Multilingual Symbolic Support for Low Levels of Literacy on the Web

EA Draffan
ead@ecs.soton.ac.uk
ECS, University of Southampton
Southampton, UK

Chaohai Ding
c.ding@soton.ac.uk
ECS, University of Southampton
Southampton, UK

Mike Wald
mw@ecs.soton.ac.uk
ECS, University of Southampton
Southampton, UK

Russell Newman
rn@russellnewman.co.uk
ECS, University of Southampton
Southampton, UK

ABSTRACT

Although literacy rates around the world have increased and there is an expectation that individuals who access web pages will be able to read their content, this is not always the case. The barriers that may be faced can be linked to the way the system is designed and content is written. There may be complex language or a layout that is dense, cluttered and lacks clear markers regarding the key points being made.

Many organizations have provided guidance for web developers and authors offering suitable ways to ensure those accessing a website or service will have a pleasurable experience. However, it appears that there are still websites hosting pages with dense text, convoluted instructions and little support for those with low levels of literacy. When considering poor reading skills, the cause may be due to many factors including a lack of education, sensory and /or intellectual impairments and specific difficulties such as dyslexia. This means that the vast majority of online content may be hard to understand for a significant proportion of the world's population. Moreover, these individuals may also lack digital skills, with little realization that assistive technologies and the availability of supportive access strategies can be helpful in these situations.

This paper aims to introduce the idea of enhancing readability of web content by using artificial intelligence (AI) techniques, such as linked data, natural language processing and image recognition to make available a wide range of automatically mapped multilingual symbols that can be used to clarify text content. In the past only a few symbol sets have been mapped and it was not possible to consider their appropriateness for text to symbol translations in a wide range of languages and cultural settings.

WebSci '20 Companion, July 6–10, 2020, Southampton, United Kingdom

ACM ISBN 978-1-4503-7994-6/20/07…$15.00
https://doi.org/10.1145/3394332.3402831

CCS CONCEPTS

• **Human-centered computing** → **Accessibility theory, concepts and paradigms**; • **Information systems** → *Personalization.*

KEYWORDS

readability, literacy support, multilingual symbols, artificial intelligence

ACM Reference Format:
EA Draffan, Chaohai Ding, Mike Wald, and Russell Newman. 2020. Multilingual Symbolic Support for Low Levels of Literacy on the Web. In *12th ACM Conference on Web Science (WebSci '20 Companion), July 6–10, 2020, Southampton, United Kingdom.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3394332.3402831

1 INTRODUCTION

As recently as 2017, UNESCO were reporting that "750 million adults – two-thirds of whom are women – still lack basic reading and writing skills". The benchmark for the 86% of those from age 15 and over who "can both read and write with understanding" is based on "a short simple statement on his/her everyday life" [7]. This does not seem to be a particularly high measure for an essential skill, with so much information being found online. UNESCO admit that many countries gather data about rates of literacy in different ways and there remains a concern about the standards achieved.

The issue arises when considering the amount of text that often appears on web pages without illustrations to aid understanding. There are over a billion websites available to online users[1], but content providers should note that readers tend to scan for key points [8] rather than read an entire page. These human behaviors have not changed according to the Nielsen Norman Group and their recent research has also shown that "reading patterns, are very similar across languages and cultures"[2]. Because people generally scan read web pages the importance of their readability in terms of ease of understanding and coping with the layout presented has become a much discussed area. It is included as a requirement in the W3C Web Content Accessibility Guidelines (WCAG 2.1) at level AAA, the highest of the three levels of compliance, which means that this requirement is often overlooked. However, the success

[1] https://news.netcraft.com/archives/category/web-server-survey/
[2] https://www.nngroup.com/articles/how-people-read-online/

criteria for 3.1.5 Reading Level states "When text requires reading ability more advanced than the lower secondary education level after removal of proper names and titles, supplemental content, or a version that does not require reading ability more advanced than the lower secondary education level, is available."[3] The techniques mentioned for offering supportive access strategies include the provision of:

- a text summary lower than secondary level
- visual illustrations, pictures, and symbols to help explain ideas, events, and processes
- a spoken version of the text
- text that is easier to read
- sign language versions of information

When considering reading skills, as opposed to literacy skills, which may encompass writing and spelling as well as numeracy, there are a range of complex strategies that need to be acquired. These include decoding skills, processing speeds for letter sound fluency as well as phonemic blending, sight word recognition and comprehension [1]. Education is key to gaining these skills as well as having the sensory and /or cognitive ability to cope with the content. Assistive technologies such as those mentioned in the WCAG techniques list, for example screen reading for those with visual impairments and text to speech for individuals with dyslexia, can also be very helpful.

