Extending a Dutch Text-to-Pictograph Converter to English and Spanish

Leen Sevens, Vincent Vandeghinste, Ineke Schuurman, Frank Van Eynde

Centre for Computational Linguistics
KU Leuven, Belgium

firstname@ccl.kuleuven.be

Abstract

We describe how a Dutch Text-to-Pictograph translation system, designed to augment written text for people with Intellectual or Developmental Disabilities (IDD), was adapted in order to be usable for English and Spanish. The original system has a language-independent design. As far as the *textual* part is concerned, it is adaptable to all natural languages for which interlingual WordNet [1] links, lemmatizers and part-of-speech taggers are available. As far as the *pictographic* part is concerned, it can be modified for various pictographic languages. The evaluations show that our results are in line with the performance of the original Dutch system. Text-to-Pictograph translation has a wide application potential in the domain of Augmentative and Alternative Communication (AAC). The system will be released as an open source product.

Index Terms: Augmentative and Alternative Communication, Pictographic Languages, Text-to-Pictograph Translation

1. Introduction

In our daily lives, we are constantly confronted with pictographs. Think of traffic signs, signs in buildings that direct visitors to the elevators, the meeting rooms, the toilets, and the emergency exits, or signs for telling people that dogs need to be kept on a leash (see Figure 1).

Figure 1: Pictographs in our daily lives.

Similar pictographs are used as a form of Augmentative and Alternative Communication (AAC). AAC assists people with severe communication disabilities to be more socially active in interpersonal interaction, learning, education, community activities, employment, volunteering, and care management. Schools, institutions, and sheltered workshops use specific pictographs that are related to everyday activities and objects to allow accessible written communication between children or adults with Intellectual or Developmental Disabilities (IDD) and their caregivers, in an offline setting.

It is undeniable that current technological advances influence our lives in various aspects. Not being able to access or use information technology is a major form of exclusion. In order to reduce social isolation, there is an acute need for digital picture-based communication interfaces that enable contact for people with IDD. Adding pictographs to text can provide help in reading and understanding the text. It is estimated that between two and five million people in the European Union could benefit from symbols or symbol-related text as a means of written communication [2].

The Dutch Text-to-Pictograph translation system that is described in Vandeghinste et al. [3] is used in the WAI-NOT[1] communication platform. WAI-NOT is a Flemish, non-profit organization that gives people with severe communication disabilities the opportunity to familiarize themselves with computers, the internet, and social media. The website makes use of an email client that automatically augments written text with a series of Beta[2] or Sclera[3] pictographs. WAI-NOT's first translation system would rely on a simple one-on-one match between the input words and the pictograph file names, usually leading to erroneous translations and leaving many words untranslated. Vandeghinste et al. [3] improved this engine by introducing linguistic analysis. Their Text-to-Pictograph translation system was made as language-independent as possible.

Within the framework of Able to Include,[4] which aims to improve the living conditions of people with IDD, we built English and Spanish versions of this system. English and Spanish being a Germanic and a Romance language, respectively, we show that the engine manages to generalize well over different European language families.

After a discussion of related work (section 2), we introduce the Beta and Sclera pictograph sets (section 3), followed by an explanation of how existing links between WordNets can be used to automatically connect pictographs to words in source languages other than Dutch (section 4). In the remainder of this paper, we describe the system's general architecture (section 5). The evaluations (section 6) show that our results are in line with the performance of the Dutch system. Section 7 shows that the Text-to-Pictograph system has a wide application potential in the domain of AAC. Finally, we describe our conclusions and future work (section 8).

2. Related work

Pictographic communication has grown from local initiatives, some of which have scaled up to larger communities. Across Europe, many pictograph sets are in place, such as Blissymbolics,[5] PCS,[6] Pictogram,[7], ARASAAC,[8] Widgit,[9] Beta, and Sclera.

[1] http://www.wai-not.be/

[2] https://www.betasymbols.com/

[3] http://www.sclera.be/

[4] http://abletoinclude.eu

[5] http://blissymbolics.org/

[6] http://www.mayer-johnson.com/category/symbols-and-photos

[7] http://www.pictogram.se/

[8] http://www.catedu.es/arasaac/

[9] https://widgit.com/

Many of the problems that written languages encounter can be overcome by the use of pictographic languages. For instance, they can be understood across language barriers[10] [4] and there is less ambiguity involved. Pictographic communication systems for remote, online communication include Messenger Visual, an instant messaging service [5], Communicator [6], Pictograph Chat Communicator III [7], and VIL, a Visual Inter Lingua [4]. Mihalcea and Leong [8] argue that the understanding of graphical sentences is similar to that of target language texts obtained by means of machine translation. Leemans [4] shows that an appropriately designed iconic language, built according to a set of fixed principles, leads to no difference in the recognition rate of icons for people of western and non-western culture, yielding an average rate of about 79%. None of these abovementioned authors, however, consider users with IDD when designing the system.

