover, since these symbols have been made for communication purposes, some symbols might have been intentionally made to look similar in design, in order to preserve coherence and clarity. The schematic grouping of symbols is a symbol set characteristic pointed out by Pampoulou and Detheridge (2007), which they claim increases symbol comprehensibility. The authors indicated that such a characteristic is seen in Widgit Symbols. Although no published evidence confirms the situation with ARASAAC, initial exploration (in Chapter 3) suggests that this characteristic is met by ARASAAC as well.

The coloured pictographic symbols provided by ARASAAC uses a range of colours using a 500 by 500 pixel image distributed in a Portable Network Graphic (PNG) format. The collected symbols have distinct outlines, often on a white or transparent background rather than a sophisticated full scene. These drawings are produced by the same organisation, thus maintaining a consistent depiction style. It is important to be aware that these symbols are unlike other general images. A lot of the sophistication that is usually present in typical images, such as variation in lightning or noise, does not exist in these symbols. However, distinctive features in some symbols are insignificant, which may make a large number of symbols similar and hard to discriminate, e.g. a large proportion contain a stick figure. However, the goal at this point is not to understand the content of the pictographic images computationally (i.e. object recognition or image captioning), but rather to find similar ones. This captured similarity can be useful as a bridge to access semantically-related glosses, which can provide context and can be used to acquire additional contextual data that can be useful for disambiguation.

A first glance over the ARASAAC symbol set reveals that those pictographs illustrating objects appear to show a consistent set of colours and few coloured backgrounds. Given the minimalism of the content of these symbols (no texture – white background), colour seems a good feature to consider when searching for similar symbols. As a result, similarity based on the approach of Rubner et al. (2000) was examined. However, using colour alone was found to be insufficient. Images may share the same set of colours by chance. For instance, a stickman has the same colour

distribution as a character. Thus, this approach was abandoned and instead an approach adopted that was based on the structure rather than the colours.

Many methods can be used to create a numeric representation of an image or local areas, such that ‘distances’ between two different images correlates with perceptual similarity between the sources. SIFT creates a descriptor of a local area of interest which is robust against scale variation (Lowe, 2004). The SIFT implementation provided with OpenCV was used for pairwise comparison between symbols. A number of SIFT descriptors is generated for each image. The descriptors generated from two symbols are then matched. Such an operation often contains a number of false matches, and so RANSAC was used to filter out these false matches following common practice. A pair of images is then considered similar if the number of images exceeds an empirically chosen threshold <sup>8</sup>. As a result, a symbol can have from zero to many similar symbols. However, a few symbols did not have an area of interest from SIFT’s perspective, and consequently there was no possibility of identifying similar symbols based on this approach.

HOG (Dalal & Triggs, 2005) generates a descriptor of a full patch in an image, rather than finding areas of interest. It is not sensitive to scale or position. Thus, it was chosen to provide an overall compact representation of the symbol. It is good at detecting symbols that are similar, but with small modifications such as gender variation, stick versions or morphological variation. HOG was used to generate a single descriptor per symbol, and cosine similarity was calculated between descriptors with an empirically chosen threshold <sup>9</sup>.

## 4.6 Evaluation

A corpus was annotated with lemmas and other morphological data, in addition to the generation of symbols. A method is needed that evaluates the accuracy of the generated tags. Unfortunately, no human reference tagged text is available to compare against for calculating precision and recall. The lack of a manually tagged corpus can be addressed by back translation. Koehn (2005) investigated back translation as a method of evaluation: translating from a source language to target language and back again to the source. He argued against this approach and showed that it gives a false high BLEU score. However, he carried out alternative ways to evaluate the outcome.

Morphological tags were compared against automatically generated tags. The evaluation was made by calculating agreement scores between corpus tags and tags generated by MADAMIRA on a random sample (lemmas and full analysis). Although this score does not measure the improvement (or decline) in performance over existing

<sup>8</sup>The lower limit for the similarity of two symbols was 25 SIFT key points

<sup>9</sup>The lower limit for the similarity of two symbols was 0.8

tools, its performance was close to tools that were privileged to use a manually tagged corpus with training data. Furthermore, the number of out of vocabulary tokens have been counted and compared. Differences have been analysed and are reported in the results chapter.

Symbol tags were evaluated by comparing contextual words of a symbol against contextual words of the associated English gloss, which can be ambiguous. Such a list will be exceedingly long and thus only words with a high associated score were kept, which was determined by the PMI score. A difference between the associated list of words with the symbol, against words associated with the English gloss, is evidence that not all senses covered by the word form were tagged with the symbol. Further examination of words in each list revealed the additional covered sense. Several examples with ambiguous glosses were inspected. An example is shown in Chapter 5.

The task of generating text given a sequence of lemmas or telegraphic input was also examined. The subsequent corpus of telegraphic and fully-formed text generated from the tagged corpus was used to train both a text-to-symbol and a symbol-to-text translator, by splitting the data into training and testing. The translator used was an LSTM implementation using Pytorch, the output being tested using BLEU score. However, an actual symbol to text system is only as good as its symbol set coverage. Symbols need to cover all lexical units in a sentence, otherwise the input will be incomplete and will falsely decrease the accuracy.

The ability to identify similar symbols was also evaluated. Quantitative observations were reported and a sample is shown in Chapter 5 with detail in Appendix C and Appendix D. An example that demonstrates the usefulness of accessing similar glosses is given in the next chapter.

## 4.7 Summary

The task of translating between text and symbols was approached using data based methods that required an annotated corpus. Having reviewed the literature, a relevant domain corpus was selected. The appropriateness of the corpus was confirmed by comparing it against other corpora using an AAC list of messages. Once the corpus was selected, the next goal was to transform this raw corpus into training data useful for translating between symbols and MSA. One objective was to avoid manual tagging and the expenses associated with it. However, tagging requires some knowledge since words in isolation are ambiguous, whether words in associated glosses or words in a sentence. As a result, the tagging approach made use of currently available translations as a source of knowledge.

Lemmatization was the first step carried out. Due to the intense morphology of Arabic, lemmatising the corpus was a key for both the text-to-symbols and symbols-to-text tasks. The lemmatisation process used evidence from the same corpus, which overcame the issue of domain mismatch. The determination was approached out of context, based on translating the equivalent word, and then tagging in context, along with figuring out the vocalisations. The resulting corpus was sufficient for the symbol to text task and a subsequent parallel corpus of telegraphic fully-formed sentences for training.

The text to symbol task required a corpus that was also tagged with symbols. Tagging with symbols needed further semantic disambiguation. The approach used an additional language, Spanish, to bridge the pairing of Arabic and English with a relevant symbol. The introduction of a third language was motivated by the need for at least two languages for disambiguation, and the lack of manual Arabic translations of glosses associated with symbols. Eventually, the corpus was tagged with ARASAAC symbols in addition to lemmas and vocalisation.

There are some limitations to having few glosses as the only textual data available with symbols. The visual content of symbols may provide some clues. To help with this, the visual content of the symbols was examined using popular computer vision algorithms to determine similar symbols. Two approaches were used and results were reported. Once similar symbols have been identified, their associated glosses provide context to these isolated glosses.

## Chapter 5

# Results

The aim of this chapter is to present the results of experiments conducted, and to analyse the data. It first presents statistics about the two parallel corpora which spanned the three languages that were used for disambiguation. Next, it shows the outcome of the experiment that was carried out to pre-process the MSA side of the corpora, which involved lemmatization and part of speech tagging as a result of a morphological analysis disambiguation. This experiment is analysed for its coverage, and statistics collected from the resulting annotated corpus are shown in some examples. Then it addresses the experiment that created a parallel monolingual corpus of telegraphic and fully-formed text as training data for the symbol to text task, showing statistics and a sample of the data. Next, the chapter shows the experiment examining the potential of a translator, which was trained using the created data, by evaluating the resulting text using the BLEU score. The next experiment was on symbol tagging, and its outcome is reported. This was evaluated qualitatively by examining a single example and discussing its overall statistics. Finally, focusing on symbols, the chapter presents results of two experiments. The first aimed at finding similar symbols using two different methods; the second aimed at identifying symbols with certain markers.

### 5.1 Corpora and Extracted Data

This section reports on the quantitative aspects of the multi-lingual corpora (see 4.3). Two parallel corpora were used, the main one for tagging and disambiguation, and a secondary one that was only used for disambiguation. The main corpus was a parallel corpus between Arabic and English. The number of tokens for Arabic was 186 million, and the average sentence length was 6.3. The English had 255 million tokens, and an average sentence length of 8.6. For tokens that appeared at least 5 times, the number of word types was 144 000 for English and 382 000 for Arabic (Table 5.1).

|                     | English     | Arabic      |
|---------------------|-------------|-------------|
| tokens              | 255 542 565 | 186 783 973 |
| types               | 537 639     | 1 288 924   |
| types $\geq 5$      | 144 469     | 382 488     |
| Average line length | 8.7         | 6.3         |
| Number of lines     | 29 442 754  |             |

Table 5.1: Arabic English Corpus Statistics

|                     | Spanish     | Arabic      |
|---------------------|-------------|-------------|
| tokens              | 207 142 283 | 164 587 851 |
| types               | 586 315     | 1 506 863   |
| types $\geq 5$      | 191 879     | 363 162     |
| Average line length | 7.9         | 6.3         |
| Number of lines     | 26 138 723  |             |

Table 5.2: Arabic Spanish Corpus Statistics

| translation | translation score | PMI score |
|-------------|-------------------|-----------|
| الكتاب      | 10052             | 12.02     |
| كتاب        | 7668              | 11.99     |
| كتابك       | 1751              | 12.30     |
| كتابي       | 1467              | 12.16     |
| كتابا       | 1428              | 12.29     |

Table 5.3: Top Arabic translations of the English word book

Word alignments were extracted from this parallel corpus and word translation models were extracted. A sample of top translations of the word ‘book’ is shown in Table 5.3 and top translations of the Arabic word “كتاب” is shown in Table 5.4. The two shown samples highlights importance of PMI score to avoid function words or taking the intersection of the two alignments. Additionally, the number of extracted pairs from the directional word alignments, excluding those with a count not less than 50 and a PMI score of at least 5, is shown in Table 5.7.

The secondary corpus was a parallel corpus between Arabic and Spanish. This resource was used in the disambiguation process needed for symbol tagging. The number of tokens for Arabic was 164 million, with an average sentence length of 6.3. The number of tokens for Spanish was 207 million, with an average sentence length of 7.9. For tokens that appeared at least 5 times, the number of word types was 192 000 for Spanish and 363 000 for Arabic. Word translation samples are shown in Table 5.6 and Table 5.5 and statistics regarding translation equivalent pairs are shown in Table 5.2.



| translation | translation score | PMI score |
|-------------|-------------------|-----------|
| book        | 7639              | 11.84     |
| a           | 3969              | 4.00      |
| the         | 1517              | 2.01      |
| 's          | 663               | 1.34      |
| of          | 338               | 1.21      |

Table 5.4: Top English translations of the Arabic word كتاب

| translation | translation score | PMI score |
|-------------|-------------------|-----------|
| libro       | 6627              | 11.67     |
| un          | 3844              | 4.51      |
| el          | 1837              | 3.07      |
| de          | 1555              | 1.82      |
| libros      | 295               | 8.48      |

Table 5.5: Top Spanish translations of the Arabic word كتاب

| translation | translation score | PMI score |
|-------------|-------------------|-----------|
| الكتاب      | 8243              | 12.06     |
| كتاب        | 6637              | 12.11     |
| كتابك       | 1556              | 12.43     |
| كتبا        | 1273              | 12.43     |
| كتابي       | 1264              | 12.26     |

Table 5.6: Top Arabic translations of the Spanish word libro

| alignment  | number |
|------------|--------|
| Eng to Arb | 124816 |
| Arb to Eng | 135360 |
| Spa to Arb | 121213 |
| Arb to Spa | 127270 |

Table 5.7: The number of extracted Pairs with frequency  $\geq 50$  and PMI score  $\geq 5$  from each alignment

## 5.2 Text Pre-Processing

After the corpus had been tokenised, normalised, and word alignments determined, a linguistic annotation experiment was carried out. Annotations were performed at the word level and involved vocalised lemma, vocalised surface form, and Buckwalter part of speech. An additional form to produce the corresponding segmented text, where attached pronouns were included only when a corresponding English pronoun was present (to avoid repetition). Each segment was augmented with its POS tags, and each token that belonged to an open class words was lemmatized. A sample of the annotation is shown in Table 5.9.

The annotations were inferred from the extracted word alignments and other surface forms aligned to the same word, as explained in section 4.4.1. For an analysis to be

| Part of speech | Count  |
|----------------|--------|
| NOUN           | 38 605 |
| ADJ            | 6 511  |
| NOUN PROP      | 5 358  |
| VERB           | 9 644  |
| ADV            | 373    |

Table 5.8: Parts of Speech in the Arabic dictionary

```

<sentence id="0">
<form type='raw sentence'>هل فكرت في المستقبل</form>
<form type='vocalised sentence'>هَلْ فَكَّرْتَ فِي الْمُسْتَقْبَلِ</form>
<form type='POS'>INTERROG_PART VERB_PERFECT+PVSUFF_SUBJ:2MS
PREP DET+NOUN</form>
<form type='glosses'>هَلْ أَنْتَ فَكَّرَ مُسْتَقْبَلِ</form>
<form type='glosses POS'>INTERROG_PART VERB_PERFECT PRON_2MS
NOUN</form>
</sentence>

```

Table 5.9: The linguistic annotation for an Arabic sentence that was a translation of “do you ever think about the future ?”

|                    |                       |
|--------------------|-----------------------|
| Lines              | 10 446 868            |
| Telegraphic Arabic | 36 674 unique tokens  |
| Full Arabic        | 353 248 unique tokens |
| Full English       | 143 654 unique tokens |

Table 5.10: Symbol to text parallel statistics

generated, there had to be at least one applicable entry in the analyser’s associated dictionary, except for proper nouns which were handled based on corpus evidence and not bounded by the dictionary. To expand the lexicon coverage, lexical entries were gathered from multiple sources: Wiktionary, AWN, and data that was part of BAMA (Buckwalter, 2002). The collection of extracted lemmas were conflated to minimise redundancy<sup>1</sup> resulting in around 54 000 forms. Table 5.8 shows the number of lemmas per part of speech class.

### 5.3 Symbol to Text Experiments

Training data was needed to train a symbol to text system. As discussed in Chapter 4, the training data was purely textual, ignoring the pictorial element. Two experiments were conducted, one to transform the tagged corpus to a parallel corpus of pairs of telegraphic sentences and their corresponding fully-formed sentences, and a second that used the resulting training data to train a translator and test its potential.

<sup>1</sup>Through handling missing diacritics and dictionary spelling variations

| Fully-Formed Sentence              | Telegraphic Sentence             |
|------------------------------------|----------------------------------|
| هَلْ فَكَّرْتَ فِي الْمُسْتَقْبَلِ | هَلْ أَنْتَ فَكَّرَ مُسْتَقْبَلِ |

Table 5.11: Fully-formed text and its corresponding telegraphic form

| target       | BLEU  |
|--------------|-------|
| Full Arabic  | 29.45 |
| Full English | 28.18 |

Table 5.12: BLEU score for unseen test segment

The fully-formed text was available as the output from the pre-processing experiments (section 5.2). On the other hand, telegraphic text was created for each sentence in the annotated corpus (using the gloss form), as described in section 4.4.2. The resulting corpus was bounded by the annotated corpora. The resulting parallel telegraphic and fully-formed sentences was over 10 million lines (Table 5.10). A pair of sentences is shown in Table 5.11.

The resulting corpora were used in a machine translation experiment to gain insight into the difficulty of the task. The source text was telegraphic Arabic while the target was fully vocalised Arabic. For comparison, an additional parallel corpus was created with the same source text (telegraphic Arabic), but with full English text as the target. The resulting two parallel corpora were used to train two translation systems using the OpenNMT Pytorch implementation (Klein et al., 2017) default model, which is an encoder-decoder two-layer LSTM model with 500 hidden units. The two systems were evaluated against an unseen segment. The evaluation was measured using the BLEU score<sup>2</sup>. When the two results were compared, it could be seen that translating into full Arabic is a challenge, even though both sides originated from the same Arabic text. Such a challenge was evident in the BLEU score obtained, which was only slightly higher than the English score<sup>3</sup>.

## 5.4 Symbols to Text Experiments

An experiment was carried out to generate training data needed for the symbol to text task. The specific symbol set chosen was ARASAAC, which had 12 662 symbols available at the time of this study. The symbols have already been discussed in Chapter 3, so it is just important to note that this open-licensed symbol set cannot be compared to any mainstream languages in size, or commercial AAC symbol sets. These would include Picture Communication Symbols (PCS) with over 45 000 symbols, and Widgit with 17 000 symbols, covering over 45 000 words in English. However, ARASAAC has shown to be a good base from which to work, with glosses that have

<sup>2</sup>Using multi-blue.perl, part of the Moses decoder

<sup>3</sup>Training time was limited to 12 hours; a longer training time is expected to improve the result

been translated from Spanish into several languages, including English. It was also found that at least 70% of the glosses were nouns, which at the time of writing is similar to that part of speech found in the collected Arabic dictionary (73%).

The experiment tagged text with symbols automatically. The approach followed made use of translations of the glosses, as well as translations of the corpus, to minimise ambiguity errors that may have resulted from a dependence on one language (section 4.4). The resulting tagged corpus was evaluated. A gold-standard data, such as a manually-tagged corpus, was not available to compare it with. The issue was investigated in section 4.6 where it was suggested that the performance could be measured by examining words associated with a symbol against words associated with the gloss with no disambiguation.