But when reading is so difficult, that the words on the page are not understandable, the use of images, icons and symbols can aid comprehension. These images can be used as a form of text to symbol translation to suggest a concept or highlight a key point. This process is one that the authors have been exploring, as this has not been achieved in a way that is customised to allow for a user's preferred language and culture. Symbols can be highly personalized to represent local environmental settings, as well as being linguistically appropriate. The types of ideographic or pictographic symbols used by those with complex communication needs have been used in the past for this purpose [4]. In fact, individuals who have severe speech and language impairments may depend on these types of augmentative and alternative forms of communication (AAC) where the symbols are their language. The gloss or label to which the symbol concept is linked provides the text to speech output on a speech generating device or the symbols are used on a paper based communication chart and the user indicates their needs and ideas by pointing to them and a communication partner can read the labels. This linking of symbols to written concepts across languages and cultures means that several symbol sets have to be mapped to offer different choices to the wide range of potential users. This aim brings with it many challenges when considering the context of a word in any language on a website and attempting to find a matching symbol.

The vocabularies of the various symbol sets are small in comparison to the number of words used in English. Adult vocabulary test takers know from 20,000–35,000 words[4]. There are up to 12,000 - 14,000 symbols in some freely available pictographic symbol sets, but only two sets have been mapped based on an international standard, so interoperability between sets is rare. The work carried out by Mats Lundälv and colleagues [2] highlighted these issues when they introduced their Concept Coding Framework (CCF) using Blissymbolics[5] and ARASAAC symbols[6]. This work has since been taken up by a group of researchers developing ways of personalizing web pages to suit user needs. The fact that Bliss characters and words form both a Universal Character Set with a growing list of unique numerical identifiers for individual concepts, as well as a lexicon-based encoding ISO standard (ISO-IR 169), provides a robust base from which the authors of this paper can work. The link with the development of the 'Personalization Semantic Explainer'[7] forms the backdrop for offering enhanced interoperability between freely available symbol sets, with an increased number of languages. The aim will be to support, not only AAC symbol users, but also those with low levels of literacy who find it hard to read content on web pages.

2 METHODOLOGY

The initial goal is to enhance web content readability by providing symbolic representations of keywords found in the text on web pages. This requires the linking of various symbol sets so that individual symbols can be mapped with their concepts into one global repository. This will provide a universal and accessible way for those supporting struggling readers to search, select and change symbols, based on preference and cultural background. An API will be provided that allows a user agent to present the symbols to a web page reader when required. Several machine learning techniques will be used to improve individual symbol interoperability.

There are several steps in the proposed symbol mapping approach, which is presented in Figure 1. Text gloss or label preparation is the first step to process all extracted symbol labels from different symbol sets by using NLP techniques. This process includes text cleaning, removal of special characters, handling of ambiguous meaning, spelling correction and the extraction of parts of speech (PoS). Once the label preparation has been completed, the second step is to map the label text to the concept entities in ConceptNet[8].

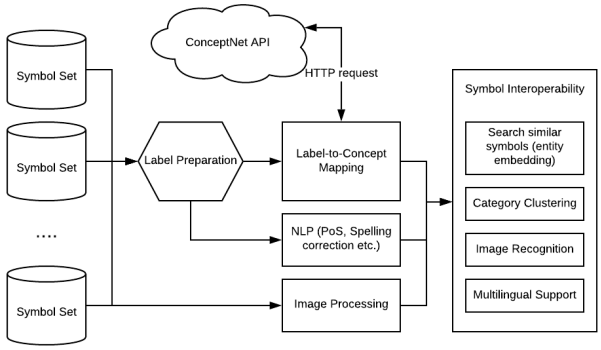

Figure 1: Symbol interoperability improvement framework

[3]https://www.w3.org/TR/UNDERSTANDING-WCAG20/meaning-supplements.html
[4]http://testyourvocab.com/blog/2013-05-10-Summary-of-results
[5]https://www.blissymbolics.org/
[6]http://www.arasaac.org/
[7]https://www.w3.org/TR/personalization-semantics-1.0/
[8]http://conceptnet.io/

ConceptNet is the knowledge graph version of the Open Mind Common Sense project, which provides the underlying source of information for symbol label mapping [6]. Compared with other lexical databases, ConceptNet provides semantic relationships between common concept entities with 78 different languages, including English, French, Arabic, Spanish, Urdu, Serbian and Chinese. With the advantages provided by the knowledge graph, the entities can be mapped based on their categories, functionalities and properties by using semantic linking. Examples of potential links include synonyms, a-form-of, part-of and related terms. These semantic links of concept entities can also be aggregated or grouped, based on inference and reasoning. Moreover, ConceptNet also provides the multilingual word embedding model, namely Numberbatch, which is built from the ground up, combining the advantages from other popular word embedding models (e.g. Glove [5] and Word2Vec [3]).