Other pictograph-based communication systems are specifically designed for people with IDD. Patel et al. [9] introduce Image-Oriented Communication Aid, an interface using the Widgit symbol set, allowing users to build picture-supported messages on a touch screen computer. Motocos [10] are image exchange devices that are designed for children with autism, including audio cues for easier understanding of the image cards. The mobile application PhotoTalk [11] aids people with aphasia by providing a digital photograph management system in support of verbal communication. Nevertheless, all these systems require face-to-face communication in an offline setting.

The use of online information technology systems as a way to enhance the quality of life of people with IDD is a recent development. For accessible, remote communication, Keskinen et al. [2] introduce SymbolChat, a platform for picture-based instant messaging, where the interaction is based on touch screen input and speech output. The Text-to-Pictograph conversion system described in Vandeghinste et al. [3] applies shallow linguistic analysis to Dutch input text and automatically generates sequences of Beta and Sclera pictographs, allowing people with IDD to read messages independently. Only few other publications related to the task of translating texts for pictograph-supported communication can be found in the literature, such as Goldberg et al. [12] and Mihalcea and Leong [8], but these systems do not translate the whole sentence or are not focused on IDD.

3. Pictographic languages

Mihalcea and Leong [8] note that complex and abstract concepts (such as *democracy*) are not always easy to depict. Some characteristics of natural languages may not be present in the pictographic languages.[11] Usually, no distinction between singular and plural is made. Tense, aspect, and inflection information is removed, and so are the auxiliaries and the articles.[12] Pictographic languages are simplified languages, that are often specifically designed for people with IDD.

Although experiments with the Pictogram set [13] have revealed that many pictographs are difficult and wrongly interpreted, a correct interpretation is easily accepted and remembered without any problem. By giving people with speech and language disorders the opportunity to familiarize themselves with the pictographs, they learn to interpret the symbols more easily. However, a deliberate effort is needed.

The Text-to-Pictograph translation system currently gives access to two pictograph sets, Sclera and Beta (see Figure 2).

Sclera pictographs[13] are mainly black-and-white pictographs, although colour is sometimes used to indicate permission (green) or prohibition (red). Over 13,000 pictographs are available and more are added upon user request. Sclera pictographs often represent *complex* concepts, such as a verb and its object (such as *to feed the dog*) or compound words (such as *carrot soup*). There are hardly any pictographs for adverbs or prepositions.

The *Beta* set[14] consists of more than 3,000 coloured pictographs. Easy recognition being one of the main objectives, Beta is characterized by its overall consistency and the use of different types of arrows and dashes (pointing to an object, indicating changes in space or time or depicting actions). Beta hardly contains any complex pictographs. Most of the pictographs represent *simplex* concepts.

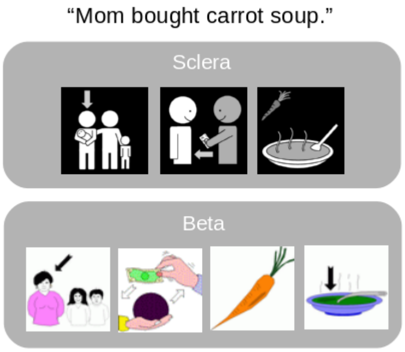


Figure 2: Example of a sentence being translated into Sclera and Beta pictographs. Tense information is removed. The Sclera translation contains a complex pictograph, namely *carrot soup*.

4. Linking pictographs to other WordNets

WordNets, lexical-semantic databases, are an essential component of the Text-to-Pictograph translation system. For the original Dutch system, Cornetto [14, 15] was used. Its English and Spanish counterparts are Princeton WordNet 3.0 [1][15] and the Spanish Multilingual Central Repository (MCR) 3.0 [16].[16] WordNets contain synsets (groupings of synonyms that have an abstract, usually numeric identifier, see Figure 3) and are designed in such way that each synset is connected to one or more lemmas.

Vandeghinste and Schuurman [17] manually linked 5710 Sclera pictographs and 2760 Beta pictographs to Dutch synsets in Cornetto.[17] An essential step in building Text-to-Pictograph translation systems for other languages is making sure that the pictographs are connected to (sets of) words in those languages.