Two symbols with the ambiguous gloss ‘nail’ were selected as a case for demonstration. The two symbols have different meanings as suggested by their depiction (Table 5.13 and Table 5.14). The list of highly associative co-occurring words was extracted from the corpus (determined using a positive PMI score as a threshold (Church & Hanks, 1989)). The list associated with each of the symbols and words are shown. Examining the list, it can be seen that a word like ‘hammer’ is associated with the relevant symbol (Table 5.13), but not the other, while ‘clippers’ and ‘finger’ are only associated with the ‘fingernail’ symbol (Table 5.14). Words associated with the noun ‘nail’ regardless of the symbol are shown in Table 5.15, which includes words related to both senses. This example showed how, by using translations, it was possible to distinguish between two different meanings that were indistinguishable relying on the English gloss alone.

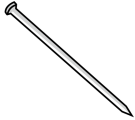
|   |       |
|---|-------|
|  |       |
| nail 2  |       |
| nail  | 13.09 |
| coffin  | 11.03 |
| hammered  | 9.17  |
| hammer  | 9.16  |
| polish  | 8.67  |
| rusty   | 7.19  |
| driven  | 7.18  |
| boot  | 7.05  |
| stepped   | 6.95  |
| tire  | 6.94  |
| final   | 6.88  |
| sticking  | 6.77  |
| gun   | 6.60  |

Table 5.13

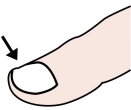
|   |       |
|---|-------|
|  |       |
| nail 2  |       |
| nail  | 13.09 |
| remover   | 12.47 |
| polish  | 12.44 |
| clippers  | 11.78 |
| varnish   | 11.66 |
| clippings   | 11.55 |
| salon   | 11.24 |
| lipstick  | 8.43  |
| tooth   | 8.41  |
| toe   | 8.32  |
| beds  | 8.31  |
| scissors  | 7.72  |
| file  | 7.10  |
| marks   | 7.03  |
| matches   | 6.56  |
| color   | 5.93  |
| finger  | 5.93  |
| broke   | 5.66  |
| skin  | 5.03  |
| dna   | 5.02  |

Table 5.14

|             |       |
|-------------|-------|
| nail (noun) |       |
| polishes    | 12.09 |
| press-on    | 11.95 |
| remover     | 11.47 |
| polish      | 11.24 |
| clippers    | 11.17 |
| varnish     | 10.78 |
| salon       | 10.67 |
| scrapings   | 10.63 |
| biter       | 10.20 |
| clippings   | 10.17 |
| salons      | 10.14 |
| coffin      | 9.82  |
| clipper     | 9.42  |
| tooth       | 9.04  |
| chipped     | 8.64  |
| hammered    | 8.30  |
| hammer      | 8.13  |
| cursing     | 7.88  |
| toe         | 7.38  |
| beds        | 7.33  |
| scissors    | 6.95  |
| carpenter   | 6.87  |
| lipstick    | 6.81  |
| rusty       | 6.69  |
| marks       | 6.15  |
| file        | 6.14  |
| fought      | 5.95  |
| boot        | 5.81  |
| final       | 5.77  |
| sticking    | 5.73  |
| driven      | 5.69  |
| stepped     | 5.64  |
| pound       | 5.63  |
| nails       | 5.60  |
| gun         | 5.49  |
| nail        | 5.37  |
| sticks      | 5.24  |
| tire        | 5.22  |
| finger      | 5.22  |
| broke       | 5.22  |
| matches     | 5.20  |
| chip        | 5.07  |

Table 5.15

## 5.5 The Visual Content

Pictographic communication symbol sets appear to have stable visual designs that can be exploited for feature detection. To evaluate the potential of the visual content in disambiguating the associated gloss, several experiments were carried out. Two experiments were undertaken to identify symbols with similar content using the SIFT and HOG approaches, as described in section 4.5. The two methods were used to quantify the pairwise similarity between symbols. A threshold was chosen for each method, based on empirical evidence, to classify a pair of symbols as similar or not. The results varied between the methods and among the symbols. Some had several similar symbols in the subset while others, based on the similarity score, seemed to be isolated.

The two approaches were compared by calculating the number of symbols that had at least one similar symbol Table 5.20 of results suggests that SIFT outperformed HOG in its ability to identify similar symbols. The example shown in Table 5.18 reveals the differences between the two approaches, by contrasting the performance of each against the same input. In this example, SIFT was able to identify few related symbols. Part of the similar symbols includes an object that is part of the main symbol, while the other part appears to have the same layout as the main symbol. In contrast, HOG identified symbols with the same layout, but failed to find symbols with a common object.

An additional example is shown in Figure 5.17 which had an input symbol that lacks any sophistication. In this case, SIFT failed to identify similar symbols, while HOG was able to find a few that had a very similar layout. In contrast, the example shown in Figure 5.16 is sophisticated and SIFT was able to identify related symbols without being affected by the position or size of a segment from the main image. Further examples for each approach can be found in (Appendix C) and (Appendix D). There are overlaps between the two methods but SIFT made use of overlapping areas to identify similar symbols, while HOG relied on the layout. The output of both can be combined.

The motivation behind identifying similar symbols is to gain awareness of the context of each symbol. The context can be represented by a collection of key words. An experiment was conducted to extract contextual data. An example is shown in Table 5.19. The table shows a symbol and its associated gloss ‘writer’ (a) on the left side of the table, and visually similar symbols as suggested by the algorithms along with their associated glosses (b), excluding ‘writer’. The two lists of glosses (a and b) were used to obtain further contextual data for the ‘writer’ symbol. The last row in the table shows the set of tokens that are associated with (a), as well as one or more of the glosses in (b). The resulting list of words is indeed semantically relevant to the ‘writer’ symbol. This example demonstrated the potential of using visual content to add more relevant contextual data.



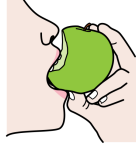
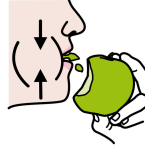

|                 |   |   |  |  |
|-----------------|---|---|--|--|
| Main symbol     | <br>raw        |   |  |  |
| Similar symbols | <br>vegetables | <br>bite | <br>chew | <br>vinegar |

Table 5.16: SIFT example




|                 |  |  |
|-----------------|--|--|
| Main symbol     | <br>with      |  |
| Similar symbols | <br>against | <br>to/for |

Table 5.17: HOG example

Some symbols had markers on them to indicate their semantic or linguistic category, section 4.2. An experiment was carried out to identify symbols having a certain marker. Three markers had been identified, as shown in Table 5.21. These markers at first sight seemed consistent in shape, colour and position across the symbol set. Identifying the existence of each marker was achieved using template matching (section 4.2), along with EMD (section 2.5.4.1) to compare colour. However, variations were noticed when the set was examined computationally, and thus slight variations in size, position and colour were allowed. The number of symbols identified for each marker are shown in Table 5.21. Empirical analysis suggests that the approach was successful in finding related symbols. Symbols with the plural marker were the largest of the three. Being aware of symbols with the plural mark can be useful in avoiding unnecessary repetition in the symbol set. The cross marker appeared less than expected, since many symbols related to medicine did not have this marker. Thus, the absence of the marker was not a reliable clue, since it did not rule out a medicine-related sense. Finding symbols with certain markers can be useful in








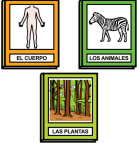



|                      |   |  |  |   |   |  |
|----------------------|---|--|--|---|---|--|
| Main symbol          | <br>tale*  |  |  |   |   |  |
| SIFT nearest symbols | <br>tale   | <br>adapted library | <br>Science books        | <br>Social books | <br>Science   |  |
| HOG nearest symbols  | <br>books* | <br>Science books   | <br>Poetry and Art books | <br>Social books | <br>Sciences' |  |

Table 5.18: Example showing HOG vs. SIFT



| Main  | Visually similar symbols   |
|---|--|
|  |  |
| (a) writer  | (b) receptionist - physiotherapist - drawer - journalist - physiotherapy             |
|   | (c) writes - freelance - journalist - novelist - editor                              |

Table 5.19: Visually similar symbols

Example shows the potential of using visual content in acquiring additional data (c)

| Method | Coverage |
|--------|----------|
| SIFT   | 60.98%   |
| HOG    | 46.16%   |

Table 5.20: Percentage of symbols with at least one similar symbol






|   |   |   |
|---|---|---|
|  |  |  |
| 1762  | 110   | 322   |

Table 5.21: ARASAAC markers and number identified

disambiguation tasks if the addition of the marker was consistent across the symbol set.

## 5.6 Summary

This chapter presented results and statistics related to the resources available, and to the experiments undertaken with those resources. The statistics showed the size of the corpus was good, which has a huge impact on the outcome since, with a large amount of data, it is more likely to provide accurate translations than erroneous ones. Results also showed a considerable amount of data had been used in the morphological analyser. It is critical to have an analyser with good coverage, especially for frequent core words, since a missing lexical entry will decrease the symbol coverage as it will prevent the translation of an associated gloss, and subsequently prevents it from being used in the tagging process. The lack of a lexical entry will limit the amount of training data generated for the symbol to text task, since the failure to identify a lemma for any token in a sentence will make it useless for the task.

The corpus was used to generate training data for both symbol to text, as well as text to symbol, tasks. However, the corpus needed pre-processing. Final annotated corpus lines longer than 10 words, as well as lines with words not covered by the collected lexicon, were abandoned. Training data for the symbol to text task was created and the corpus annotated with symbols. Statistics on the resulting corpus were reported. This resulted in a 10 million line corpus sufficient for training a machine translation system. However, such a portion is small compared to the original corpus, which is over 29 million lines, and further analysis is needed. The resulting training data was used to train a symbol to text translation system, and compared with English as another target language. Results were surprisingly close, which highlights the difficulty of translating telegraphic text into a fully-formed sentences, in other words, translating into English from Arabic telegraphic text was close to translating Arabic telegraphic text to full Arabic, based on the resulting BLEU scores.

Symbols were the focus of attention when creating training data for a text to symbol tagger or translator. However, the symbols initially lacked Arabic translations for their associated glosses. The translation was obtained using translation equivalents extracted from two pairs of languages. The significance of the disambiguation

approach was illustrated through the list of words associated with two symbols having the same gloss but different meanings.

To further improve the disambiguation process, the visual content of symbols was tackled. SIFT and HOG were used to identify similar symbols. SIFT showed more success, as evidenced by the percentage of symbols that had at least one similar symbol that did not share the same gloss. But HOG was able to identify some symbols where SIFT failed, due to a lack of sophistication. Being aware of similar symbols can be useful in automatically obtaining related words, which can contribute to the disambiguation process as well as increase recall when used in a retrieval system. Associated markers, such as the cross mark can be used for disambiguation. However, the number of symbols with a semantic marker is insignificant.

Results in this chapter show the potential of the approach followed in automatically creating training data for translating to and from symbols. Outcomes indicate significant morphological analyser coverage and lemmas subsequently identified. The evidence supports the use of another language for disambiguation, while analysing visual content demonstrated the potential for reducing ambiguity and consequently increasing the relevancy of tagged symbols. The next chapter discuss observations and limitations of this approach.

## Chapter 6

# Discussion

This research focused on creating a symbol annotated corpus needed for translating between symbols and text as part of technologies targeting the AAC community. The goal has been to create a corpus automatically with no human intervention. The tagging was carried over corpora that matches daily face-to-face conversation. The research treated MSA as representative of a natural language alongside a symbol set. MSA is morphologically rich and to some degree agglutinative, and raw text cannot be directly tagged with symbols but needs pre-processing. Lemmatisation is key to the annotation process and thus significant attention has been paid to this task. Of course, from a computational point of view, lemmatisation alone is not enough to determine the symbol, due to sense ambiguity. Two disambiguation tasks were created as part of the tagging process, one to determine the lemma, and the other to determine the relevant symbol. The former made use of English translations while the latter made use of two translations, English and Spanish. The resulting tagged corpus was impressive. The graphical content of the symbols was also examined to check whether similarity in graphical content could be an indicator for meaning of the associated gloss, and results suggest that such knowledge can further improve the tagging process. This chapter discusses different aspects of the research problem and the outcomes.

## 6.1 Corpus

The choice of corpus is a dominant factor for the performance of a translation task. It must be relevant to the domain and in a form and size useful for developing the system required. The selected corpus was a collection of subtitles extracted from movies and television programmes, which was collected by Lison and Tiedemann (2016), and the content was available in many languages. A significant part of these subtitles are direct dialogues between individuals which appear to be relevant to the needs of an AAC user to support him in daily face-to-face conversations.

Since the corpus is a critical step in solving the problem of translating between symbol and text, it was important to evaluate the selected corpus before any further steps. To carry out the evaluation, a sample of text that represents the domain was needed. However, such a sample was not available in Arabic as there appeared to be no available examples from those working with AAC users, and data collection from within the community was not possible due to time constraints. As an alternative, since many translations of the same corpus were available, the evaluation was carried out over the English side of the corpus. A small sample of adult AAC user sentences collected by Beukelman and Gutmann (1999) was used for the evaluation. Although the list was originally collected for adults with Amyotrophic Lateral Sclerosis (ALS), the content appears to be general and applicable to many AAC users. The evaluation compared several available corpora against this list of AAC messages (Beukelman & Gutmann, 1999). Results given in section 4.3 suggested that a subtitle corpus was far more relevant to the AAC domain than any other corpora examined. The sample was an important tool when searching for a relevant corpus.

Researchers have highlighted the importance of small talk in our daily conversations across different age groups (Ball et al., 1999; King et al., 1995). Small talk are utterances that do not carry much information but are used to engage in a social conversation. These utterances are unlikely to be found in corpora collected from news articles, encyclopaedia, or other formal documents. Therefore, selecting a conversational corpus is very important so that these utterances are available to the user. It should be noted that small talk utterances are repeated many times in such a corpus, but this redundancy should be retained, reflecting how common they are, which is useful information for AAC devices.

The differences between conversational corpora and other corpora is not only in the content but also in their structure. For instance, spoken utterances are expected to be shorter. Rice et al. (2010) examined children up to nine years old, and found the mean length for the oldest age group to be 4.99 words per utterance. On the other hand, adults in their forties have been shown to produce utterances with means of 9.49 to 13.60, depending on the topic (Nippold et al., 2014). Examining the subtitle corpus, the average sentence length was found to be 8.6 for English and 6.3 for Arabic. The corpus still contains some much longer sentences, often with more than one clause. On the other hand, many lines were single word utterances. Single word utterances were found to be more common in AAC messaging (Smith & Grove, 2003). The average sentence length was significantly shorter than those observed in newswire corpora. For example, Habash et al. (2011) reported an average length of 25 and 33 for Arabic and English respectively. This variation in the number of tokens per line is a strong indicator of how different a conversational corpus is compared to a newswire corpus. Thus, the linguistic tools required need to be trained on data that matches the domain to ensure plausible outcomes.

The content of the corpus was further examined. Since the corpus was not a collection of typical conversations, it often included unusual or unexpected events. The genre for the sources of these subtitles appeared to be mixed (no explicit indication of the genre), and as a result, some topics were over-represented (e.g. violence) compared to typical conversations between two individuals. For example, the word ‘walk’ appeared about 40 000 times while ‘kill’ appeared around 112 000 times, which is more than double.

The selected corpus provides conversational utterances, but some useful contextual data is not available, as the corpus did not provide indications of who had spoken each phrase or sentence (e.g. their gender, age, or role), when (e.g. lunch time or before bed), and in what setting (e.g. home or shopping centre). This missing information could be useful for AAC systems in order to improve their output. Furthermore, the corpus was originally a collection of files each a sequence of sentences ordered chronologically. However, this sequence was not exploited, although the wider context (preceding the current sentence) could be useful for disambiguation.

A limitation of this corpus is that these are not uninterrupted conversations between two people (as the data available for training dialogue systems). There are often shifts from one scene to another which may involve different groups of people, in different settings, and another time. The corpus did not provide any indication of these shifts. The absence of this information was considered to be a limitation that had to be accepted in the absence of a large set of AAC messages in either English or Arabic.

The Arabic subtitles were often translations of text originally written in some foreign language. Text translated into MSA does not necessarily match messages that have been written in MSA originally. Researchers have pointed out that there is a difference between original and translated text (Kurokawa et al., 2009; Lembersky et al., 2012) and were able to train classifiers that discriminate between them with high accuracy (Kurokawa et al., 2009). This difference, between translated and original text, may have a negative impact on the performance. Furthermore, local and important cultural or religious events are significantly under-represented in the corpus, compared to similar Western events. For example ‘Christmas’ appeared 17,942 times, while ‘Ramadan’ 86, ‘Hajj’ 73, and ‘Eid’ 88 times. The corpus did contain noise such as ill-formed machine translations, errors in sentence alignments and translations into a local dialect, but the majority was in an acceptable MSA form. Furthermore, the availability of translations in other languages was a valuable resource in disambiguating Arabic text.

Examining the vocabulary of the selected corpora showed that the Arabic vocabulary was more than double that of English, partly due to morphology and the number of prefixes and suffixes that can be attached to stems. Foreign words transliterated into Arabic also contributed to the Arabic vocabulary size, since there is no standard form, but usually multiple acceptable forms for a single foreign word. The reason behind this

is because a pair of languages do not usually share the same sounds, and so matching between sounds across the two languages is not one-to-one, and often there is no agreed map between the two. Proper nouns are plentiful in the English subtitles, and are often transliterated into several forms, which increases the vocabulary significantly. In general, the expanded vocabulary results in the Arabic side suffering from sparse data compared with English, which highlights the importance of lemmatisation.

## 6.2 Lemmatisation and Morphological Analysis

Lemmatisation is an essential step that has attracted much attention. The process of determining the lemma made use of evidence from the same corpus to avoid errors that result from domain mismatch. For instance, the distribution of verb inflections is likely to be different in conversational data than in long articles. The process groups multiple Arabic surface forms by their corresponding aligned English word. It collects statistics from the group of surface forms, which is a key to determine the likely lemma. The outcomes were promising, given how challenging Arabic text is.