The use of ConceptNet and word embedding provided a semantic similarity measurement between different symbols, which was at the heart of the process used in the early stages of the repository development. However, preliminary results showed that there were a few problems with the current approach. For example, the label for the symbol 'car' also produced the symbol for a horse and cart and a carousel when using the ARASAAC symbol set as a test search. Neither result would have been helpful in a text to symbol translation, where a specified form of transport was required.

A decision was made to include image recognition as a supporting strategy to provide an increased amount of data directly related to the visual representation of the symbols. The early stage results have showed that some objects in the symbol picture can be detected and recognized by computer vision algorithms. The example demonstrated in Figure 2 shows how objects in the horse and cart symbol have been detected and recognized, such as wheels to denote a form of transport, but when the symbol for 'car' is analyzed the word is found with 62% certainty as well as the wheels. As result, the proposed approach will be used to improve symbol

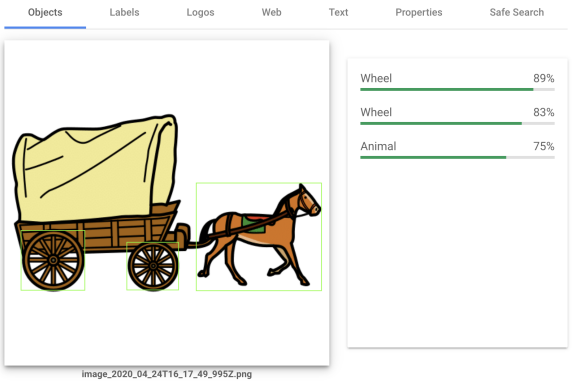

Figure 2: Symbol Image recognition (Google Vision AI)

interoperability across different symbol sets and also contribute to the enhancement of web content readability for end users.

3 RESULTS AND DISCUSSION

The work on the harmonization of the various symbols sets is still in progress. Nevertheless, the authors have discovered that depending on a solely semantically based linkage of concepts can lead to symbols not being found, due to failures when different parts of speech are used, but are derived from the same concept. A symbol for the verb 'to be' in the present tense 'is' would be found, but not 'was' or 'will'. However, these may be selected by an AAC user with a modifier, such as an arrow in one direction for past and in another direction for future. Another issue that occurred was where a label had multiple words, where only one should have been used to represent the concept, such as 'it'. These two problems happened with 17 percent of the 100 frequently used core words in English, published online by Hill and Romich[9] and used by AAC professionals. 17,388 symbols were mapped to ConceptNet and the full results have yet to be analysed. An initial scan through the concept list showed that confusions for potential users would arise where there were two or more symbols for one label. This was especially so if this was a homonym e.g. the word 'can' i.e. to be able or 'can' as a tin can. If a word like 'make' is used in a sentence, this could also be represented by different symbols, one meaning 'it is a requirement' – to make someone do something and another for 'the ability to create something'. This is obviously a problem that occurs in automatic language translation, but to a lesser extent, because context can be taken into account. As only a few symbols are usually used to signify some of the key words in a sentence, each one has to be as representative of the actual meaning as possible (Figure 3). Therefore, whereas initial work using ConceptNet with semantics produced a 70% chance of a good symbol to label match, the proposed combination of machine learning algorithms including word embedding and image recognition using deep neural networks has the potential to offer increased accuracy for text to symbol matches.

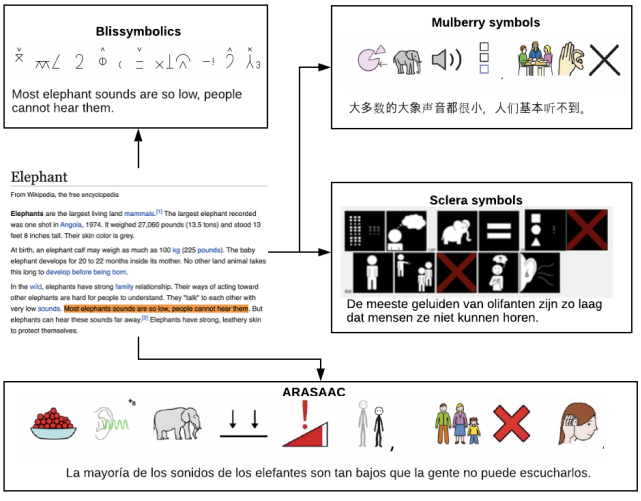

Figure 3: Sample text from Wikipedia supported by a choice of four different symbol sets in four languages