[10] Although cultural differences remain.

[11] We use the term *pictographic language* in order to refer to the combination of individual pictographs, that belong to a specific *pictograph set*, into a larger meaningful structure.

[12] There are some exceptions. Beta, for instance, contains the Dutch articles.

[13] Freely available under Creative Commons License 2.0.

[14] The coloured pictographs can be obtained at reasonable prices, while their black-and-white equivalents are available for free.

[15] http://wordnet.princeton.edu/

[16] http://adimen.si.ehu.es/web/MCR/

[17] As a Cornetto license can no longer be obtained, the authors will transfer these links to the Open Source Dutch WordNet (http://wordpress.let.vupr.nl/odwn/).

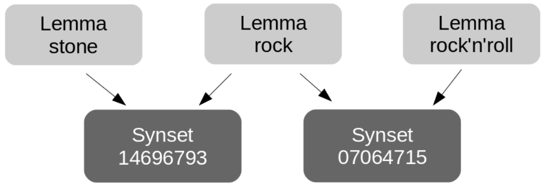


Figure 3: An example of a lemma, *rock*, having different meanings and belonging to different synsets. Two synsets are shown here.

Manually linking thousands of pictographs all over again would be a very time-consuming procedure. Instead, by transferring the connections automatically (see Figure 4), this process can be sped up drastically.

Sevens et al. [18] note that connections between WordNets are an important resource in knowledge-based multilingual language processing. The already mentioned Cornetto database for Dutch, used to build the Dutch Text-to-Pictograph translation system, contains connections to the English Princeton WordNet. We describe how we automatically connected Beta and Sclera pictographs to synsets in Princeton WordNet 3.0 in section 4.1.

Many WordNets nowadays contain high-quality links between the source language's synsets and Princeton WordNet 3.0, which is often viewed as the *central* WordNet. Princeton WordNet 3.0 now also plays this central role in our Text-to-Pictograph translation system. Having obtained the links between Beta and Sclera pictographs and Princeton WordNet 3.0, it becomes possible to automatically assign pictographs to synsets in any WordNet that has decent connections with Princeton WordNet,[18] allowing us to quickly build Text-to-Pictograph translation systems for many other languages. For example, with the English pictograph connections in place, a mapping between the pictographs and Spanish synsets in MCR 3.0 became possible. This process is described in section 4.2.

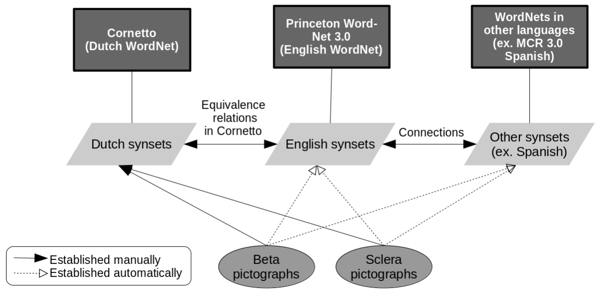


Figure 4: Making Princeton WordNet 3.0 the central WordNet of the Text-to-Pictograph translation system and transferring the links to the MCR 3.0 for Spanish.

4.1. Connecting pictographs to Princeton WordNet 3.0

Cornetto's *equivalence relations* establish connections between Dutch and English synsets in Princeton WordNet. These relations have originally been established semi-automatically by Vossen et al. [19], filling the database with more than 80000 links between Dutch and English synsets.

Sevens et al. [18] showed that a considerable amount of the original links were highly erroneous, making them not yet very reliable for multilingual processing. By using these equivalence relations, we would risk assigning pictographs to unrelated synsets in Princeton WordNet 3.0. In the case of a Dutch synset being wrongly connected to an English synset, writing a message in English would allow the system to generate pictographs that depict another concept. Therefore, we used the filtered,[19] more reliable connections that were established by Sevens et al. [18].

As a result, it became possible to automatically assign a large amount of Sclera and Beta pictographs to English synsets in Princeton WordNet 3.0. However, 154 (5.58%) Beta pictographs and 288 (5.04%) Sclera pictographs still had to be connected manually, either because the original equivalence relation was rejected by the filtering algorithm, or because the Dutch compound word corresponded to multiple words in English and forced us to treat the pictograph as a complex pictograph[20] in English (such as the Dutch word *vanillesuiker*, meaning *vanilla sugar* in English). In some rare cases, no equivalent English concept existed in the WordNet for an existing Dutch concept (for instance, the fictional character *Zwarte Piet* or typical kinds of food such as *choco*, which can roughly be translated as *chocolate spread*).