However, the lemmatisation task requires the availability of a lexicon, in addition to awareness of possible morphological forms, prefixes and suffixes. The approach chosen was to generate possible hypotheses and rank them based on collected statistics. As a result, the accuracy of the process was limited by the coverage of the lexicon. If the actual lexeme is not part of the lexicon, it will not be among the generated hypotheses, and another lexeme may be accidentally selected. Thus, a lexicon with good coverage is key to the success of the lemmatization process. The coverage was expanded by collecting lexical entries from several sources (Wiktionary and WordNet, in addition to Buckwalter), to expand the dictionary used by the morphological analyser.

The lemmatization process assumes that text is tokenised and each token is processed individually. Tokenisation was done beforehand by inserting a space between every adjacent Arabic alphabetic and non-alphabetic symbols. However, the corpus contained some cases where the space between two tokens was missing. This problem was also pointed out by Buckwalter (2004). The majority of these cases encountered occurred when a word was preceded by a short token (two letters) ending with a non-concatenative letter. When examining word lists grouped by their English equivalent token, there would often be a correct form with a higher count. Thus, this common tokenisation error could be avoided if the correct form exists within a list, or more frequently, the untokenised form has no plausible analysis. The approach followed was to assume that attached preceding tokens could be one of three function words: **لَا** or **لِ** or **لِ**. This list could be expanded as necessary. For instance, the word ‘hundred’ was often attached to the preceding number (textual form) and could be handled in a similar fashion.

The corpus contained common spelling errors that were too frequent to be neglected. Such an issue was not special to this corpus but had also been observed in corpora from a collection of newswire data (Buckwalter, 2004). The majority of cases were alternating between various alef forms, final Yaa forms, and final ‘Haa’ versus ‘Taa marbuta’. Buckwalter solved these issues by adding to the morphological analyser’s dictionary, additional entries that covered these variations. However, this affects the accuracy of the lemmatisation process. As a result, various alef forms are always normalised because this is such a common error. Other errors were only addressed once hypothesised lemmas were ranked, and they were only corrected if they resulted in more likely lemma given the corresponding English form.

The process of determining the lemma was based on a single English word, and other surface forms aligned to that English word. However, a few cases were not resolved, which have similar surface forms and are often related semantically, but whose vocalised lemma is not the same. Also, determining the vocalised lemma of a few function words was a challenge since they usually do not correspond to any specific English word. These cases were handled through some rules, but the outcome was not reliable and led to inaccurate vocalised output. However, once lemmas are ranked, the final lemma per instance should take the context into account whenever ambiguities are not resolved. The final decision can be based on the likely form, given the context suggested by tools trained on some fully-vocalised text. This step of in-context lemmatisation was not carried out during this research, and is left for future work.

Lemmatisation is not the only task for textual processing. Determining the full morphological analysis of words which reveals possible attachments is needed, such as pronouns, which require a relevant symbol. This analysis is key to recovering the vocalised form. Unfortunately, knowing the main stem’s lemma of a surface word is not always sufficient to determine its full analysis. The local context of both English and Arabic were used to determine the likely analysis. The surrounding pronouns, e.g. subject pronoun on the English side, were often key in resolving surface form ambiguities. The most common ambiguities, which are difficult to resolve, are forms that contain a single second person pronoun. Determining whether the pronoun is masculine or feminine is a challenge since they are indistinguishable in English, as well as in their Arabic normalised form. Some clues on the Arabic side have been exploited, such as the analysis of any other verb in the second singular form or the gender of the following adjective or proper noun. Unresolved ambiguities are currently handled by randomly choosing one form. Another challenge was distinguishing between passive and active forms, since in their normalised forms, it is not easy to determine, and this was also left for future work.

### 6.3 Symbol to Text Translation

The task of translating symbols to text is defined as a user composing a message of one or more symbols followed by the translation process generating fully-formed text. The input stream is not a sequence of bare images but rather each image is associated with one or more glosses representing the meaning conveyed. As discussed in Chapter 4, from the translator's point of view, the problem is seen as a translation from isolated glosses into fully-formed sentences, without taking the pictographic symbols into account. The input list of glosses could be described as telegraphic sentences (Karberis & Kouroupetroglou, 2002; McCoy et al., 1998). Hunnicutt (1984) pointed out three kinds of grammar used with Blissymbols: telegraphic style, Bliss syntax, and natural language grammar, e.g. English. He described telegraphic grammar as lacking function words and giving less attention to word order. A symbol user may lack literacy skills, which prevents them from following the spoken rules for grammar, and this justifies the assumption of telegraphic input.

In order to train a translator that translates telegraphic text into fully-formed text, training data was needed. Training data is a parallel corpus, one side covering the input, in this case telegraphic text, the other side giving the fully-formed text. The telegraphic data was generated from the fully-formed sentences from the corpus. Preparing the telegraphic text required stripping out non-essential tokens that do not carry meanings, such as function words. This required awareness of speech tags to identify function words. As a result, prepositions, conjunctions and determiners were excluded.

Also, morphological analysis was needed (which was available at this point) to handle clitics and identify lemmas. Arabic pronouns are not always an independent token; they attach to the verb whether representing subjects or objects and thus need to be segmented and replaced with their full form since pronouns have their own symbols. Furthermore, Arabic is a partially null-subject language. The subject pronoun is not always explicit for verbs in the active form. The dropped pronoun can be inferred from the morphological form of the verb. However, it is hard to automatically determine whether the subject is explicit or implicit, and this has always been a concern for researchers annotating Arabic text (Hajic et al., 2004; Maamouri et al., 2004; Palmer et al., 2008). Pro-drop was found to occur in 30% of Arabic Treebank sentences (Palmer et al., 2008). This property occurs in many languages, e.g. Italian. This affects the choice of symbols; in some cases, a symbol representing the dropped subject had to be added. As a result, there is a symbol with a non-existent equivalent in the corresponding text. The dropped pronoun can be determined using a syntactic parser. However, with the availability of translation equivalent text in a language that does not have the same pro-drop issue, such as English, Arabic verbs will be aligned to a pronoun in addition to the verb whenever the subject is implicit or attached, which is



| Arabic token | English translation                   |
|--------------|---------------------------------------|
| اوافق        | i_PRP agree_VBP                       |
| يعطيك        | gives_VBZ you_PRP                     |
| ساخذ         | li_PRP 'm_VBP going_VBG to_TO take_VB |
| لكنه         | but_CC it_PRP                         |

Table 6.1: The null subject problem and agglutination showing the difference between Arabic and English, using the English Penn Treebank tagset

enough evidence to show that a pronoun was dropped. Table 6.1 illustrates the issue, where “I agree” is aligned to a single token and the verb has no pronoun suffixes.

The corpus contained some rather complex utterances, with multiple clauses and relative pronouns, which may not be relevant to the task. A process of filtering out complex utterances was carried out to speed up training and for a better estimation of the amount of data that actually matches the task. The resulting data was divided into a training set and a test set. A translator was trained with the training data and then an evaluation was carried out using the test set and the results recorded. However, this task is difficult since lots of the information needed to generate the output is missing from the input. For instance, the input will often provide no clue about a verb’s tense. As result, there are many plausible outputs given a particular input (covering various verb tenses and structures), but only a subset matches the user’s needs. The evaluation, however, was a comparison against one translation and is a common issue faced by those working in MT. It is worrying in the symbol to text case since the input is incomplete, which increases the space of acceptable translations.

## 6.4 Symbol Tagging

Translating from text to symbols requires the use of a specific symbol set or other semantics tags. The process undertaken here was to use a specific symbol set rather than general semantic tags. The problem is that there are many symbol sets, which differ in their depiction approach (pictographic vs. ideographic), drawing style, colour choices, number of dimensions, concept coverage, target language and target culture. They also vary in their licenses. The ARASAAC symbol set was chosen, but the approach followed was based on the characteristics commonly found in other popular communication symbol sets such as Widgit and PCS. The availability of multiple glosses in several languages is common across symbol sets, and some provided categorisation (linguistic or semantic) of the glosses. Therefore, because of similarities between pictographic symbol sets, it is suggested that automatic tagging, rather than manual tagging, be used for the generation of training data, because it can be repeated with several symbol sets.

right - something - way - people - thing - anything - look  
 - guy - home - someone - lot - thanks - one - kind - mom  
 - today - work - shit - idea - everyone - anyone - phone -  
 men - kid - tomorrow - stuff - fuck - minute - everybody  
 - somebody - chance - mind - fun - point - hour - story -  
 number - matter - tonight - nobody - game - deal - yeah -  
 side - couple - end - month - call - hey - child

Table 6.2: Top 50 nouns missing from the symbol set in order of frequency

gone - let - need - thank - mean - happen - believe - stay -  
 meet - excuse - suppose - wan - live - check - shut - hope -  
 huh - care - seem - fuck - mind - hate - trust - kid - promise -  
 felt - welcome - figure - end - wear - wake - become - matter  
 - imagine - relax - bear - prove - fell - spend - involve - set -  
 damn - wonder - expect - name - step - accord - deal - owe  
 - appreciate

Table 6.3: Top 50 verbs missing from the symbol set in order of frequency

ARASAAC is quite large, giving the impression that the symbol set will cover most frequently used words. However, once it had been examined against the corpus, it was clear that the set surprisingly lacked some essential glosses. There were often less frequent or more specialised glosses that had been included. For instance, at the time of writing, there was no gloss for ‘home’ yet there was a gloss for related concepts such as ‘home insurance’, ‘nursing home’ and ‘mobile home’. In this particular case, there was a gloss ‘house’ that could probably be used as a symbol for ‘home’, but this process cannot be automated. Table 6.4 demonstrates the problem by showing symbols containing the substring ‘story’ and no symbol for ‘story’ was found.

Another striking example was the lack of a symbol associated with the verb ‘need’, yet there was a more specific symbol ‘I need help’. Also, some symbols that represent the plural form of a word are available while the corresponding singular form is missing, such as kid and child which are missing, while ‘children’ and ‘kids’ are part of the symbol set. Some concepts are covered, but not with all possible part of speech classes, such as ‘welcome’ which has not been associated with the verb class. Table 6.2 shows the top 50 nouns, and Table 6.3 the top 50 verbs, that are missing from the symbol set when examined against the corpus, in frequency order. This limits the number of tokens tagged with a symbol, which in turn limits the number of Arabic glosses extracted. However, these limitations to the symbol set may be resolved in the near future, since the set is often updated by adding new symbols or glosses to existing symbols.

A number of symbols are associated with glosses that are not a single token but

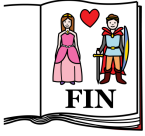



|   |   |  |   |
|---|---|--|---|
|  |  |  |  |
| and that is the<br>end of the story   | storyteller   | storyteller  | storyteller   |

Table 6.4: All symbols in the set having the substring ‘story’, showing the lack of a symbol for ‘story’

multiple tokens, which poses a challenge to the tagging process. Some glosses are fully-formed utterances, such as ‘I need help’, which can be used – in tagging the corpus – when an exact sequence match is found. However, this case prevents using the symbol with another pronoun or proper noun, e.g ‘he needs help’ or ‘Tom needs help’. In this particular example, dropping the subject pronoun, i.e. ‘need help’, may solve the issue, which can be done automatically. However, the content of the associated image may have an element that suggests the first person singular pronoun which results in a confusing symbol if the pronoun is dropped. Additionally, with multi-word glosses, intervening tokens are also possible, such as ‘I need medical help’, which is hard to handle based on sequence matching, and an intervening token may also need its own symbol. Furthermore, the word order may not always match. For instance, a symbol associated with ‘clean the glasses’ (Figure 6.1) which may be relevant to more than one word order, e.g ‘my glasses need to be cleaned’ as well as ‘I need to clean my glasses’, which requires a bundle of words to match, rather than a sequence. Handling multiwords is a challenge which NLP researchers often face (Sag et al., 2002).

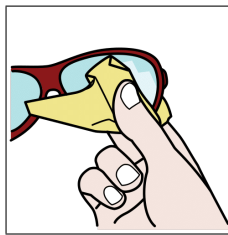


Figure 6.1: An ARSAAC symbol with the gloss ‘clean the glasses’

The process of determining the relevant symbol was based on matching glosses in two languages, to avoid errors that may result from matching based on a single gloss. Thus, to reduce ambiguities as much as possible, bilingual glosses associated with a symbol are matched against bilingual contexts in the corpus. However, MSA was not sufficiently covered in the symbol set, so Spanish was used as a pivot language between the Arabic context and the symbol. This approach minimised the chances of ambiguity and improvements were shown over single glosses.

Glosses which are of concern are those whoser ambiguity is a result of homonymy, as opposed to polysemy. Using more than one language significantly minimises this issue.

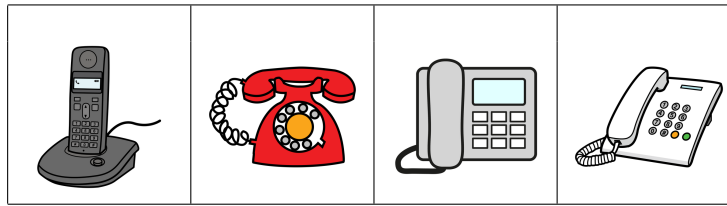


Table 6.5: Several symbols for the gloss ‘telephone’, showing no strong similarity

However, there are few symbols with glosses that are cross-linguistically ambiguous, which means that glosses in another language are not sufficient for disambiguation. This set of problematic glosses is supposedly a much smaller number than those that are ambiguous in one language. There is still a need for a mechanism to discover them for further processing. A multilingual conventional dictionary can be used to identify ambiguous glosses. However, the result is often overwhelmingly numerous, since dictionaries, such as WordNet, list many senses that are related and arise from polysemy (school as a building vs. school as an institution), and the focus is on the detection of unrelated meanings expressed in a single word form, i.e. homographs. In a perfect world, this can be discovered simply by finding all pairs of glosses (in two languages) that have been associated with more than one symbol. In practice, this does not work since the same meaning can be represented in multiple depictions, and not all senses a gloss may have are expected to be covered in the symbol set. The issue can be minimised by extracting contextual data for each symbol and using it to choose the best match.

Spanish was chosen as the second language for disambiguation in this research. However, distant languages are likely to perform better in disambiguation. Resnik and Yarowsky (1999) examined several languages, and suggested that there is a correlation between the distance between two languages and the likelihood that ambiguities in each language are lexicalized differently. They showed that 54% of sense distinctions are lexicalized differently (from English) in Arabic and 50% in Spanish. However, more distant languages such as Korean and Japanese were found to exceed 80%. Thus, the choice of language can make a considerable difference. Zhong and Ng (2010) used Chinese translations to infer WordNet senses for aligned English words, to expand WSD training data. The resulting tool performed remarkably well. Examining the potential of other languages in resolving symbol ambiguity is an important issue for future research. For instance, tagging symbols with Japanese glosses beforehand may reduce ambiguities beyond the current outcomes.

## 6.5 Visual Content of Symbols

The content of the pictographic symbols has been examined to gain insight into whether it can provide clues into the semantics of a symbol. Initial observations of the

ARASAAC symbol set (Chapter 3) showed that there were certain similarities, for example, between adjectives and nouns on specific topics. The literature cites many methods for extracting quantitative features that allow the distance between two images to be computed, which also reflects how similar they are to human perception. A threshold can be set, based on observation, to decide whether two symbols are similar or not. SIFT and HOG are two methods that have been used here, and results are encouraging (Appendix C and D).

Although symbols are simple images with often no background, the approach used here was able to identify visually similar symbols. Chapter 5 showed how the two methods differ. SIFT is able to identify symbols with common objects without being affected by size and position, which is often the case across symbols. However, an object must have some sophistication and not be represented with a few straight lines, such as a square. On the other hand, HOG was able to identify near-identical symbols; the same overall structure and major components are in approximately the same location with slight differences, such as different hair style. Both approaches are useful, but probably for different tasks. SIFT can be used to identify semantically-related symbols, while HOG can be used to identify almost identical symbols. Thus, HOG can be helpful to identify redundant symbols when, for instance, the number of symbols needs to be reduced. Hashing methods, such as the MD5 message-digest algorithm, is sensitive to slight variations (e.g. one pixel change), and as a result misses many symbols that are almost identical to the human eye.

Under the assumption that visually similar symbols are semantically similar symbols, glosses of similar symbols can provide clues to the context. For instance, the gloss ‘park’ can be ambiguous, but once knowing other glosses associated with similar symbols which include ‘playtime’ and ‘slide’, the ambiguity is resolved. Furthermore, neighbouring glosses can be used to collect more words that relate to the context. This was achieved by selecting two glosses and finding highly associative words with both. The availability of additional context is important since no context has been provided for the symbols. The data can be useful in a situation where multiple symbols are possible for a word or words in context, by ranking possible symbols based on the overlap between the target and contextual words associated with each symbol.

A single gloss can be associated with many symbols. Identifying whether a pair of symbols sharing a gloss are similar or not can be useful to know before the tagging process. The number of pairs sharing the same set of glosses (in two languages) was found to be 5 304. When similar pairs were filtered out, the numbers significantly decreased to 1 159. The number of pairs that shared at least one gloss (in two languages) was found to be 9 768, while filtering similar pairs reduced it to 3 122. This reduces the manual work needed to label these symbols with additional data related to their meaning. This highlights the issue that a proportion of the symbol set has content and meaning that is already in the symbol set but with slight variations.

However, the absence of similarity between symbols sharing a gloss is not an indication that the depicted meaning is different. For example, Figure 6.5 shows an example of symbols all showing a telephone but are understandably not considered visually similar. Such knowledge can only be built using machine learning methods, which requires plenty of images per gloss.

Awareness of visual similar content can be useful for many tasks and not only translation. It can be a useful tool for symbol maintainers and carers who need a faster method to be aware of the content of the symbol set. It provides valuable information for a symbol retrieval system by providing more textual content for each symbol and similar symbols that might be related to the topic. Manual annotation or categorisation for symbols can also benefit from knowing similar symbols by tagging all similar symbols at once.