[9]https://aaclanguagelab.com/resources/100-high-frequency-core-word-listwords

There are several limitations to these ideas including the lack of freely available symbol sets with sufficient vocabularies and so in the course of the trials the intention is to include more symbol sets developed in different languages. This would allow for an increase in training data for the ConceptNet and word embedding approach, as well as improved results when using image recognition. In Figure 4, a symbol for park or playground resulted in only one element of the image being picked up and tagged as 'packaged goods', but there are several other symbol sets available with similar images that could be incorporated in the process. The use of image recognition and pattern classification will also improve symbol clustering for topic categorization and future research on context sensitive text to symbol and symbol to text work.

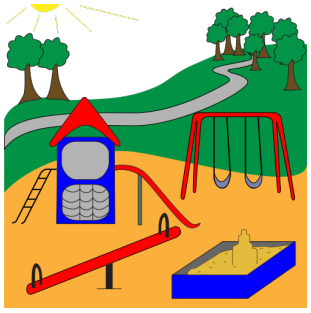

Figure 4: Playground or park recognized as packaged goods

However, there remain concerns around complete multilingual mapping, which also needs to be addressed, as some of the ConceptNet lexicons are incomplete, as are the culturally sensitive symbol sets with translations. This has an impact on less frequently used languages and it has also been found that some of the translations already available for the symbols sets are not always accurate.

4 CONCLUSION

Over several years researchers have attempted to harmonize AAC symbol sets that would allow for interoperability, meaning they could be used for text to symbol and symbol to text translations with ease. Invariably there has been the inescapable realization that much of the work entails human endeavor with a considerable amount of understanding to cope with the various differences between each symbol set. However, with the increased use of artificial intelligence some of the hurdles can be overcome. It is also accepted that there has already been a considerable amount of work carried out to ensure the standardization of Blissymbolics and the mapping against the ARASAAC symbol set, along with recent work on 'standard semantics to enable user-driven personalization'. Building on this work and using the latest AI techniques it should be possible to present stakeholders with a means of using a group of freely available harmonized multilingual AAC symbol sets for content clarification. Furthermore, the results of this work aim to support those with complex communication difficulties by providing chart building support using the linked symbol sets from the repository. This will mean users can access free symbols of their choice for use on assistive technologies and those supporting struggling readers or individuals who have low levels of literacy can access symbols to explain key words on the web.

ACKNOWLEDGMENTS

The authors would like to thank The Alan Turing Institute and The Web Science Institute at the University of Southampton for their sponsorship of their Alan Turing Pilot Project about AI and Inclusion[10].

REFERENCES

[1] Roxanne F. Hudson, Paige C. Pullen, Holly B. Lane, and Joseph K. Torgesen. 2009. The complex nature of reading fluency: A multidimensional view. *Reading and Writing Quarterly* 25, 1 (2009), 4–32. https://doi.org/10.1080/10573560802491208

[2] Mats Lundälv and Sandra Derbring. 2012. AAC Vocabulary Standardisation and Harmonisation. In *International Conference on Computers for Handicapped Persons*. Springer, 303–310.

[3] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*. International Conference on Learning Representations, ICLR. arXiv:1301.3781

[4] Eliada Pampoulou and Cate Detheridge. 2007. The role of symbols in the mainstream to access literacy. , 15–21 pages. https://doi.org/10.1108/17549450200700004

[5] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. (2014), 1532–1543. https://doi.org/10.3115/V1/D14-1162

[6] Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. (dec 2016). arXiv:1612.03975 http://arxiv.org/abs/1612.03975

[7] UNESCO Institute for Statistics. 2017. Literacy Rates Continue to Rise from One Generation to the Next. *Unesco* 2016, 45 (2017), 1–13. http://uis.unesco.orghttp://on.unesco.org/literacy-map.

[8] Harald Weinreich, Hartmut Obendorf, Eelco Herder, and Matthias Mayer. 2008. Not quite the average: An empirical study of Web use. *ACM Transactions on the Web* 2, 1 (feb 2008), 1–31. https://doi.org/10.1145/1326561.1326566

[10] urlhttps://www.turing.ac.uk/research/research-projects/ai-and-inclusion