4.2. Connecting pictographs to the Spanish MCR 3.0

The MCR 3.0 integrates in the same EuroWordNet framework WordNets from five different languages, namely English, Catalan, Spanish, Basque, and Galician. Words in one language are connected to words in any of the other languages through Inter-Lingual-Indexes. Sevens et al. [18] showed that the links between English and Spanish synsets were correctly established, making it possible for us to create highly reliable connections between Beta and Sclera pictographs and Spanish synsets. This exact same process can be done for any language's WordNet that establishes reliable links to Princeton WordNet 3.0.

5. The Text-to-Pictograph translation system for English and Spanish

In this section, we describe how a textual message is converted into a sequence of Sclera or Beta pictographs [3] (see Figure 5), with an application to English and Spanish. The source text first undergoes shallow linguistic analysis (section 5.1). For further processing, two routes can be taken. The semantic route is only applied to content words (nouns, verbs, adjectives, adverbs) that are present in the WordNets. It consists of linking the source text to synsets in the databases (section 5.2) and retrieving the pictographs that are connected to these synsets (section 5.3). The direct route (section 5.4), which runs in parallel with the semantic route, contains specific rules for appropriately dealing with pronouns, and it uses a dictionary for parts-of-speech that are not present in the WordNets. The system contains a handful of parameters (section 5.5), which were tuned beforehand (section 5.6). Finally, as explained in section 5.7, an optimal sequence of pictographs is selected.

[18] A full list can be found on http://globalwordnet.org/wordnets-in-the-world/

[19] Filtering was done by using large bilingual dictionaries.

[20] A pictograph that is connected to multiple synsets instead of just one synset. For example, the pictograph depicting vanilla sugar is connected to both the synset that contains the lemma *vanilla* and the synset that contains the lemma *sugar*.

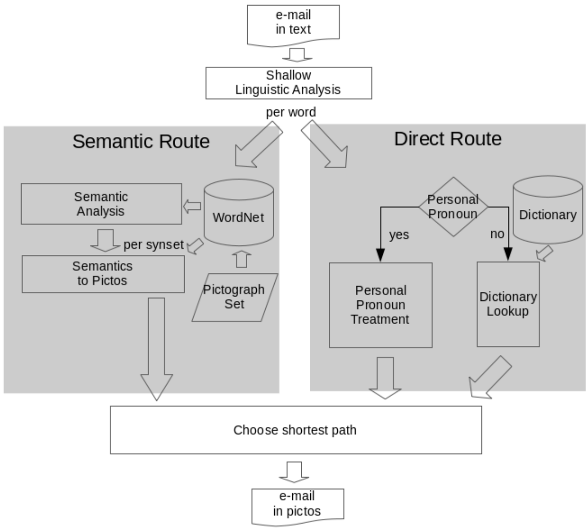
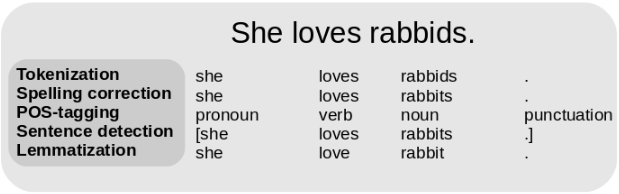


Figure 5: Architecture of the translation engine.

5.1. Shallow linguistic analysis

The source text undergoes shallow linguistic processing, consisting of several sub-processes (see Figure 6). This process is analogous to the linguistic processing step in the original Dutch tool.



She loves rabbids.

Tokenization	she	loves	rabbids	.
Spelling correction	she	loves	rabbits	.
POS-tagging	pronoun	verb	noun	punctuation
Sentence detection	[she	loves	rabbits	.]
Lemmatization	she	love	rabbit	.

Figure 6: An example of shallow linguistic processing.

First, *tokenization* is applied to split the punctuation signs from the words, with the exception of the hyphen/dash and the apostrophe, as they often belong to the word.

As the targeted users have different levels of illiterateness, basic *spelling correction* (one deletion, one insertion, one substitution)[21] aids in finding the correct variant of words that do not appear in the lexicon[22] and the list of first names.[23]

Next, *part-of-speech tagging* is applied. For English, we used HunPos [20], an open source tagger, using the English training data (with Penn Treebank tags[24]) made available on its website.[25] For Spanish, part-of-speech tagging (with TreeTagger tags[26]) and lemmatization are done in one step with TreeTagger [21].[27] TreeTagger is available for a large variety of European languages.