The investigation so far has assumed pictographic symbols. Some other symbols sets cannot be handled in the same way. For instance, Blissymbolics was designed in a way that makes it easy to reproduce or draw on paper. As a result, drawings are simple and are not expected to provide features useful for pairwise compression. However, Blissymbols are constructed from Bliss-characters that form Bliss-words. Thus, having overlapping characters can be an indication of semantic relatedness between two symbols. Therefore, the method used for visual content awareness needs to be relevant to the overall design of the symbol set.

## 6.6 Text to Symbols Translation

The automatic tagging approach has made training data available. However, is the process of finding relevant symbols a translation or a tagging task? MT is a complicated process because the output has a different order, and word correspondence between the source and the target is not one-to-one. Sense ambiguity must be handled implicitly. The output needs follow the grammatical rules of the target language, which may require adding function words and ensuring agreement between dependent tokens. Symbol communication on the other hand, apart from Blissymbols, often do not have their own syntax, but rather depend on the spoken language for their order. As a result, the task of producing training data might be considered a tagging process rather than a translation. However, the word to symbol relationship is not one-to-one, but rather a single word may be tagged with no symbols, or many symbols, and a single symbol may cover one or many words.

The data is thus ready to be used to develop a symbol tagger. NLP involves many tagging tasks such as POS tagging, WSD classifiers and named entity recognition (NER). There are many approaches to the tagging task and recent approaches are based on various neural network architectures that can operate at word level or at

character level or both. For instance, a bi-directional LSTM with CNN (Chiu & Nichols, 2016) was an architecture used for tagging that learned word features as well as character features. Recent approaches are based on the transformer (Vaswani et al., 2017) such as the architecture proposed by Baevski et al. (2019), which outperformed NER state of the art results. Having the data, the process of training and evaluation is straightforward and left for future work.

## 6.7 Summary

This chapter discussed several aspects of the problem of symbol and text translation and carried out experiments. The choice of corpora is important. Conversational data contains small talk utterances, which are important for face-to-face conversation. It is also structured differently to other types of text. Since the corpus was collected from movie and TV scripts, the distribution of some of the words may not reflect typical conversations. Also, some contextual information is missing from the corpus, which might improve the performance of AAC systems. The Arabic translation used in this research was in an acceptable form, but did not cover some local cultural concepts. MSA is rich in morphology and agglutinative and as a result morphological analysis is an essential task.

Text pre-processing was carried out before creating the training data. This included lemmatization and further morphological analysis. Lemmatization made use of local evidence to avoid domain mismatch errors. The coverage of the lexicon is a critical aspect of the lemmatization approach and should cover all vocabulary used frequently. The approach used for lemmatization was also useful in correcting spelling errors. Few lemmatization cases were hard to address, and increased data is expected to further improve the outcomes. Additional morphological analysis was made knowing the lemma and accessing the wider bilingual context.

Symbol to text translation and related experiments were discussed. The translator was trained using data from the corpus after creating a fully-formed and a telegraphic transformed text. The task of translation is challenging since the input is incomplete and, as a result, the space of possible valid translations is large and only a few may match the user's need.

Translating from text to symbols was investigated. The ability to determine the relevant symbol to words in context requires knowledge that was missing. Multiple corpus translations were exploited to minimise errors that result from ambiguous glosses. Symbols with multi-word glosses were a challenge. Limitations include missing applicable glosses and also symbols. Further contextual data was obtained through awareness of symbols with similar visual content. This contextual data can be used to determine the best matching symbol for a given word in context.

Multiple tasks were completed in order to achieve the results aimed for. The outcomes are promising, with a clear indication that it is possible to achieve a sufficiently successful symbol to text and text to symbol training data approach with limited corpora, whether it is a small pictographic symbol set or a highly morphologically rich language such as Arabic.



## Chapter 7

# Conclusion and Future Work

Graphical symbols are sometimes used for communication when there are barriers that prevent an individual from communicating using the various forms of spoken or signed languages. Users of graphical symbols need to interact with others around them who may not be familiar with symbols, and they also need to access resources expressed in local spoken languages. Therefore, it is essential they have an approach to automatically translate to and from graphical symbols, which could have a great impact on their quality of life. The translation process needs to resolve ambiguity to determine relevant symbols and generate plausible text in the local language. Such a process can be developed using data based methods. However, data tagged with symbols is not a resource yet available in many languages and graphical symbol sets. The aim of this research was to address the lack of such a corpus by generating training data useful for training a translation system. The generation process avoids the need for manual tagging, which would be costly in time and effort needed to produce such a resource. This study focused in particular on Arabic and the ARASAAC symbol set.

## 7.1 Conclusions

This thesis first reviewed the relevant literature in Chapter 2. Then publications on symbol to text and text to symbol showed that most relied on existing language-specific tools. This research therefore focused on making the data needed available, while highlighting the importance of data. Related aspects of language processing were assessed, such as word sense disambiguation (WSD), an issue somewhat similar to symbol tagging but with a different kind of inventory, and the use of translation to overcome the missing manually annotated resources. Previous research concerned with machine translation was briefly described, which highlighted the sophistication and advances made in the field that could be beneficial to AAC once the data is available.

Many communication symbol sets are available. The ARASAAC symbol set was selected, and examined in Chapter 3. This set was chosen due to its licencing freedom, its availability, size and the pictographic depiction style. Having a symbol set, along with its associated glosses, allows a user to enter a word and find pictographic symbols based on text matching between the query and the glosses, which can result in several matching symbols that may or may not have the same meaning. However, such a resource is not sufficient when there is a word or phrase in context that needs to be translated to a relevant symbol that conveys the same meaning, due to the lack of knowledge needed for disambiguating the exact word meaning. It is also is not adequate for building a system that translates a set of symbols into text, either for speech synthesis or for understanding by another person, due to the absence of sequences of text needed to learn the grammar of the target language. Therefore, further data was needed.

In order for a machine to grasp the grammar of the target language, plenty of sentences of various lengths were needed. The relationship between text and symbol can be learned through observing many sentences and their corresponding symbols. Hence, a textual corpus was needed, and such a corpus need to have corresponding symbol messages. This resource is not currently available and but it is essential, as described in Chapter 4.

Several steps were carried out to generate the training data needed. First, a corpus needed to be collected from scratch or an available base corpus needed to be chosen. Once a corpus is available, it needs to be linked with communication symbols. However, the choice of corpus is not arbitrary, but needs to match the domain and be large enough to be useful for machine learning.

In order to match the domain, a search was initially conducted for a corpus of actual AAC messages collected from users. However, such a resource large enough for the task was not available. A search was then made for a corpus that is conversational and large. A corpus of subtitles (TV and film) expressed in the target language was chosen. Finding and using domain-relevant data is an important step. In order to confirm the relevancy of this corpus to the target task, an experiment was carried out that made use of a list of messages prepared by experts in the field (given in Appendix A). Such a list has an important role in determining the best corpus. Unfortunately, such a resource may not be readily available in other languages, but it is valuable for evaluating available corpora, and obtaining such a resource ahead of time is recommended. The experiment showed that the corpus (the English translation) was significantly better at matching the domain than other publicly-available corpora. The corpus was also big enough and translations were available in many languages, which was an important aspect for this research.

The raw corpus is not sufficient on its own and required tagging with various grammatical markers, including symbols. The proposed solution was an automatic tagging approach, eliminating the need for manual tagging, which is costly in time and effort, and allowed the tagging process to be repeated for any pair of languages and symbol set. The approach created plausible tags by finding ways to minimise ambiguity. This approach therefore has the ability to create training data for developing systems that can translate to and from symbols.

The tags needed were not only symbols but other linguistic information such as lemmas. The main challenge from a machine point of view, when automatically tagging a corpus, is the ambiguity of the text. This came in two forms: word lemma disambiguation and word pictograph disambiguation. The tagging process can make use of existing linguistic taggers, such as part of speech taggers and WSD taggers. However, these tools were not developed to handle conversational data, which is significantly different than other type of corpora such as newswire corpora. This is especially true for highly ambiguous text such as Arabic, where several inflections of a verb appear identical due to the absence of diacritics and are often not preceded by subject pronouns, which can be a key in determining the verb surface form. Apart from domain mismatch, WSD taggers cover a limited vocabulary and there is no WSD Arabic tool that is sufficient for disambiguation. Therefore, the disambiguation needed for tagging relied mainly on the availability of translations in other languages.

Determining the exact lemma for a surface word in context may not be an issue for English, which can be resolved using a POS tagger. However, it is a challenge for Arabic. Knowing the lemma is essential for creating training data for two main reasons. First, symbols are often only tagged with the lemma form of a word. Secondly, identifying the lemma is a key in the restoration of absent diacritics which are crucial for speech synthesis in speech generating devices. Therefore, the first step to be carried out was tagging with the lemma.

The lemma disambiguation process requires an inventory of lemmas, preferably with their metadata. It also requires a process that suggests possible lemmas for a given surface form. Given several options, a single lemma was selected based on a ranking criteria. The knowledge needed for ranking was obtained by making use of an English translation equivalent corpus that was initially sentence-aligned. For the purpose of disambiguation, word alignments were obtained. Grouping several surface Arabic forms by a single English translation equivalent provided evidence that allowed the determination of the likely lemma, regardless of the context of each instance. Once the lemma was identified, further morphological data was determined for each instance from both the context and the lemma.

Thus, tagging process did not make use of any existing tools trained on annotated data. Instead, it used dictionaries, a rule-based analyser, a parallel word-aligned

English text, and heuristics. The results were encouraging because it now allows automatic morphological tagging for any domain-specific text, as long as a parallel English text and domain-specific dictionary exists. The process avoids errors of existing taggers that can be a result of domain mismatch or insufficient evidence provided by the monolingual context. The outcome can be further improved by using an existing analyser as a backup whenever ambiguity cannot be resolved. The resulting tagged corpus, detailed in Chapter 5, can be useful not only for symbol and text translation, but also for word based communication and word sense disambiguation.

The translation process from text to symbols needs tagging the corpus with symbols. However, awareness of symbols is not necessary for translating from symbols to text, since glosses alone can be used to generate a translation. An input symbol message represented by glosses can be simulated by transforming a sentence into a telegraphic form. Thus, it becomes a translation from telegraphic text to fully-formed text. Consequently, this training data can be directly used with any symbol set. An experiment was carried that made use of this monolingual parallel corpus by training a translator and testing it, as discussed in Chapter 5. However, translating from morphologically poor languages to morphologically rich languages is known to be a challenge in machine translation, since much information is missing, e.g. the adjective gender. In practice, the translation task is even harder due to potential errors in word order and the many possible glosses a specific symbol may have.

The task of translating text into symbols requires awareness of the actual symbols. Even having symbols with associated glosses and a corpus tagged with lemmas is not sufficient for symbol tagging, due to potential ambiguities. As with the lemma disambiguation, the symbol disambiguation was based on translations. However, MSA glosses were not available. Relying on the English equivalent to find the corresponding symbol is not sufficient, because this English word on its own can be ambiguous.

This was solved by introducing an additional language for disambiguation. The choice of language required coverage of the symbol set through glosses, and also some available parallel corpora that included Arabic. Spanish was chosen to operate together with English for disambiguation. For each English/Arabic aligned pair of words, a relevant symbol is one with an English gloss matching the English text word and a Spanish gloss that is the translation equivalent of the Arabic text word. The part of speech tag on the English side was also used to minimise ambiguity. The resulting tagged corpus, reported in Chapter 5, was examined comparing symbol textual contexts with gloss textual contexts, and the positive impact confirmed. A few cases remain an issue where both the English and Spanish glosses are cross-linguistically ambiguous.

To further improve the disambiguation needed for symbol tagging, the visual content of the selected symbols was examined. The goal was to obtain textual context by comparing the glosses of visually similar symbols. This was motivated by the scarcity

and ambiguity of associated glosses. Experiments showed that some algorithms were better at identifying similar symbols than others (see Chapter 5). Glosses in the symbol's neighbourhood were used to obtain further contextual data. Resulting contextual data could be used to further improve the tagging process, for instance by ranking symbols based on their similarity with the current context.

This approach in deriving the tagged corpus will be of value to future research into symbol to text translations. The process of tagging is not specific to the chosen symbol set, and as a result, the same corpora can be tagged with various symbol sets at no cost. The approach is not limited to a certain language as long as an inventory of lemmas and a morphological rule-based analyser is available.

Chapter 6 discussed the crucial selection of tools and resources to make sure they match the domain. The size of the corpus is also vital, since it needs to be large enough for reliable translation equivalent pairs to be extracted, and that noise can be filtered out. The average frequency for an English word in the corpus used was over 400, and using a smaller parallel corpus is unlikely to yield plausible results. A few basic monolingual processing tools are essential for pre-processing. For instance, awareness of the part of speech tags for at least one language had a positive impact on the disambiguation process. Another tool was responsible for generating all possible morphological analyses. The importance of these monolingual tools grows with the complexity of the target language. For example, rich morphological languages will need a morphological analyser to handle the large vocabulary, often with sparse counts.

## 7.2 Limitations

Several limitations of this research need to be acknowledged. First, Modern Standard Arabic (MSA) was the language version used for symbol glosses and the corpora. It is not the colloquial spoken language, despite the need for conversational corpora. Arabic speakers do not use MSA in informal settings, and as a result MSA may not sound natural to them. There are many local Arabic dialects with little documented grammars about each, which makes them difficult to process computationally.

The choice of corpus imposed two limitations. First, the majority of subtitles were originally written in English and translated to other languages, including MSA. Thus, the resulting syntax might be biased toward an English-like syntax. Also, the content of the corpus does not cover cultural aspects of the Arabic community. Secondly, such a corpus does not provide information about the structure and content of a symbol message an AAC author may compose, and such knowledge remains missing.

Other limitations include the coverage of the symbol set with respect to words in the corpus. Although, the symbol set was large, many concepts are depicted in many

alternative ways, which increased the size of the symbol set without widening its reach. Other symbols are tagged with long utterances, which are not as frequent as words, and thus their usability from a tagging perspective is limited. Further, existing glosses associated with each symbol do not include possible synonyms and the majority of symbols are associated with one gloss only. Finally, the dictionary that provides lemmas gave plausible coverage but some lexical items were missing, which meant that corresponding lemmas in the corpus were left with no symbol tag even when a relevant symbol was present in the dataset.

### 7.3 Future Work

Despite these promising results, there is space for further improvement. In practice, the tagging process is an iterative process and not a single event. Erroneous output can be collected and used to improve the tagging process. Further, challenging symbols can be identified and a manual tagging task, focusing on only those few symbols, can be carried out to further improve the accuracy.

Further improvements can be obtained by not relying on existing glosses provided with the symbols. This was an obstacle to benefiting from the actual potential coverage of the symbol set. Thus, a manual review of the associated glosses is recommended before the tagging process, so that additional glosses can be added to existing symbols and erroneous ones corrected. The process of expanding associated glosses can be improved by knowing semantically-related words as informed by a relevant corpus that can be identified through vector space models (Mikolov et al., 2013; Turney & Pantel, 2010).

Awareness of visually similar symbols in advance can speed up the tagging process. This process of the expansion and correction of glosses can be manually done with at least two languages, and implicitly transferred through parallel corpora to other languages. Such a suggestion may appear to contradict the automatic tagging approach but it does not. Symbols number in the thousands, but words that needs tagging are in the hundreds of millions (for instance the selected corpus section 5.1), and so the tagging effort for symbols is limited, yet significant improvement is expected. Further, additional metadata can be added to symbols, such as related topics. However, when tagging the corpus, it is vital to discriminate between the two types of tags, – glosses and related words – since each has its own role and should not be mixed.

Other future directions could be repeating the tagging process and changing the symbol set or the language. Although ARASAAC was the pictographic symbol set chosen, due to its coverage and availability, other symbol sets and languages can be used to generate training data and consequently systems are needed to support AAC use in different settings. In addition to AAC devices and applications, these systems can be integrated into text simplification and web page translation tools, designed for

people with cognitive impairments who have also been aiming to aid literacy skills. This allows the content to be linked with AAC symbols to meet the user's need<sup>1</sup>.

The specific target group in mind when choosing the corpus was Arabic speakers. The corpus chosen was Modern Standard Arabic, with the aim of helping a large population of Arabic speakers. However, future work may look into a way of making use of the current findings with local Arabic dialect dictionaries for tagging a local dialect corpus with relevant symbols. Research into Arabic NLP initially focused on MSA, but later paid attention to other dialects, such as Levantine and Egyptian dialects (Habash et al., 2013; Maamouri et al., 2006), and recently to several more dialects. For example, the Gumar Corpus (Khalifa et al., 2016) is a collection of novels written in various dialects spoken in the Arabian peninsula. Also, a corpus of dialects has been created by translating "1,045 concepts with an average of 45 words from 25 cities per concept" from English or French sentences in the travel domain into several Arabic dialects (Bouamor et al., 2018). The same approach can be followed to create an AAC corpus by, for example, extracting samples of sentences from the subtitle corpus with high probability and translating them to local dialects. The future plan is to focus on major Saudi dialects.

## 7.4 Summary

This research investigated the possibilities of automatically tagging an Arabic corpus with relevant communication symbols given the lack a tagged corpus. The resulting corpus is suitable for any data based machine learning tool. The tagging process involved the use of an open licensed pictographic symbol library that was analysed in detail for its visual attributes and other associated data. The approach taken to tag the MSA corpus was by using a collection of subtitles that had been shown to closely match the AAC user's needs. The tagging process made use of translations to identify relevant symbols and MSA vocalised lemmas. The approach adopted for the symbol to text task was a translation from telegraphic to fully-formed text, and subsequently a monolingual corpus was created. The visual content of symbols were used to obtain further contextual data that could be useful for the disambiguation process. The resulting data was examined and promising results were investigated. The limitations are related to the choice of language, the corpus, the symbol set and the dictionary. However, the processes involved have shown that it is not only possible to better support text to symbol and symbol to text in multiple languages, but to also start the journey of giving AAC users enhanced automatic personalised symbol to text translations on the web with the mapping of symbols across different symbol sets and languages, as was suggested by Mats Lundälv with the Concept Coding Framework over ten years ago.