The Text-to-Pictograph translation system works on the sentence level. Although most messages sent by the users only contain one sentence, *sentence detection* is applied. Segmentation is based on full stops, which will eventually correspond to line breaks in the resulting pictographic representation.

The next step is *lemmatization*, which requires a language-specific treatment. For English, we built a lemmatizer based on a list of English token/part-of-speech combinations and their lemma.[28] As mentioned before, for Spanish, part-of-speech tagging and lemmatization are done with TreeTagger.

One additional adaptation concerns the treatment of the Spanish *pro-drop* phenomenon (which occurs in all Romance languages, with the exception of French), meaning that personal pronouns in subject position are usually omitted (unless emphasis is given). Translating such a message into pictographs would leave us with no subject, as the pictographic representations of words are based on the lemma form and do not retain any grammatical information. However, person information can be inferred from the verb in the source sentence. We wrote a set of rules that explicitly adds the personal pronouns in the message before converting it into a series of pictographs.[29] When a matching personal pronoun is already found within a window of three words (since adverbs or pronouns can appear between the subject and the verb), these rules are not applied (see Figure 7).

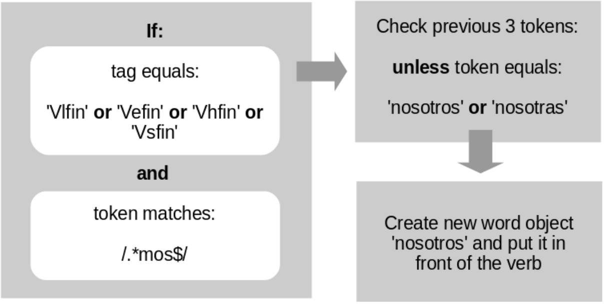


Figure 7: An example of a pro-drop rule. The tags correspond to *finite lexical verb*, *finite estar (to be)*, *finite haber (to have)*, and *finite ser (to be)*. The token has to end on *-mos*, which indicates a first person plural form. *Nosotros* and *nosotras* correspond to the English pronoun *we*.

5.2. Semantic analysis

The first step in the semantic analysis consists of the detection of words with a negative polarity, such as *not/no* and *no/ningún*. When such a word is found, the system looks for its head (a verb or a noun) and adds the value *negative* to its polarity feature.

For each word in the source text, the system returns all possible WordNet synsets (see section 4). The synsets are filtered,

[21] We are currently designing a spelling corrector that is specifically tailored towards Dutch text written by people with IDD. Our approach does not rely on the use of parallel corpora (erroneous text - corrected text). Therefore, it can also prove to be useful for spelling correction in other languages.

[22] http://www.anc.org/SecondRelease/frequency2.html (for English) and http://corpus.leeds.ac.uk/frqc/internet-es-forms.num (for Spanish)

[23] http://www.quietaffiliate.com/free-first-name-and-last-name-databases-csv-and-sql (for English and Spanish)

[24] http://www.cis.upenn.edu/ treebank

[25] https://code.google.com/p/hunpos/downloads/list

[26] http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/data/spanish-tagset.txt

[27] http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

[28] http://www.anc.org/

[29] When the verb is a third person singular or plural, these rules are not applied, as its subject could be a noun phrase. This problem can be solved by applying deeper grammatical analysis. Gender information (*he*, *she*, *it*), however, cannot be inferred from the verb alone and requires deeper semantic knowledge.

keeping only those where the part-of-speech tag of the synset matches the part-of-speech tag of the word.

Certain links between lemmas and synsets can be disabled in order to remove unwanted, often sexual meanings of common words, which are not appropriate for some groups of users (such as one meaning of the word *member*).

5.3. Retrieving the pictographs

The WordNet synsets described in section 5.2 are used to connect pictographs to natural language text. This way, the lexical coverage of the system is greatly improved, as pictographs are connected to sets of words that have the same meaning, instead of just individual words. Additionally, if a synset is not covered by a pictograph, the links between synsets can be used to look for alternative pictographs with a similar meaning. For instance, the *hyperonymy* relation can be used if no pictograph is found for a concept that is too specific (such as *rabbit* for *cottontail*, see Figure 8). The *antonymy* relation, indicating that synsets are the opposite of each other, selects a pictograph of the antonym, along with a negation pictograph (such as *not sick* for *recovered*). The *XPos* relation concerns similar words with a different part-of-speech tag (such as the adjective *female* for *woman*). However, using pictographs through *synset propagation* (making use of the WordNet relations) is controlled by parameters or penalties for not using the proper concept (see section 5.5).

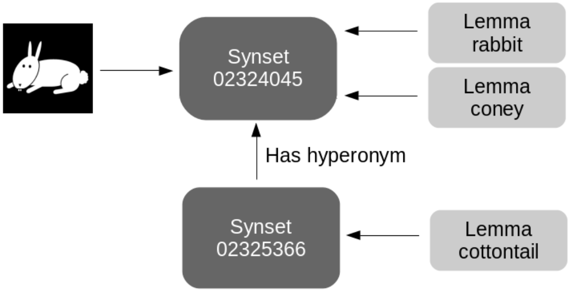


Figure 8: When a specific word, such as *cottontail*, does not have a pictograph connected to its synset, WordNet relations will be used to find a similar concept and display its pictograph instead. The synset for *rabbit* and *coney* (a synonym of *rabbit*) is found.

5.4. The direct route

The English and Spanish WordNets contain nouns, verbs, adjectives, and adverbs. To deal with pronouns and words that have a part-of-speech tag that is not covered by the WordNets, the direct route is introduced.

To make sure that *personal and possessive pronouns* are covered, they are given an explicit treatment. Person, gender, and number information can be obtained during the part-of-speech tagging process, resulting in correct pictograph translations.

The *dictionary* provides a direct link between the *token/lemma/tag* and the names of the pictographs. The *tag* field and either the *lemma* or *token* field can be left underspecified. For instance, in Sclera, there is a direct link between the lemma *hey* and the pictograph *hallo-zeggen-2.png* (*to say hello*), while the verb *miss* needs an additional *verb* tag to avoid confusion with the noun. The dictionary is used to cover any words that are missing from the database, because their part-of-speech tag is not included in the WordNet database (such as various types of greetings), or because the concept is too recent (such as *tablet*), among other things.

5.5. The parameters

For every word in the sentence, the system checks whether one or more pictographs can be found for it and whether the use of these pictographs is subject to a penalty. Penalties correspond to parameters that were tuned beforehand.

The first set of parameters (*hyperonym* penalty, *antonym* penalty, and *XPos* penalty) concern the maximum distance (*threshold* parameter) allowed between the original text and the pictographic message in terms of synset relations (see section 5.3).

The second set of parameters is related to the *numeric features* of the pictographs (*no number* and *wrong number*), as some pictographs make a distinction between singular or plural concepts (such as *oog.png*, depicting one eye, and *ogen.png*, depicting two eyes).

The last set of parameters determines the behaviour as to *the route to take*. An *out-of-vocabulary* parameter penalizes for leaving a content word untranslated, while the *direct route* parameter is a negative penalty (i.e. a bonus) for choosing the direct route over the semantic route.

Furthermore, the use of complex pictographs, which reunite multiple concepts within one pictograph (see section 3), will be preferred by the system over the separation of those concepts. The shorter the pictographic translation is, the higher it will be scored by the system (see section 5.7).

5.6. Tuning the parameters

The parameters that are mentioned in section 5.5 are tuned for every natural language/pictographic language pair. Ideally, tuning would be based on emails or text messages written by people with IDD. These messages are usually short, tend to refer to everyday life and very often contain spelling mistakes, like tweets.[30] As we did not have a large corpus of messages written by the targeted users at our disposition, we selected 75 English tweets and 75 Spanish tweets based on the following criteria: the messages should contain at least 8 words, they have to refer to personal experiences (no citations or lyrics), and they are allowed to contain spelling mistakes or lack punctuation marks. The tweets were retrieved by searching for messages containing the hash tags *#school/#escuela*, *#love/#amor*, *#family/#familia*, *#happy/#feliz*, and *#sad/#triste*.

For both languages, we manually translated, to the best of our ability, all tweets into Beta and Sclera pictographs. We built a local hill climber that varies the parameters (see section 5.5) when running the Text2Picto script on each of the four test sets (from English and Spanish to Beta and Sclera). The BLEU metric [22] was used as an indicator of relative improvement. In order to maximize the BLEU score, we ran five trials of a local hill climbing algorithm for each natural language/pictographic language pair. We did this until BLEU converged onto a fixed score after several thousands of iterations. Each trial was run with random initialization values, while varying the parameters between certain boundaries and with a granularity (size of the parameter steps) of one in order to cover different areas of the search space. From these trials, we took the best scoring parameter values for all four language/pictographic language pairs.