---

<sup>1</sup><https://www.w3.org/TR/personalization-semantics-content-1.0/#symbol-explanation>







## Appendix A

# Message List

### **Generic Message List for AAC users with ALS**

Prepared by David Beukelman and Michelle Gutmann

November, 1999

---

#### **Greetings**

Hi  
Hello  
Good morning  
Good to see you.

#### **Opening Questions**

What's new?  
How are you today?  
What's happening with you?

#### **Responses**

I'm OK.  
Could be better.  
I am getting along.  
Not very good today.  
I like that.  
I don't understand.  
I don't know.  
I don't think so.  
It doesn't matter, I guess.  
It is important to me.  
It is not that important.  
I am sorry to hear that.  
Really?

#### **Conversational Continuers**

Really  
Alright  
Isn't that wonderful (great)  
That's good

I see  
I know it  
Okay  
Yeah  
Good  
Uhhuh

### **Conversational Turnarounds & Extenders**

What about you?  
What do you think about that?  
What have you been doing?  
Tell me about your family.  
That's interesting, tell me more.  
Thank you.  
You're welcome

### **Resolving Communication Breakdowns**

I changed my mind  
Let's try that again  
Let's do it another time.  
Tell me you what think that I said.

### **Personal Care**

I need you to...  
I would like for you to...  
I need some help with...  
Can that wait until another time?  
Just a minute, I'm not finished.  
When will you be back.

### **Good-byes and Farewells**

Thanks for stopping by.  
Come back again.  
Great to see you again.  
See you soon.  
Good night  
Good-bye  
Use of Telephone  
I'd like to talk to·

This is (Name), I have a speech problem. I use a machine to talk. Please be patient.

The number I am dialing is \_ \_ \_ \_ \_

How are you?

I'll talk to you soon..

Call me back when you can.

Do you understand me?

### **Meeting New People**

Hi, I'm (Name). I can hear and understand everything that you say. I have ALS/Lou

Gehrig's disease and I have trouble speaking. I use this machine to communication. Give me a minute.

Please tell me if you don't understand what I am saying.

### **Health and Safety**

This is an emergency.

Get help now!

I need suction!

### **Vocabulary for Support Groups or Conversing with Others about ALS**

Having this disease has made me...

I worry about...

I fear the loss of...

I can't think about...

It makes me really mad that ...

I am determined to...

One good thing about this is...

### **Clinic Appointments**

I need to see the doctor about...

I need to make an appointment for...

My seating, wheelchair/computer isn't working.

I have noticed that...

What's next.

I need information about...

## Appendix B

# Examples Showing Arabic Morphological Variations

A subset of Arabic words (247 forms) that have been aligned the word ‘read’ at least 5 times showing only forms based on the verb قَرَأَ

قرات - اقرا - قراءة - تقرا - القراءة - قراته - يقرأ - اقراها - قراتها - قرا - اقراه - قراءتها - لقراءة - قراءته - اقراي - بقراءة - ساقرا - نقرا - تقراي - تقراه - تقرأين - تقرأها - قرائتها - قرانا - للقراءة - قراتي - قرائته - اقربي - يقرأون - وقرات - واقرا - يقرأها - ستقرا - اقراء - قراه - اقربها - يقرأه - قراتم - قراتهم - سيقرا - القراءه - ساقراه - قراة - قراها - تقرأين - قرائة - تقريه - تقرأين - لقراءته - ساقراها - اقراه - وقراءة - قراءه - قراوا - اقراوا - بالقراءة - اقرات - وتقرا - اقراهم - بقراءته - لاقرا - لتقرا - تقريها - قراتيه - لقراءتها - اقراء - اقراي - القراة - بقراءتها - يقرأوا - تقرأون - اقريها - سنقرا - اقربها - ليقرأ - قراتي - تقرأى - نقراه - تقرأين - اقربيه - بقرائتها - تقريه - لتقرا - لتقرا - تقريها - قراتيه - قراءتك - القراءة - ستقراين - نقراها - لاقراها - قرانا - وقرا - لتقراه - قرائتهم - اتقرا - يقرؤون - لقرائتها - قراتك - ونقرا - قراءت - اقراي - تقراهم - لاقراه - وقراته - ويقرأ - لقرائته - وقراتها - قرائتك - والقراءة - تقريها - وساقرا - قرائتي - اقروها - واقراي - قراء - ستقراه - واقراها - قراناها - اقرووا - لاتقرا - تقريه - بقراءة - فلتقرا - سيقراها - تقريها - قرائه - لتقراها - اقراة - ليقرأها - ستقراها - اقروه - قراة - سيقراه - اقراها - ستقريين - وقرانا - يقران - اقراو - يقرأ - قرئت - بقراة - قراتما - تقرأن - اقراي - اقراها - واقربي - قرائتهم - اقربهم - بقراهه - لقراة - سيقراون - بقراته - قراءتي - سنقراها - تقري - يقرأهم - تقراوه - اقراه - ليقرأه - يقرؤوا - قراتي - يقرأوا - قراهم - اقربه - تقريهم - لقراة - للقراءه - واقراه - تقراة - اقريها - قرؤوا - تقريها - اقراي - القراء - ساقروها - تقراي - ساقراهم - فلتقرا - وتقرأين - وقراءته - قراوه - لتقراي - فاقرا - يقراني - اقراه - بقراتها - اقروها - اقروا - وتقراي - اتقراين - يقرأونها - لتقراها - ماقراته - تقروون - وقراءتها - قرواها - تقروه - يقرأونه - قراة - لقراءه - لتقريه - تقريه - بقرائتهم - اقري - قراة - اقراه - اقراها - ساقروه - قارئ - اقريهم - اقراهم - قراتموه - اقراي - تقريها - لتقراه - اقراك - لتقري - يقرأها - قراو - يقرؤن - ستقراون - اقروا - وساقراها - لقراته - اقراه - قريت - وتقرأها - ساقراء - ويقرأون - واقراه - ليقرأوا - ستقراي - لقرات - استقرا - ساقريها - تقرائه - تقراو - سنقراه - القرائه - اقري - ساقراء - قريها - وقراي - اقراه

A subset of Arabic words (163 forms) that have been aligned the word ‘go’ at least 5 times showing only forms based on the verb ذَهَبَ

ذهب - الذهب - نذهب - تذهب - ساذهب - اذهبي - يذهب - للذهاب - اذهبوا - ذهبت - سذهب  
- تذهبي - لنذهب - بالذهب - ستذهب - تذهين - ذهب - يذهبون - اذهبي - اذهبا - واذهب -

سيذهب - ذهبنا - سندهين - تذهبوا - وتذهب - يذهبوا - تذهبي - ونذهب - لتذهب - ذاهب -  
 لاذهب - والذهاب - فلتذهب - لاتذهب - ذاهبة - تذهبون - ذهاب - فلنذهب - ذاهبون - اذهبوا -  
 ذهابك - واذهي - لذهاب - وساذهب - ليذهب - ذهبتى - تذهبا - سيذهبون - ويذهب - ذهبوا - ذهابي -  
 - فاذهب - تذهبان - وسنذهب - اذهبن - ويذهبون - ذهبتن - سندهين - لاتذهبي - لتذهبي - الذهب -  
 ذهابا - يذهبن - فلتذهبي - يذهبان - وذهبت - واذهبوا - ذهابنا - فسادهب - لاندذهب - ذاهبه - فاذهي -  
 - ذاهبين - ذهبتنا - وتذهبين - تذهبن - لاتذهبين - ذهبتى - سندهيان - يذهبا - ليذهبوا - وتذهبي -  
 سندهبي - وستذهب - لذهبت - وذهب - اذهب - انذهب - اذهبت - ذاهبا - لتذهبوا - فليذهب -  
 يذهبوا - فلتذهبوا - تذهبوا - ذهبوا - ذاهبان - استذهب - لذهابك - فالتذهب - ولنذهب - لاتذهبوا -  
 ذهابه - واذهي - سيذهبوا - ذهابكم - وسيذهب - سيذهبان - ويذهبوا - بذهاب - لتذهبي - والذهب -  
 فاذهبوا - فذهبت - فستذهب - بذهابك - لاذهب - فتذهب - واذها - سندهبي - اذهاب - فسيذهب -  
 - اذهب - وستذهبين - لذهب - اتذهبين - بذهابي - ويذهبن - ذهابى - غذهب - لاتذهبون - فالتذهب -  
 - لا يذهب - لتذهبين - لاتذهبي - فنذهب - اسنذهب - للذهب - وذهبوا - وتذهبون - سذهبوا -  
 وسيذهبون - اذهبه - ذهبي - لا يذهبون - نذهب - ذهبن - سيذهبن - سذهب - ذهبتوا - ذهابهم - بذهابه -  
 - فلتذهبي - تذهبه - فذهب - بذهابنا - ذاهبات - وتذهبي - اذهبنا - بالذهب - فلتذهب - ليذهب

A subset of Arabic words (347 forms) that have been aligned the word 'say' at least 5 times showing only forms based on the verb قَالَ

قول - تقول - قلت - يقولون - قول - يقول - القول - قل - قوله - قال - تقل - ساقول - اقل - تقولي -  
 نقول - اقول - تقولين - تقوله - قولي - قالت - لنقل - ستقول - قلها - لا قول - يقل - قالوا - لقول - قلته -  
 يقول - سيقول - اقولها - لتقوله - ساقوله - يقال - قولها - لا قوله - قولك - واقول - تقولها - قولها - ستقوله -  
 - قاله - قلتي - ويقول - يقولونه - تقولينه - ستقولين - يقولوا - تقولى - لتقول - وتقول - لنقول - سيقولون -  
 لقوله - للقول - ليقول - تقولون - قولوا - ساقولها - بقول - ويقولون - فلنقل - قولى - سنقول - تقولينه - وقل -  
 - سيقوله - قالت - ليقوله - نقل - نقوله - لاتقل - تقولها - قله - تقولوا - يقولها - لتقولينه - ستقولينها -  
 ونقول - اقلت - قلتها - لقلت - لاتقولي - قلنا - اتقول - تقولينها - تقلها - قلتيه - وقلت - قلتي - لتقولي -  
 - ماتقوله - بالقول - فساقول - قالوا - قالوه - نقولها - وقول - بقوله - وتقولين - وساقول - قوله - اقال -  
 ليقال - لقولها - وقولي - قوليه - تقوله - ماقولك - اقله - قولكم - يقولونه - فلنقل - لاتقول - قلتم - قولوا -  
 يقولان - يقلن - لنقله - تقولوا - اقلت - لا قولها - تقولونه - قلت - تقولان - اقلها - قالا - قليها -  
 مايقولونه - فقل - بقولك - يقولونها - اتقولين - سنقوله - ستقولي - فاقول - ليقولوا - وقالوا - فقلت - لقولي -  
 - قائل - وتقولي - مايقولون - وقال - ستقوليه - ستقولون - فقله - قولها - تقله - تقال - قلى - ولنقل -  
 ماتقولينه - لتقولينه - قالها - سيقولونه - قيل - ماقله - وستقول - والقول - ماتقول - وساقولها - تقولونها -  
 يقولن - بقولها - لتقل - قولوا - اقله - ليقولوه - يقولوا - لقولك - وسيقول - ماقلته - لا يقول - اقولك - وقلها -  
 - ستقولها - ماستقولها - يقولوا - ليقولها - لتقوله - مايقال - وقولوا - ماقول - فقولي - اقول - قلناه - لا قول -  
 - قولهم - تقولوه - سيقولوا - اقولها - قلتما - ساقوله - وسيقولون - ويقولوا - تقولانه - بقولي - ماساقوله -  
 سيقولها - قائلة - ليقل - يقوله - قلن - قلتيها - يقلها - فلتقله - لا قوله - فتقول - ستقوله - لتقولى - لتقولها -  
 - وسنقول - ستقولى - المقوله - قولتي - يقله - قائلين - يقولها - فلتقولي - لقال - ماتقولين - وساقوله -  
 قللك - لاقل - سيقال - وستقولين - لتقولين - ستقولان - تقلن - ستقولونه - لانقول - استقول - يقولونه -  
 لاتقولها - قلته - يقلون - فليقل - قالتها - لاتقلها - مانقولها - ستقولينها - ماقولكم - تقولها - لتقولها -  
 يقلنه - سيقولان - يقولانه - القائل - وقالت - فلتقلها - وتقولها - لاتقولين - اقالوا - فسيقول - وقولى -  
 ماقلت - مايقوله - فقولي - لاتقولها - سنقولها - وقولها - ماستقولينه - وقلنا - تقولوا - لتقولوه - ليقولون -  
 بتقول - اقولك - لاتقولى - فلتقولها - مقوله - ويقال - فالنقل - ماقوله - وقله - قولكم - لقوله - فلنقول -  
 ليقولونه - قله - فيقول - ولنقول - ماتقوليه - مايقول - هتقول - ماسيقوله - ولاتقل - ولاقول - لتقولها -  
 قالوها - وتقولى - لتقولوا - لقولى - باقول - لقلت - قولته - لا يقولون - قولتي - فلتقولها - تقولك - سيقولوه

- يقولنه - اقولهم - يقولك - لتقال - فيقولون - وقلها - فلتقولوا - استقولين - فستقول - واقوله - سيقوله -  
 قولتيه - لتقولاه - لتقولونه - نتقابل - اقولهما - فتقولين - تقولية - ستقوليهما - قلى - ليقولا - فلتقوليه - قائل  
 - ستقولوه - انقول - وتقولون - قالاه - ماقولة - وتقولا - تقلي - بقولون - ستقولانه - ماستقول - ستقل -  
 سنقوله - ماقالوا - وقلتي - فلتقول - فلتقوليهما - اتقولون - قلتيه - ليقولوها - فساقله - ايقول - ماساقول





Appendix C

SIFT ARASAAC Examples





|                 |   |
|-----------------|---|
| main symbol     | <br>to plan  |
| similar symbols |  to plan  to plan  to organise |

Table C.1






|                 |  |
|-----------------|--|
| main symbol     | <br>to dance  |
| similar symbols |  to dance  to dance  movement disability*  to dance |

Table C.2






|                 |  |
|-----------------|--|
| main symbol     | <br>crisis*   |
| similar symbols |    <br>lower    to improve    economy*    <NONE> |

Table C.3





|                 |   |
|-----------------|---|
| main symbol     | <br>to sit on the edge   |
| similar symbols |   <br>to sit on the edge    to breathe underwater    to sit on the edge |

Table C.4





|                 |  |
|-----------------|--|
| main symbol     | <br>sailor*   |
| similar symbols |   <br>mariner*    tug*    to tug |

Table C.5

Appendix D

HOG ARASAAC Examples




|                 |  |
|-----------------|--|
| main symbol     | <br>to plan   |
| similar symbols |  <br>to plan    to organise |

Table D.1





|                 |   |
|-----------------|---|
| main symbol     | <br>to dance   |
| similar symbols |   <br>to dance    to dance    to dance |

Table D.2


|                 |   |
|-----------------|---|
| main symbol     | <br>economic crisis* |
| similar symbols |   |

Table D.3




|                 |   |
|-----------------|---|
| main symbol     | <br>to sit on the edge   |
| similar symbols |  <br>to sit on the edge    to sit on the edge |

Table D.4



|                 |   |
|-----------------|---|
| main symbol     | <br>naval officer* |
| similar symbols | <br>seaman*        |

Table D.5

## References

- Agirre, E., Ansa, O., Hovy, E., & Martinez, D. (2001). Enriching WordNet concepts with topic signatures. arXiv:cs/0109031. <http://arxiv.org/abs/cs/0109031>
- Al-Arifi, B., Al-Rubaian, A., Al-Ofisan, G., Al-Romi, N., & Al-Wabil, A. (2013). Towards an Arabic Language Augmentative and Alternative Communication Application for Autism. In A. Marcus (Ed.), *Design, User Experience, and Usability. Health, Learning, Playing, Cultural, and Cross-Cultural User Experience* (pp. 333–341). Springer Berlin Heidelberg. [https://doi.org/10.1007/978-3-642-39241-2\\_37](https://doi.org/10.1007/978-3-642-39241-2_37)
- Al-Haj, H., & Lavie, A. (2012). The impact of Arabic morphological segmentation on broad-coverage English-to-Arabic statistical machine translation. *Machine Translation*, 26(1-2), 3–24. <https://doi.org/10.1007/s10590-011-9101-1>
- Allen, J. F. (2003). Natural language processing. *Encyclopedia of computer science* (pp. 1218–1222). John Wiley; Sons Ltd.
- Almahairi, A., Cho, K., Habash, N., & Courville, A. (2016). First result on arabic neural machine translation. arXiv e-prints, arXiv–1606.
- Al-Onaizan, Y., Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, D., Och, F.-J., Purdy, D., Smith, N. A., & Yarowsky, D. (1999). Statistical machine translation. Final Report, JHU Summer Workshop, 30.
- Alotaiby, F. A., Alkharashi, I. A., & Foda, S. G. (2009). Processing large Arabic text corpora: Preliminary analysis and results. in *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, 78–82.
- Alrabiah, M., Al-Salman, A., & Atwell, E. S. (2013). The design and construction of the 50 million words KSUCCA.
- Alsari, N. A. M., Alshair, A. M., Almalik, S. A., & Alsa'ad, S. S. (2020). A survey on the awareness, accessibility and funding for augmentative and alternative communication services and devices in Saudi Arabia. *Disability and Rehabilitation: Assistive Technology*, 0(0), 1–7. <https://doi.org/10.1080/17483107.2020.1736651>
- Al-Sughaiyer, I. A., & Al-Kharashi, I. A. (2004). Arabic morphological analysis techniques: A comprehensive survey. *Journal of the American Society for Information Science and Technology*, 55(3), 189–213. <http://onlinelibrary.wiley.com/doi/10.1002/asi.10368/full>