[30] https://twitter.com/

		With proper names		Without proper names	
Condition	**Precision**	**Recall**	**F-Score**	**Recall**	**F-Score**
Sclera					
Baseline	71.37%	61.25%	65.92%	62.25%	66.50%
Add frequent concepts	93.30%	71.95%	81.25%	73.04%	81.94%
Rel. improv.	*30.73%*	*17.47%*	*23.26%*	*17.33%*	*23.22%*
Beta					
Baseline	75.08%	70.63%	72.78%	71.71%	73.36%
Add frequent concepts	82.56%	85.07%	83.80%	86.14%	84.31%
Rel.improv.	*9.96%*	*20.45%*	*15.14%*	*20.12%*	*14.93%*

Table 1: Manual evaluation of the English system

		With proper names		Without proper names	
Condition	**Precision**	**Recall**	**F-Score**	**Recall**	**F-Score**
Sclera					
Baseline	73.84%	57.63%	64.74%	58.30%	65.16%
Add frequent concepts	93.31%	82.17%	87.38%	83.14%	87.93%
Rel. improv.	*26.37%*	*42.58%*	*34.97%*	*42.61%*	*34.95%*
Beta					
Baseline	83.48%	60.83%	70.38%	61.26%	70.66%
Add frequent concepts	94.64%	86.01%	90.12%	86.83%	90.57%
Rel.improv.	*13.37%*	*41.39%*	*28.05%*	*41.74%*	*28.18%*

Table 2: Manual evaluation of the Spanish system

5.7. Selecting the optimal path

An A* algorithm[31] calculates the optimal pictographic sequence for the source text. Its input is the pictographically annotated source message, together with the pictographs' penalties, depending on the number and kind of synset relations the system had to go through to connect them to the words.

The algorithm starts with a queue containing an empty path that still has all the input words left to process. In every step, the currently best scoring pictograph path is extended. We check whether there are any pictographs, with their corresponding penalties, connected to the next word that has to be processed.[32] New paths are thus created by adding the retrieved pictograph to the list of the already matched pictographs. All possible paths are added to the queue. The queue is sorted by lowest estimated cost and the best scoring path is extended. This process is repeated until the first queue element no longer has any words left to process.

When encountering words that have their *antonym* feature set to *negative* (see section 5.2), we insert the negation pictograph.

6. Evaluation

At the time of our evaluation, we did not yet have a corpus of messages written by people at IDD at our disposition. An evaluation set was built using the selection procedure as described in section 5.6. A total of 50 English tweets and 50 Spanish tweets were retrieved.

After having obtained the system's output translations for every message from the evaluation set, we performed a manual verification with one judge, who removed untranslated non-content words (such as *just*, *although*, and *it* in English). This allowed calculating the recall. For each of the translated words, she judged whether the pictograph generated was the correct pictograph, in order to calculate precision. As proper names occur rather frequently in online environments, we have calculated recall and F-score with and without proper names, in the latter case removing all proper names from the output. Precision remains the same in both conditions. In the case where proper names are included, they are not converted into pictographs, affecting recall negatively. In applications, similar to an option that is currently available in the WAI-NOT environment, proper names occurring in the contact lists of the users can be converted into the photographs that are attached to user profiles, resulting in more personalized messages.

Using the automatic pictograph connections that Sevens et al. [18] created by using the links between Cornetto synsets and Princeton WordNet synsets and the links between Princeton WordNet synsets and Spanish MCR synsets, a baseline system could be built. This system, which is not subject to any post-editing actions in the WordNet databases, leaves us with F-Scores of 66.50% and 73.36% for Sclera and Beta, respectively, for English text without proper names. For Spanish, F-Scores of 65.16% and 70.66% are obtained. A decent baseline system was thus created by making use of the previously available WordNet relations.

To improve the English and Spanish systems, we added or edited the 500 most frequently used words according to the Dutch WAI-NOT corpus,[33] in order to cover the specific vocabulary that the target group uses to address their peers or caregivers. For each one of these words, we translated them into English and Spanish and checked whether the right pictograph was connected to its synset. If this was not the case, we disabled the erroneous pictographs or created new pictograph connections. Sometimes, the pictograph dictionary (direct route) was used to add missing words to the database, such as different types of greetings. As a result, the English system currently yields F-Scores of 81.94% and 84.31% for Sclera and Beta, respectively, while the Spanish system reaches F-Scores of 87.93% and 90.57%, both for text in which proper names are omitted.