- American Speech-Language-Hearing Association. (1993). Definitions of communication disorders and variations [relevant paper].  
<https://doi.org/10.1044/policy.RP1993-00208>
- Archer, L. A. (1977). Blissymbolics-a nonverbal communication system. *Journal of Speech and Hearing Disorders*, 42(4), 568–579.  
<https://doi.org/10.1044/jshd.4204.568>
- Attia, M. A. (2007). Arabic tokenization system. *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, 65–72.
- Badr, I., Zbib, R., & Glass, J. (2008). Segmentation for English-to-Arabic Statistical Machine Translation. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, 153–156. <http://dl.acm.org/citation.cfm?id=1557732>
- Baevski, A., Edunov, S., Liu, Y., Zettlemoyer, L., & Auli, M. (2019). Cloze-driven pretraining of self-attention networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5360–5369. <https://doi.org/10.18653/v1/D19-1539>
- Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.
- Baker, B., Hill, K., & Devylder, R. (2000). Core vocabulary is the same across environments. *California state university at northridge conference*, 3–8.
- Balandin, S., & Iacono, T. (1999). Crews, wusses, and whoppas: Core and fringe vocabularies of Australian meal-break conversations in the workplace. *Augmentative and Alternative Communication*, 15(2), 95–109.
- Ball, L., Marvin, C., Beukelman, D., Lasker, J., & Rupp, D. (1999). Generic talk use by preschool children. *Augmentative and Alternative Communication*, 15(3), 145–155. <https://doi.org/10.1080/07434619912331278685>
- Banajee, M., Dicarlo, C., & Stricklin, S. B. (2003). Core vocabulary determination for toddlers. *Augmentative and Alternative Communication*, 19(2), 67–73.
- Beard, A. (2018). Speech, language and communication: A public health issue across the lifecourse. *Paediatrics and Child Health*, 28(3), 126–131.  
<https://doi.org/https://doi.org/10.1016/j.paed.2017.12.004>
- Belinkov, Y., & Glass, J. (2015). Arabic diacritization with recurrent neural networks. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2281–2285.
- Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *The journal of machine learning research*, 3, 1137–1155.
- Beukelman, D., & Gutmann, M. (1999). Generic message list for AAC users with ALS. Retrieved March 1, 2021, from  
[https://cehs.unl.edu/documents/secd/aac/vocablists/ALS\\_Message\\_List1.pdf](https://cehs.unl.edu/documents/secd/aac/vocablists/ALS_Message_List1.pdf)

- Beukelman, D., & Mirenda, P. (2013). *Augmentative and alternative communication: Supporting children and adults with complex communication needs* (4th ed.). Baltimore, MD: Paul H. Brookes (4th). Paul H. Brookes Publishing Company.
- Bilmes, J., & Kirchhoff, K. (2003). Factored language models and generalized parallel backoff. *Companion Volume of the Proceedings of HLT-NAACL 2003-Short Papers*, 4–6.
- Binger, C., & Light, J. (2006). Demographics of Preschoolers Who Require AAC. *Language, Speech, and Hearing Services in Schools*, 37(3), 200–208.  
[https://doi.org/10.1044/0161-1461\(2006/022\)](https://doi.org/10.1044/0161-1461(2006/022))
- Black, R., Reddington, J., Reiter, E., Tintarev, N., & Waller, A. (2010). Using NLG and sensors to support personal narrative for children with complex communication needs. *Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies*, 1–9.  
<http://dl.acm.org/citation.cfm?id=1867751>
- Black, R., Waller, A., Reiter, E., & Tintarev, N. (2012). Automatic Utterance Generation for Personal Narrative—System Development and Feasibility Experiences. *Communication Matters National Symposium*. Leicester, UK, 23–25.
- Black, R., Waller, A., Reiter, E., & Turner, R. (2008). Tell me about your day: Creating novel access to personal narrative. *Communication Matters Symposium 2008*.
- Black, R., Waller, A., Tintarev, N., Reiter, E., & Reddington, J. (2011). A mobile phone based personal narrative system. *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility - ASSETS '11*, 171. <https://doi.org/10.1145/2049536.2049568>
- Black, R., Waller, A., Turner, R., & Reiter, E. (2012). Supporting personal narrative for children with complex communication needs. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 19(2), 15.  
<http://dl.acm.org/citation.cfm?id=2240163>
- Bliss, C. K. (1949). *Semantography: A non-alphabetical symbol writing, readable in all languages; a practical tool for general international communication, especially in science, industry, commerce, traffic, etc., and for semantical education, based on the principles of ideographic writing and chemical symbolism*. Institute for Semantography.
- Bloomberg, K., Karlan, G. R., & Lloyd, L. L. (1990). The comparative translucency of initial lexical items represented in five graphic symbol systems and sets. *Journal of Speech, Language, and Hearing Research*, 33(4), 717–725.  
<https://doi.org/10.1044/jshr.3304.717>
- Bond, F., & Foster, R. (2013). Linking and extending an open multilingual Wordnet. *Proceedings of the 51st Annual Meeting of the Association for Computational*

- Linguistics (Volume 1: Long Papers), 1352–1362.  
<https://aclanthology.org/P13-1133>
- Bondy, A. S., & Frost, L. A. (1994). The Picture Exchange Communication System. *Focus on Autistic Behavior*, 9(3), 1–19.  
<https://doi.org/10.1177/108835769400900301>
- Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A., & Oflazer, K. (2018). The MADAR Arabic dialect corpus and lexicon. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.  
<https://www.aclweb.org/anthology/L18-1535>
- Brants, T., Popat, A. C., Xu, P., Och, F. J., & Dean, J. (2007). Large language models in machine translation. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 858–867.  
<https://www.aclweb.org/anthology/D07-1090>
- Briechele, K., & Hanebeck, U. D. (2001). Template matching using fast normalized cross correlation. In D. P. Casasent & T.-H. Chao (Eds.), *Optical pattern recognition xii* (pp. 95–102). SPIE. <https://doi.org/10.1117/12.421129>
- Britz, D., Goldie, A., Luong, M.-T., & Le, Q. (2017). Massive exploration of neural machine translation architectures. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1442–1451.
- Brown, K., & Miller, J. (2013a). *Concept*. The cambridge dictionary of linguistics. Cambridge University Press. <https://doi.org/10.1017/CBO9781139049412>
- Brown, K., & Miller, J. (2013b). *Lemmatization*. The cambridge dictionary of linguistics. Cambridge University Press.  
<https://doi.org/10.1017/CBO9781139049412>
- Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Mercer, R. L., & Roossin, P. (1988). A statistical approach to language translation. *Coling Budapest 1988 Volume 1: International Conference on Computational Linguistics*.
- Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1991). Word-Sense Disambiguation Using Statistical Methods. *29th Annual Meeting of the Association for Computational Linguistics*, 264–270.  
<https://doi.org/10.3115/981344.981378>
- Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2), 263–311.  
<https://www.aclweb.org/anthology/J93-2003>
- Brown, P. F., Della Pietra, V. J., Desouza, P. V., Lai, J. C., & Mercer, R. L. (1992). Class-based n-gram models of natural language. *Computational linguistics*, 18(4), 467–480.



- Buckwalter, T. (2002). Buckwalter arabic morphological analyzer version 1.0. linguistic data consortium. University of Pennsylvania, LDC Catalog No.: LDC2002L49.
- Buckwalter, T. (2004). Issues in Arabic orthography and morphology analysis. *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, 31–34. <http://dl.acm.org/citation.cfm?id=1621813>
- Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3), 510–526. <https://doi.org/10.3758/BF03193020>
- Cannon, B., & Edmond, G. (2009). A few good words using core vocabulary to support nonverbal students. *The ASHA Leader*, 14(5), 20–23. <https://doi.org/10.1044/leader.ftr4.14052009.20>
- Chang, S. K., Costagliola, G., Orefice, S., Polese, G., & Baker, B. R. (1992). A methodology for iconic language design with application to augmentative communication. *Proceedings IEEE Workshop on Visual Languages*, 110–116. <https://doi.org/10.1109/WVL.1992.275776>
- Chang, S.-K., Orefice, S., Polese, G., & Baker, B. R. (1993). Deriving the meaning of iconic sentences for augmentative communication. *Visual Languages, 1993.*, *Proceedings 1993 IEEE Symposium on*, 267–274. [http://ieeexplore.ieee.org/xpls/abs\\_all.jsp?arnumber=269608](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=269608)
- Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., & Robinson, T. (2014). One billion word benchmark for measuring progress in statistical language modeling. *Fifteenth Annual Conference of the International Speech Communication Association*.
- Chen, S. F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4), 359–394.
- Chiu, J. P., & Nichols, E. (2016). Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4, 357–370. [https://doi.org/10.1162/tacl\\_a\\_00104](https://doi.org/10.1162/tacl_a_00104)
- Cho, K., van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder–decoder approaches. *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 103–111. <https://doi.org/10.3115/v1/W14-4012>
- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734. <https://doi.org/10.3115/v1/D14-1179>
- Church, K. W., & Gale, W. A. (1991). Concordances for parallel text. *Proceedings of the seventh annual conference of the UW centre for the new OED and text research*, 40–62.

- Church, K. W., & Hanks, P. (1989). Word Association Norms, Mutual Information, and Lexicography. 27th Annual Meeting of the Association for Computational Linguistics, 76–83. <https://doi.org/10.3115/981623.981633>
- Cleverdon, C. (1967). The cranfield tests on index language devices. Readings in information retrieval (pp. 47–59). Morgan Kaufmann Publishers Inc.
- Communication Matters. (n.d.). Glossary. Retrieved July 16, 2021, from <https://www.communicationmatters.org.uk/what-is-aac/glossary/>
- Cooper, H., Holt, B., & Bowden, R. (2011). Sign Language Recognition. In T. B. Moeslund, A. Hilton, V. Krüger, & L. Sigal (Eds.), *Visual Analysis of Humans* (pp. 539–562). Springer London. [https://doi.org/10.1007/978-0-85729-997-0\\_27](https://doi.org/10.1007/978-0-85729-997-0_27)
- Creer, S., Enderby, P., Judge, S., & John, A. (2016). Prevalence of people who could benefit from augmentative and alternative communication (AAC) in the UK: Determining the need. *International Journal of Language & Communication Disorders*, 51(6), 639–653. <https://doi.org/10.1111/1460-6984.12235>
- Cuadros, M., & Rigau, G. (2006). Quality assessment of large scale knowledge resources. *Proceedings of the 2006 conference on empirical methods in natural language processing*, 534–541. <https://www.aclweb.org/anthology/W06-1663>
- Dada, S., Huguet, A., & Bornman, J. (2013). The Iconicity of Picture Communication Symbols for Children with English Additional Language and Mild Intellectual Disability. *Augmentative and Alternative Communication*, 29(4), 360–373. <https://doi.org/10.3109/07434618.2013.849753>
- Dalal, N., & Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 1, 886–893. <https://doi.org/10.1109/CVPR.2005.177>
- Dalal, N., Triggs, B., & Schmid, C. (2006). Human Detection Using Oriented Histograms of Flow and Appearance. In A. Leonardis, H. Bischof, & A. Pinz (Eds.), *Computer Vision – ECCV 2006* (pp. 428–441). Springer. [https://doi.org/10.1007/11744047\\_33](https://doi.org/10.1007/11744047_33)
- Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers*, 28(1), 125–127. <https://doi.org/10.3758/BF03203646>
- Demasco, P. W., & McCoy, K. F. (1992). Generating text from compressed input: An intelligent interface for people with severe motor impairments. *Communications of the ACM*, 35(5), 68–78.
- Dempster, M., Alm, N., & Reiter, E. (2010). Automatic generation of conversational utterances and narrative for Augmentative and Alternative Communication: A prototype system. *Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies*, 10–18. <http://dl.acm.org/citation.cfm?id=1867752>

- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- Diab, M., Hacıoglu, K., & Jurafsky, D. (2004). Automatic tagging of Arabic text: From raw text to base phrase chunks. *Proceedings of HLT-NAACL 2004: Short Papers*, 149–152. <http://dl.acm.org/citation.cfm?id=1614022>
- Diab, M., & Resnik, P. (2002). An Unsupervised Method for Word Sense Tagging using Parallel Corpora. *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, 255–262. <https://doi.org/10.3115/1073083.1073126>
- Ding, C., Halabi, N., Al-Zaben, L., Li, Y., Draffan, E. A., & Wald, M. (2015). A web based multi-linguists symbol-to-text AAC application. *Proceedings of the 12th Web for All Conference*, 24. <http://dl.acm.org/citation.cfm?id=2746674>
- Draffan, E. A., Wald, M., Halabi, N., Sabia, O., Zaghoulani, W., Kadous, A., Idris, A., Zeinoun, N., & Banes, D. (2015). Generating acceptable Arabic Core Vocabularies and Symbols for AAC users. *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, 91–96. <http://eprints.soton.ac.uk/384316/>
- Draffan, E., Wald, M., Halabi, N., Kadous, A., Idris, A., Zeinoun, N., Banes, D., & Lawand, D. (2015). A voting system for aac symbol acceptance. *Proceedings of the 17th International ACM SIGACCESS Conference on Computers Accessibility*, 371–372. <https://doi.org/10.1145/2700648.2811374>
- Dye, R., Alm, N., Arnott, J. L., Harper, G., & Morrison, A. I. (1998). A script-based AAC system for transactional interaction. *Natural Language Engineering*, 4(1), 57–71.
- Dye, R., Alm, N., Arnott, J. L., Murray, I. R., & Harper, G. (1998). ScripTalker - An AAC System Incorporating Scripts. *Proceedings of the TIDE Congress (Technology for Inclusive Design and Equality)*.
- Edmonds, P., & Agirre, E. (2008). Word sense disambiguation. *Scholarpedia*, 3(7), 4358. <https://doi.org/10.4249/scholarpedia.4358>
- Edmonds, P. (2005). Lexical disambiguation. *Encyclopedia of Language and Linguistics*. Ed. by Keith Brown. 2nd edition. Oxford: Elsevier Science, 607–623.
- Edran, A. F. (2002). Picture communication system: A method to support language acquisition to students with moderate to severe disabilities. California State University, Long Beach.
- Eisele, A., & Chen, Y. (2010). MultiUN: A multilingual corpus from united nation documents. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. [http://www.lrec-conf.org/proceedings/lrec2010/pdf/686\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2010/pdf/686_Paper.pdf)

- El Kholi, A., & Habash, N. (2010). Techniques for Arabic morphological detokenization and orthographic denormalization. *LREC 2010 Workshop on Language Resources and Human Language Technology for Semitic Languages*, 45–51.
- Elhadad, M., & Robin, J. (1996). An overview of SURGE: A reusable comprehensive syntactic realization component. *Eighth international natural language generation workshop (posters and demonstrations)*.  
<https://www.aclweb.org/anthology/W96-0501>
- Elkateb, S., Black, W., Rodríguez, H., Alkhalifa, M., Vossen, P., Pease, A., & Fellbaum, C. (2006). Building a wordnet for arabic. *Proceedings of The fifth international conference on Language Resources and Evaluation (LREC 2006)*.
- Ethnologue. (2021, February 21). The most spoken languages worldwide in 2021 (by speakers in millions) [Graph]. Statista. Retrieved July 16, 2021, from <https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide>
- Fallon, K. A., Light, J. C., & Paige, T. K. (2001). Enhancing vocabulary selection for preschoolers who require augmentative and alternative communication (AAC). *AMERICAN JOURNAL OF SPEECH-LANGUAGE PATHOLOGY*, 10(1), 81–94. [https://doi.org/10.1044/1058-0360\(2001/010\)1](https://doi.org/10.1044/1058-0360(2001/010)1)
- Farghaly, A., & Shaalan, K. (2009). Arabic natural language processing. *ACM Transactions on Asian Language Information Processing*, 8(4), 1–22.  
<https://doi.org/10.1145/1644879.1644881>
- Fellbaum, C. (2010). WordNet. In R. Poli, M. Healy, & A. Kameas (Eds.), *Theory and Applications of Ontology: Computer Applications* (pp. 231–243). Springer Netherlands.  
[http://www.springerlink.com/index/10.1007/978-90-481-8847-5\\_10](http://www.springerlink.com/index/10.1007/978-90-481-8847-5_10)
- Fellbaum, C., & Vossen, P. (2012). Challenges for a multilingual wordnet. *Language Resources and Evaluation*, 46(2), 313–326.  
<https://doi.org/10.1007/s10579-012-9186-z>
- Fenson, L., Dale, P. S., Reznick, J. S., Thal, D., Bates, E., Hartung, J. P., Pethick, S., & Reilly, J. S. (1993). *MacArthur communicative development inventories: User's guide and technical manual*. San Diego: Singular.
- Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6), 381–395.  
<https://doi.org/10.1145/358669.358692>
- Forchini, P. (2012). *Movie language revisited. Evidence from multi-dimensional analysis and corpora*. Peter Lang.
- Francis, W. N., & Kucera, H. (1979). *Brown corpus manual*. Letters to the Editor, 5(2), 7.
- Gale, W. A., Church, K. W., & Yarowsky, D. (1992). A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26(5/6), 415–439. <http://www.jstor.org/stable/30204634>

- Gao, Q., & Vogel, S. (2008). Parallel implementations of word alignment tool. *Software engineering, testing, and quality assurance for natural language processing*, 49–57.
- Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., & Smith, N. A. (2011). Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 42–47.  
<https://www.aclweb.org/anthology/P11-2008>
- Goldbart, J., & Marshall, J. (2004). “pushes and pulls” on the parents of children who use aac. *Augmentative and Alternative Communication*, 20(4), 194–208.  
<https://doi.org/10.1080/07434610400010960>
- Goldberg, A. B., Rosin, J., Zhu, X., & Dyer, C. R. (2009). Toward text-to-picture synthesis. *NIPS 2009 Mini-Symposia on Assistive Machine Learning for People with Disabilities*. <http://pages.cs.wisc.edu/~jerryzhu/pub/ttpNIPS09.pdf>
- Goldberg, A. B., Zhu, X., Dyer, C. R., Eldawy, M., & Heng, L. (2008). Easy as ABC?: Facilitating pictorial communication via semantically enhanced layout. *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, 119–126.
- Goshtasby, A., Gage, S. H., & Bartholic, J. F. (1984). A two-stage cross correlation approach to template matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(3), 374–378.  
<https://doi.org/10.1109/TPAMI.1984.4767532>
- Grove, N., & Walker, M. (1990). The Makaton Vocabulary: Using manual signs and graphic symbols to develop interpersonal communication. *Augmentative and Alternative Communication*, 6(1), 15–28.  
<https://doi.org/10.1080/07434619012331275284>
- Gu, J., Cho, K., & Li, V. O. (2017). Trainable greedy decoding for neural machine translation. *2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*, 1968–1978.
- Habash, N., & Hu, J. (2009). Improving Arabic-Chinese statistical machine translation using English as pivot language. *Proceedings of the Fourth Workshop on Statistical Machine Translation*, 173–181.  
<http://dl.acm.org/citation.cfm?id=1626467>
- Habash, N., Olive, J., Christianson, C., & McCary, J. (2011). Machine translation from text. In J. Olive, C. Christianson, & J. McCary (Eds.), *Handbook of natural language processing and machine translation: Darpa global autonomous language exploitation* (pp. 133–397). Springer New York.  
[https://doi.org/10.1007/978-1-4419-7713-7\\_2](https://doi.org/10.1007/978-1-4419-7713-7_2)
- Habash, N., & Rambow, O. (2009). MADA TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging,

- stemming and lemmatization. Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR).
- Habash, N., & Roth, R. (2009). Catib: The columbia arabic treebank. Proceedings of the ACL-IJCNLP 2009 conference short papers, 221–224.
- Habash, N., Roth, R., Rambow, O., Eskander, R., & Tomeh, N. (2013). Morphological analysis and disambiguation for dialectal arabic. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 426–432.
- Habash, N., & Sadat, F. (2006). Arabic preprocessing schemes for statistical machine translation. Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, 49–52.
- Habash, N. Y. (2010). Introduction to arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1), 1–187.  
<https://doi.org/10.2200/S00277ED1V01Y201008HLT010>
- Hajic, J., Smrz, O., Zemánek, P., Šnidauf, J., & Beška, E. (2004). Prague Arabic dependency treebank: Development in data and tools. Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools, 110–117. [http://ufal.mff.cuni.cz/project/padt/PADT\\_1.0/docs/papers/2004-nemlar-padt.pdf](http://ufal.mff.cuni.cz/project/padt/PADT_1.0/docs/papers/2004-nemlar-padt.pdf)
- Heintz, I. (2014). Language Modeling. In I. Zitouni (Ed.), *Natural Language Processing of Semitic Languages* (pp. 161–196). Springer Berlin Heidelberg.  
[http://dx.doi.org/10.1007/978-3-642-45358-8\\_5](http://dx.doi.org/10.1007/978-3-642-45358-8_5)
- Higginbotham, D., Shane, H., Russell, S., & Caves, K. (2007). Access to AAC: Present, past, and future. *AAC: Augmentative and Alternative Communication*, 23(3), 243–257. <https://doi.org/10.1080/07434610701571058>
- Huang, J., Kumar, S., Mitra, M., Zhu, W.-J., & Zabih, R. (1997). Image indexing using color correlograms. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 762–768.  
<https://doi.org/10.1109/CVPR.1997.609412>
- Huer, M. B. (2000). Examining perceptions of graphic symbols across cultures: Preliminary study of the impact of culture/ethnicity. *Augmentative and Alternative Communication*, 16(3), 180–185.  
<https://doi.org/10.1080/07434610012331279034>
- Hunnicutt, S. (1984). Bliss symbol-to-speech conversion: "Bliss-talk". 25, 058–077.
- Ide, N., Erjavec, T., & Tufis, D. (2002). Sense discrimination with parallel corpora. Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions, 61–66.  
<https://doi.org/10.3115/1118675.1118683>
- Ide, N., & Wilks, Y. (2006). Making Sense About Sense. In E. Agirre & P. Edmonds (Eds.), *Word Sense Disambiguation: Algorithms and Applications* (pp. 47–73). Springer Netherlands. [https://doi.org/10.1007/978-1-4020-4809-8\\_3](https://doi.org/10.1007/978-1-4020-4809-8_3)



- Irvine, A., & Callison-Burch, C. (2013). Combining bilingual and comparable corpora for low resource machine translation. *Proceedings of the Eighth Workshop on Statistical Machine Translation*, 262–270.  
<https://www.aclweb.org/anthology/W13-2233>
- Jones, F. W., Long, K., & Finlay, W. M. L. (2007). Symbols can improve the reading comprehension of adults with learning disabilities. *Journal of Intellectual Disability Research*, 51(7), 545–550.  
<https://doi.org/10.1111/j.1365-2788.2006.00926.x>
- Kahlon, N. K., & Singh, W. (2021). Machine translation from text to sign language: A systematic review. *Universal Access in the Information Society*.  
<https://doi.org/10.1007/s10209-021-00823-1>
- Kane, S. K., Morris, M. R., Paradiso, A., & Campbell, J. (2017). "at times avuncular and cantankerous, with the reflexes of a mongoose": Understanding self-expression through augmentative and alternative communication devices. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1166–1179.  
<https://doi.org/10.1145/2998181.2998284>
- Karberis, G., & Kouroupetroglou, G. (2002). Transforming spontaneous telegraphic language to Well-Formed greek sentences for alternative and augmentative communication. *Methods and Applications of Artificial Intelligence* (pp. 155–166). Springer.  
[http://link.springer.com/chapter/10.1007/3-540-46014-4\\_15](http://link.springer.com/chapter/10.1007/3-540-46014-4_15)
- Khalifa, S., Habash, N., Abdulrahim, D., & Hassan, S. (2016). A large scale corpus of gulf arabic. *10th International Conference on Language Resources and Evaluation, LREC 2016*, 4282–4289.
- Kilgarrieff, A., & Rosenzweig, J. (2000). Framework and Results for English SENSEVAL. *Computers and the Humanities*, 34(1), 15–48.  
<https://doi.org/10.1023/A:1002693207386>
- Kilgarrieff, A. (2006). Word Senses. In E. Agirre & P. Edmonds (Eds.), *Word Sense Disambiguation: Algorithms and Applications* (pp. 29–46). Springer Netherlands. [https://doi.org/10.1007/978-1-4020-4809-8\\_2](https://doi.org/10.1007/978-1-4020-4809-8_2)
- King, J., Spoeneman, T., Stuart, S., & Beukelman, D. (1995). Small talk in adult conversations: Implications for AAC vocabulary selection. *Augmentative and Alternative Communication*, 11(4), 260–264.  
<https://doi.org/10.1080/07434619512331277399>
- Kirchhoff, K., Vergyri, D., Bilmes, J., Duh, K., & Stolcke, A. (2006). Morphology-based language modeling for conversational arabic speech recognition. *Computer Speech & Language*, 20(4), 589–608.  
<https://doi.org/https://doi.org/10.1016/j.csl.2005.10.001>

- Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. (2017). OpenNMT: Open-source toolkit for neural machine translation. *Proceedings of ACL 2017, System Demonstrations*, 67–72. <https://www.aclweb.org/anthology/P17-4012>
- Klepousniotou, E. (2002). The Processing of Lexical Ambiguity: Homonymy and Polysemy in the Mental Lexicon. *Brain and Language*, 81(1), 205–223. <https://doi.org/10.1006/brln.2001.2518>
- Kneser, R., & Ney, H. (1995). Improved backing-off for m-gram language modeling. *1995 International Conference on Acoustics, Speech, and Signal Processing*, 1, 181–184 vol.1. <https://doi.org/10.1109/ICASSP.1995.479394>
- Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *MT Summit X*.
- Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511815829>
- Koehn, P. (2020). *Neural machine translation*. Cambridge University Press. <https://doi.org/10.1017/9781108608480>
- Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, 48–54.
- Kurokawa, D., Goutte, C., & Isabelle, P. (2009). Automatic detection of translated text and its impact on machine translation. *Proceedings of MT-Summit XII*, 81–88.
- Langer, S., & Hickey, M. (1999). Augmentative and Alternative Communication and Natural Language Processing: Current research activities and prospects. *Augmentative and Alternative Communication*, 15(4), 260–268. <http://informahealthcare.com/doi/abs/10.1080/07434619912331278795>
- Langkilde, I., & Knight, K. (1998). Generation that exploits corpus-based statistical knowledge. *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, 704–710. <http://dl.acm.org/citation.cfm?id=980963>
- Laubscher, E., & Light, J. (2020). Core vocabulary lists for young children and considerations for early language development: A narrative review. *Augmentative and Alternative Communication*, 36(1), 43–53.
- Lembersky, G., Ordan, N., & Wintner, S. (2012). Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, 38(4), 799–825. [https://doi.org/10.1162/coli\\_a\\_00111](https://doi.org/10.1162/coli_a_00111)
- Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3, 211–225.
- Lewis, D. (1997). Reuters-21578 text categorization test collection, distribution 1.0. <http://www.research.att.com>. <https://ci.nii.ac.jp/naid/10022174263/en/>
- Lewis, J. (1995). Fast normalized cross-correlation. *Vision Interface*, 120–123.



- Light, J. (1988). Interaction involving individuals using augmentative and alternative communication systems: State of the art and future directions. *Augmentative and Alternative Communication*, 4(2), 66–82.  
<https://doi.org/10.1080/07434618812331274657>
- Light, J., & McNaughton, D. (2012). Supporting the communication, language, and literacy development of children with complex communication needs: State of the science and future research priorities. *Assistive Technology*, 24(1), 34–44.  
<https://doi.org/10.1080/10400435.2011.648717>
- Light, J., & McNaughton, D. (2014). Communicative competence for individuals who require augmentative and alternative communication: A new definition for a new era of communication? [PMID: 30952185]. *Augmentative and Alternative Communication*, 30(1), 1–18. <https://doi.org/10.3109/07434618.2014.885080>
- Lin, S.-C., Tsai, C.-L., Chien, L.-F., Chen, K.-J., & Lee, L.-S. (1997). Chinese language model adaptation based on document classification and multiple domain-specific language models. *Proc. 5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, 1463–1466.
- Lison, P., & Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 923–929.
- Lloyd, L., & Blischak, D. (1992). Aac terminology policy and issues update. *Augmentative and Alternative Communication*, 8(2), 104–109.  
<https://doi.org/10.1080/07434619212331276153>
- Lopez, A. (2008). Statistical machine translation. *ACM Comput. Surv.*, 40(3).  
<https://doi.org/10.1145/1380584.1380586>
- Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3), 319–344.  
<https://doi.org/10.1075/ijcl.22.3.02lov>
- Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2), 91–110.  
<https://doi.org/10.1023/B:VISI.0000029664.99615.94>
- Lundälv, M., & Derbring, S. (2012a). AAC Vocabulary Standardisation and Harmonisation. In K. Miesenberger, A. Karshmer, P. Penaz, & W. Zagler (Eds.), *Computers Helping People with Special Needs* (pp. 303–310). Springer Berlin Heidelberg.  
[http://link.springer.com/chapter/10.1007/978-3-642-31534-3\\_46](http://link.springer.com/chapter/10.1007/978-3-642-31534-3_46)
- Lundälv, M., & Derbring, S. (2012b). Towards General Cross-Platform CCF Based Multi-modal Language Support. In K. Miesenberger, A. Karshmer, P. Penaz, & W. Zagler (Eds.), *Computers Helping People with Special Needs* (pp. 261–268). Springer Berlin Heidelberg.  
[http://link.springer.com/chapter/10.1007/978-3-642-31534-3\\_40](http://link.springer.com/chapter/10.1007/978-3-642-31534-3_40)

- Lundälv, M., Derbring, S., Mühlenbock, K. H., Brännström, A., Farre, B., & Nordberg, L. (2014). Inclusive AAC: Multi-modal and multilingual language support for all (P. Encarnação, Ed.). *Technology and Disability*, 26(2-3), 93–103. <https://doi.org/10.3233/TAD-140407>
- Lundälv, M., Mühlenbock, K., Farre, B., & Brännström, A. (2006). SYMBERED - a symbol-concept editing tool. *Proceedings of the fifth international conference on language resources and evaluation (LREC'06)*.
- Lyons, J. (1995). *Linguistic semantics: An introduction*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511810213>
- Ma, X., & Hovy, E. H. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. *CoRR*, abs/1603.01354. <http://arxiv.org/abs/1603.01354>
- Maamouri, M., Bies, A., Buckwalter, T., Diab, M. T., Habash, N., Rambow, O., & Tabessi, D. (2006). Developing and using a pilot dialectal arabic treebank. *LREC*, 443–448.
- Maamouri, M., Bies, A., Buckwalter, T., & Mekki, W. (2004). The penn arabic treebank: Building a large-scale annotated arabic corpus. *NEMLAR conference on Arabic language resources and tools*, 27, 466–467.
- Mairesse, F., & Young, S. (2014). Stochastic Language Generation in Dialogue using Factored Language Models. *Computational Linguistics*, 40(4), 763–799. [https://doi.org/10.1162/COLI\\_a\\_00199](https://doi.org/10.1162/COLI_a_00199)
- Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.
- Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In A. F. Gelbukh (Ed.), *Computational linguistics and intelligent text processing* (pp. 171–189). Springer Berlin Heidelberg.
- Manurung, R., O'Mara, D., Pain, H., Ritchie, G., & Waller, A. (2006). Building a Lexical Database for an Interactive Joke-Generator. *LREC*, 1738–1741.
- Markov, A. A. (1954). The theory of algorithms. *Trudy Matematicheskogo Instituta Imeni VA Steklova*, 42, 3–375.
- Marvin, C., Beukelman, D., & Bilyeu, D. (1994). Vocabulary-use patterns in preschool children: Effects of context and time sampling. *Augmentative and Alternative Communication*, 10(4), 224–236.
- McClure, M., & Rush, E. (2007). Selecting symbol sets: Implications for AAC users, clinicians, and researchers. *annual meeting of american speech-language-hearing association, boston*.
- McCoy, K., Demasco, P., Jones, M., Pennington, C., & Rowe, C. (1990). Applying natural language processing techniques to augmentative communication systems. *Proceedings of the 13th conference on Computational linguistics-Volume 3*, 413–415. <http://dl.acm.org/citation.cfm?id=991235>

- McCoy, K. F. (1997). Simple NLP Techniques for Expanding Telegraphic Sentences. *Natural Language Processing for Communication Aids*.  
<https://www.aclweb.org/anthology/W97-0503>
- McCoy, K. F., & Hershberger, D. (1999). The role of evaluation in bringing NLP to AAC: A case to consider. *AAC: New Directions in Research and Practice*, 105–122.
- McCoy, K. F., McKnitt, W. M., Peischl, D., Pennington, C., Vanderheyden, P., & Demasco, P. W. (1994). AAC-user therapist interactions: Preliminary linguistic observations and implications for Compansion. *Proceedings of the RESNA*, 94, 129–131.
- McCoy, K. F., Pennington, C. A., & Badman, A. L. (1998). Compansion: From research prototype to practical integration. *Natural Language Engineering*, 4(1), 73–95.
- Mehl, M. R., Vazire, S., Ramirez-Esparza, N., Slatcher, R. B., & Pennebaker, J. W. (2007). Are women really more talkative than men? *SCIENCE*, 317(5834), 82.  
<https://doi.org/10.1126/science.1139940>
- Menai, M. E. B. (2014). Word sense disambiguation using evolutionary algorithms – application to arabic language. *Computers in Human Behavior*, 41, 92–103.  
<https://doi.org/10.1016/j.chb.2014.06.021>
- Mihalcea, R. (2007). Using Wikipedia for automatic word sense disambiguation. *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, 196–203. <https://www.aclweb.org/anthology/N07-1025>
- Mihalcea, R., & Leong, C. W. (2008). Toward communicating simple sentences using pictorial representations. *Machine Translation*, 22(3), 153–173.  
<https://doi.org/10.1007/s10590-009-9050-0>
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 3111–3119.  
<http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality>
- Mikolov, T., Deoras, A., Kombrink, S., Burget, L., & Černocký, J. (2011). Empirical evaluation and combination of advanced language modeling techniques. *Twelfth annual conference of the international speech communication association*.
- Miller, G. A., Leacock, C., Teng, R., & Bunker, R. T. (1993). A semantic concordance. *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*. <https://www.aclweb.org/anthology/H93-1061>
- Mirenda, P., & Locke, P. A. (1989). A comparison of symbol transparency in nonspeaking persons with intellectual disabilities. *Journal of Speech and Hearing Disorders*, 54(2), 131–140.

- Mitchell, M., Han, X., Dodge, J., Mensch, A., Goyal, A., Berg, A., Yamaguchi, K., Berg, T., Stratos, K., & Daumé, H., III. (2012). Midge: Generating Image Descriptions from Computer Vision Detections. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 747–756. <http://dl.acm.org/citation.cfm?id=2380816.2380907>
- Mitchell, M., & Sproat, R. (2012). Discourse-based modeling for aac. *Proceedings of the Third Workshop on Speech and Language Processing for Assistive Technologies*, 9–18. <http://dl.acm.org/citation.cfm?id=2392858>
- Mizuko, M. (1987). Transparency and ease of learning of symbols represented by Blissymbols, PCS, and Picsyms. *Augmentative and Alternative Communication*, 3(3), 129–136. <https://doi.org/10.1080/07434618712331274409>
- Moorcroft, A., Scarinci, N., & Meyer, C. (2019). A systematic review of the barriers and facilitators to the provision and use of low-tech and unaided aac systems for people with complex communication needs and their families. *Disability and Rehabilitation: Assistive Technology*, 14(7), 710–731. <https://doi.org/10.1080/17483107.2018.1499135>
- Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2). <https://doi.org/10.1145/1459352.1459355>
- Navigli, R., Litkowski, K. C., & Hargraves, O. (2007). SemEval-2007 Task 07: Coarse-Grained English All-Words Task. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, 30–35. <https://www.aclweb.org/anthology/S07-1006>
- Newell, A., Langer, S., & Hickey, M. (1998). The rôle of natural language processing in alternative and augmentative communication. *Natural Language Engineering*, 4(01), 1–16. <https://doi.org/null>
- Newell, A. F., & Alm, N. (1994). Developing AAC technologies: A personal story and philosophy. *European journal of disorders of communication*, 29(4), 399–411.
- Nippold, M. A., Cramond, P. M., & Hayward-Mayhew, C. (2014). Spoken language production in adults: Examining age-related differences in syntactic complexity. *Clinical Linguistics & Phonetics*, 28(3), 195–207. <https://doi.org/10.3109/02699206.2013.841292>
- Novais, E. M. d., & Paraboni, I. (2012). Portuguese text generation using factored language models. *Journal of the Brazilian Computer Society*, 19(2), 135–146. <https://doi.org/10.1007/s13173-012-0095-1>
- Och, F. J., & Ney, H. (2000). Improved Statistical Alignment Models. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 440–447. <https://doi.org/10.3115/1075218.1075274>
- Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1), 19–51. <https://doi.org/10.1162/089120103321337421>

- Och, F. J., Tillmann, C., & Ney, H. (1999). Improved alignment models for statistical machine translation. *Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, 20–28. <http://mtgroup.ict.ac.cn/~liuqun/protected/papers/Och1999Improved-liuqun.doc>
- Oh, A. H., & Rudnicky, A. I. (2002). Stochastic natural language generation for spoken dialog systems. *Computer Speech & Language*, 16(3-4), 387–407. [https://doi.org/10.1016/S0885-2308\(02\)00012-8](https://doi.org/10.1016/S0885-2308(02)00012-8)
- Palmer, M., Babko-Malaya, O., Bies, A., Diab, M. T., Maamouri, M., Mansouri, A., & Zaghouni, W. (2008). A Pilot Arabic Propbank. *LREC*. [http://repository.dlsi.ua.es/242/1/pdf/880\\_paper.pdf](http://repository.dlsi.ua.es/242/1/pdf/880_paper.pdf)
- Palmer, M., Dang, H. T., & Fellbaum, C. (2007). Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13(2), 137–163. <https://doi.org/10.1017/S135132490500402X>
- Pampoulou, E., & Detheridge, C. (2007). The role of symbols in the mainstream to access literacy. *Journal of Assistive Technologies*, 1(1), 15–21. <https://doi.org/10.1108/17549450200700004>
- Paolieri, D., & Marful, A. (2018). Norms for a Pictographic System: The Aragonese Portal of Augmentative/Alternative Communication (ARASAAC) System. *Frontiers in Psychology*, 9. <https://doi.org/10.3389/fpsyg.2018.02538>
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. <http://dl.acm.org/citation.cfm?id=1073135>
- Pascanu, R., Gulcehre, C., Cho, K., & Bengio, Y. (2014). How to construct deep recurrent neural networks: *Proceedings of the second international conference on learning representations (iclr 2014)*. 2nd International Conference on Learning Representations, ICLR 2014.
- Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, 1310–1318.
- Pasha, A., Al-Badrashiny, M., Diab, M., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O., & Roth, R. M. (2014). MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland. [http://www.lrec-conf.org/proceedings/lrec2014/pdf/593\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2014/pdf/593_Paper.pdf)
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep

- learning library. *Advances in neural information processing systems*, 32, 8026–8037.
- Patel, R., & Khamis-Dakwar, R. (2005). An AAC training program for special education teachers: A case study of Palestinian Arab teachers in Israel. *Augmentative and Alternative communication*, 21(3), 205–217.
- Petrov, S., Das, D., & McDonald, R. (2012). A universal part-of-speech tagset. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2089–2096.  
[http://www.lrec-conf.org/proceedings/lrec2012/pdf/274\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf)
- Pino, A. (2014). Augmentative and alternative communication systems for the motor disabled. *Disability informatics and web accessibility for motor limitations* (pp. 105–152). IGI Global.
- Rambow, O., Bangalore, S., & Walker, M. (2001). Natural Language Generation in Dialog Systems. *Proceedings of the First International Conference on Human Language Technology Research*. <https://www.aclweb.org/anthology/H01-1055>
- Reiter, E., Turner, R., Alm, N., Black, R., Dempster, M., & Waller, A. (2009). Using NLG to help language-impaired users tell stories and participate in social dialogues. *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, 1–8. <https://aclanthology.org/W09-0601>
- Renvall, K., Nickels, L., & Davidson, B. (2013). Functionally relevant items in the treatment of aphasia (part I): Challenges for current practice. *Aphasiology*, 27(6), 636–650.
- Resnik, P., & Yarowsky, D. (1999). Distinguishing systems and distinguishing senses: New evaluation methods for Word Sense Disambiguation. *Natural Language Engineering*, 5(2), 113–133. <https://doi.org/10.1017/S1351324999002211>
- Rice, M. L., Smolik, F., Perpich, D., Thompson, T., Rytting, N., & Blossom, M. (2010). Mean length of utterance levels in 6-month intervals for children 3 to 9 years with and without language impairments. *Journal of Speech, Language, and Hearing Research*, 53(2), 333–349.
- Ritchie, G., Manurung, R., Pain, H., Waller, A., Black, R., & O'Mara, D. (2007). A practical application of computational humour. *Proceedings of the 4th International Joint Conference on Computational Creativity*, 91–98.
- Rivest, R. (1992). Rfc1321: The md5 message-digest algorithm.  
<https://doi.org/10.17487/RFC1321>
- Rodríguez, H., Farwell, D., Farreres, J., Bertran, M., Alkhalifa, M., Martí, M. A., Black, W., Elkateb, S., Kirk, J., Pease, A., Vossen, P., & Fellbaum, C. (2008). Arabic wordnet: Current state and future extensions. *Proceedings of The Fourth Global WordNet Conference*, Szeged, Hungary.
- Rogati, M., McCarley, S., & Yang, Y. (2003). Unsupervised Learning of Arabic Stemming Using a Parallel Corpus. *Proceedings of the 41st Annual Meeting of*



- the Association for Computational Linguistics, 391–398.  
<https://doi.org/10.3115/1075096.1075146>
- Rosenfeld, R. (2000). Two decades of statistical language modeling: Where do we go from here. *Proceedings of the IEEE*, 2000.
- Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The Earth Mover’s Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2), 99–121. <https://doi.org/10.1023/A:1026543900054>
- Ryding, K. C. (2014). *Arabic: A linguistic introduction*. Cambridge University Press.  
<https://doi.org/10.1017/CBO9781139151016>
- Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. In A. Gelbukh (Ed.), *Computational linguistics and intelligent text processing* (pp. 1–15). Springer Berlin Heidelberg.
- Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
- Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11), 2673–2681.  
<https://doi.org/10.1109/78.650093>
- Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1), 97–123. <https://www.aclweb.org/anthology/J98-1004>
- Sekine, S. (1997). The Domain Dependence of Parsing. *Fifth Conference on Applied Natural Language Processing*, 96–102. <https://doi.org/10.3115/974557.974572>
- Sevens, L., Jacobs, G., Vandeghinste, V., Schuurman, I., & Van Eynde, F. (2016). Improving Text-to-Pictograph Translation Through Word Sense Disambiguation. *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, 131–135. <https://doi.org/10.18653/v1/S16-2017>
- Sevens, L., Vandeghinste, V., Schuurman, I., & Van Eynde, F. (2015a). Extending a Dutch Text-to-Pictograph Converter to English and Spanish. *6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 110.  
<http://www.aclweb.org/anthology/W/W15/W15-51.pdf%5C#page=117>
- Sevens, L., Vandeghinste, V., Schuurman, I., & Van Eynde, F. (2015b). Natural Language Generation from Pictographs. *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, 71–75.
- Sigafoos, J. (2010). Introduction to the Special Issue on Augmentative and Alternative Communication. *Journal of Developmental and Physical Disabilities*, 22(2), 101–104. <https://doi.org/10.1007/s10882-010-9197-x>
- Smith, M., & Grove, N. (2003). Asymmetry in input and output for individuals who use augmentative and alternative communication. *Communicative Competence of Individuals Who Use Augmentative and Alternative Communication*, 163–195.

- Snyder, B., & Palmer, M. (2004). The English all-words task. Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, 41–43.  
<https://www.aclweb.org/anthology/W04-0811>
- Stephenson, J., & Linfoot, K. (1996). Pictures as communication symbols for students with severe intellectual disability. *Augmentative and Alternative Communication*, 12(4), 244–256.  
<https://doi.org/10.1080/07434619612331277708>
- Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. Seventh international conference on spoken language processing.
- Stolcke, A., Zheng, J., Wang, W., & Abrash, V. (2011). SRILM at sixteen: Update and outlook. Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop, 5.  
<http://www-speech.sri.com/projects/srilm/papers/asru2011-srilm.pdf>
- Stricker, M. A., & Orengo, M. (1995). Similarity of color images. *Storage and Retrieval for Image and Video Databases III*, 2420, 381–392.  
<https://doi.org/10.1117/12.205308>
- Sutton, A., Gallagher, T., Morford, J., & Shahnaz, N. (2000). Relative clause sentence production using augmentative and alternative communicationsystems. *Applied Psycholinguistics*, 21(4), 473–486. <https://doi.org/10.1017/S0142716400004033>
- Symbol, n.1. (2021). In *Oxford English Dictionary*. Oxford University Press. Retrieved July 16, 2021, from <https://www.oed.com/view/Entry/196197?rskey=SIXvQO&result=1&isAdvanced=false>
- Taji, D., Habash, N., & Zeman, D. (2017). Universal dependencies for arabic. Proceedings of the Third Arabic Natural Language Processing Workshop, 166–176.
- Tintarev, N., Reiter, E., Black, R., Waller, A., & Reddington, J. (2016). Personal storytelling: Using Natural Language Generation for children with complex communication needs, in the wild... *International Journal of Human-Computer Studies*, 92-93, 1–16. <https://doi.org/10.1016/j.ijhcs.2016.04.005>
- Tkalcic, M., & Tasic, J. (2003). Colour spaces: Perceptual, historical and applicational background. *The IEEE Region 8 EUROCON 2003. Computer as a Tool.*, 1, 304–308 vol.1. <https://doi.org/10.1109/EURCON.2003.1248032>
- Todman, J., Elder, L., & Alm, N. (1995). Evaluation of the content of computer-aided conversations. *Augmentative and Alternative Communication*, 11(4), 229–235.
- Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, 252–259.
- Trinh, H., Waller, A., Vertanen, K., Kristensson, P. O., & Hanson, V. L. (2012). iSCAN: A phoneme-based predictive communication aid for nonspeaking



- individuals. Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility, 57–64.
- Tufis, D., Ion, R., & Ide, N. (2005). Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets. arXiv:cs/0503024. <http://arxiv.org/abs/cs/0503024>
- Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37, 141–188. <https://doi.org/10.1613/jair.2934>
- Tuset, P., Barberán, P., Janer, L., Buscà, E., Delgado, S., & Vilà, N. (2010). Messenger visual: A pictogram-based IM service to improve communications among disabled people. Proceedings of the 6th Nordic Conference on Human-Computer Interaction Extending Boundaries - NordiCHI '10, 797. <https://doi.org/10.1145/1868914.1869032>
- United Nations. (2006). Convention on the rights of persons with disabilities. New York: United Nations.
- Vandeghinste, V., & Schuurman, I. (2014). Linking pictographs to synsets: Sclera2Cornetto. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 9, 3404–3410.
- Vandeghinste, V., Sevens, I. S. L., & Eynde, F. V. (2017). Translating text into pictographs. *Natural Language Engineering*, 23(2), 217–244. <https://doi.org/10.1017/S135132491500039X>
- Vandeghinste, V., Sevens, L., & Schuurman, I. (2018). Pictograph Translation Technologies for People with Limited Literacy. Poster session, 4.
- Vanderheyden, P. B., Demasco, P. W., McCoy, K. F., & Pennington, C. A. (1996). A preliminary study into schema-based access and organization of re-usable text in aac. Proceedings of the RESNA Conference, 59–61.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems, 6000–6010.
- Vertanen, K., & Kristensson, P. O. (2011). The imagination of crowds: Conversational AAC language modeling using crowdsourcing and large data sources. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 700–711. <http://dl.acm.org/citation.cfm?id=2145514>
- Vicente, A., & Falkum, I. L. (2017). Polysemy. <https://doi.org/10.1093/acrefore/9780199384655.013.325>
- Viglas, C., & Kouroupetroglou, G. (2002). An open machine translation system for augmentative and alternative communication. *Computers Helping People with Special Needs* (pp. 699–706). Springer. [http://link.springer.com/chapter/10.1007/3-540-45491-8\\_135](http://link.springer.com/chapter/10.1007/3-540-45491-8_135)

- Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Vossen, P., Maks, I., Segers, R., & Vliet, H. V. D. (2008). Integrating lexical units, synsets and ontology in the Cornetto database. In *LREC'08*.
- Waller, A., Balandin, S. A., O'Mara, D. A., & Judson, A. D. (2005). Training aac users in user-centred design. *Proceedings of the 2005 International Conference on Accessible Design in the Digital World*, 2.
- Waller, A., & Black, R. (2012). Personal storytelling for children who use augmentative and alternative communication. *Using Storytelling to Support Children and Adults with Special Needs*, 111–119.
- Waller, A., Francis, J., Tait, L., Booth, L., & Hood, H. (1999). The WriteTalk project: Story-based interactive communication. *Assistive Technology on the Threshold of the New Millennium*, 6, 180.
- Waller, A., & Jack, K. (2002). A predictive Blissymbolic to English translation system. *Proceedings of the fifth international ACM conference on Assistive technologies*, 186–191. <http://dl.acm.org/citation.cfm?id=638283>
- Waller, A., Menzies, R., Herron, D., Prior, S., Black, R., & Kroll, T. (2013). Chronicles: Supporting conversational narrative in alternative and augmentative communication. *IFIP Conference on Human-Computer Interaction*, 364–371.
- Waller, A., & Newell, A. F. (1997). Towards a narrative-based augmentative communication system. *International Journal of Language & Communication Disorders*, 32(S3), 289–306.
- Watson, J. C. (2002). *The phonology and morphology of arabic*. Oxford University Press on Demand.
- Wiegand, K. (2013). Semantic disambiguation of non-syntactic and continuous motion text entry for AAC. *ACM SIGACCESS Accessibility and Computing*, (105), 38–43. <http://dl.acm.org/citation.cfm?id=2444808>
- Wiegand, K., & Patel, R. (2012a). Non-syntactic word prediction for AAC. *Proceedings of the Third Workshop on Speech and Language Processing for Assistive Technologies*, 28–36. <http://dl.acm.org/citation.cfm?id=2392860>
- Wiegand, K., & Patel, R. (2012b). SymbolPath: A continuous motion overlay module for icon-based assistive communication. *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*, 209–210. <http://dl.acm.org/citation.cfm?id=2384957>
- Wilks, Y., & Stevenson, M. (1998). The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation. *Natural Language Engineering*, 4(2), 135–143. <https://doi.org/10.1017/S1351324998001946>
- Witkowski, D., & Baker, B. (2012). Addressing the content vocabulary with core: Theory and practice for nonliterate or emerging literate students. *Perspectives on Augmentative and Alternative Communication*, 21(3), 74–81.

- Wu, J., Cui, Z., Sheng, V. S., Zhao, P., Su, D., & Gong, S. (2013). A Comparative Study of SIFT and its Variants. *Measurement Science Review*, 13(3), 122–131. <https://doi.org/10.2478/msr-2013-0021>
- Yang, Y., Teo, C. L., Daumé, H., III, & Aloimonos, Y. (2011). Corpus-guided Sentence Generation of Natural Images. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 444–454. <http://dl.acm.org/citation.cfm?id=2145432.2145484>
- Yarowsky, D., Ngai, G., & Wicentowski, R. (2001). Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. *Proceedings of the First International Conference on Human Language Technology Research*. <https://www.aclweb.org/anthology/H01-1035>
- Zeroual, I., Lakhouaja, A., & Belahbib, R. (2017). Towards a standard part of speech tagset for the arabic language. *Journal of King Saud University - Computer and Information Sciences*, 29(2), 171–178. <https://doi.org/https://doi.org/10.1016/j.jksuci.2017.01.006>
- Zheng, L., Yang, Y., & Tian, Q. (2018). SIFT Meets CNN: A Decade Survey of Instance Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5), 1224–1244. <https://doi.org/10.1109/TPAMI.2017.2709749>
- Zhong, Z., & Ng, S. (2010). It Makes Sense: A Wide-Coverage Word Sense Disambiguation System for Free Text. *Proceedings of the ACL 2010 System Demonstrations*, 78–83. <https://www.aclweb.org/anthology/P10-4014>
- Zhu, X., Goldberg, A. B., Eldawy, M., Dyer, C. R., & Strock, B. (2007). A text-to-picture synthesis system for augmenting communication. *AAAI*, 7, 1590–1595. <http://www.aaai.org/Papers/AAAI/2007/AAAI07-252.pdf>