These results are comparable to the manual evaluations for Dutch [3]. The authors obtain F-Scores of 87.16% and 87.27% for Sclera and Beta translations of Dutch IDD text, respectively.

7. Application potential

The Text-to-Pictograph translation system will be released as an open source product, allowing developers to build pictograph-supported AAC applications and web browser extensions.

The pictographs are not meant to replace written text. They can be used as a stepping stone towards a better comprehension of written content.

Since textual content on the web, in particular long or difficult words, is sometimes very challenging for the target group to deal with, a Text-to-Pictograph translation system in the form of a web browser extension could be a welcome addition for many users. Web browser extensions are programs that extend the functionality of a web browser. For instance, by hovering over a difficult word, the program could show the pictographic representation of that word. This idea has already been implemented by the creators of Widgit, although their Point system[34] does not make use of semantic networks to simplify extension to additional languages.[35]

The system offers the possibility for family members, caregivers, and teachers to build pictographic messages more easily. Browsing large databases to find the appropriate icons is a long and tedious job, that can be facilitated by automatically translating a textual message into a series of pictographs. This

[31] A pathfinding algorithm that uses a heuristic to search the most likely paths first.

[32] If a complex pictograph is retrieved, the system checks whether the other synsets that belong to that complex pictograph are connected to any of the remaining words to process. If this is the case, the word that is linked to that synset is removed from the list of words to process.

[33] A corpus containing more than 40000 e-mails sent by users with IDD and their caregivers. Most e-mails are about their everyday life.

[34] https://widgit.com/products/online.htm

[35] We thank the anonymous reviewers for this observation.

way, pictograph-supported instructions, schedules and menus will become easier to construct. Text-to-Pictograph translation will also allow the family members and caregivers to send pictographic e-mails to the target group, making it simpler to communicate in an online setting, where the use of written text would normally cause big difficulties.

Within the Able to Include framework, a mobile app is currently being developed to address a variety of scenarios in which pictographs offer support. The tool will also integrate text-to-speech and text simplification technologies. The user can choose a technology (or a combination of technologies, such as text simplification followed by translation into pictographs) that he or she feels most comfortable with.

While our system is initially focused on users with IDD (since the tool was developed on the request of WAI-NOT, a website for people with disabilities), its general architecture can be reused in various other contexts, such as education, language learning for non-native speakers, and translation into sign languages.

8. Conclusions and future work

We have shown how the Dutch Text-to-Pictograph translation system can be extended towards other languages. To implement new languages, only a few components are required: decent connections between the source language's WordNet and the Princeton WordNet 3.0 (as we have shown for Spanish), a language-specific part-of-speech tagger and lemmatizer, a new set of parameters to optimize the system's performance and possibly some additional rules to deal with language-specific properties.

Future work will consist of improving the English and Spanish systems. Proper word sense disambiguation will have to be applied, as the system currently only takes the most frequent sense for a given word. We will look into possibilities for better spelling correction, specifically tailored towards text written by people with cognitive disabilities, and simplification of the pictographic output. Finally, the inverse relation, pictograph-to-text translation, will also be taken care of, allowing users to create textual messages by selecting a series of pictographs [23].

In collaboration with Faculty of Psychology and Educational Sciences of KU Leuven and our Able to Include partners, the pictograph translation system will be tested by the target group. The results will give us better insights concerning the usability of the engine.

Analysis of text written by English and Spanish users with IDD will reveal which concepts are missing from the databases and we will continue to improve the coverage of the system.

9. References

[1] G. Miller, R. Beckwidth, C. Fellbaum, D. Gross, and K. Miller, "Introduction to WordNet: An On-line Lexical Database," *International Journal of Lexicography*, vol. 3, no. 4, pp. 235–244, 1990.

[2] T. Keskinen, T. Heimonen, M. Turunen, J. Rajaniemi, and S. Kauppinen, "SymbolChat: A Flexible Picture-based Communication Platform for Users with Intellectual Disabilities," *Interacting with Computers*, vol. 24, no. 5, pp. 374–386, 2012.

[3] V. Vandeghinste, I. Schuurman, L. Sevens, and F. Van Eynde, "Translating Text into Pictographs," *Natural Language Engineering*, Accepted.

[4] P. Leemans, *VIL: A Visual Inter Lingua*. Dissertation. Worcester Polytechnic Institute., 2001.

[5] P. Tuset, J. Barbern, P. Cervell-Pastor, and C. Janer, "Designing Messenger Visual, an Instant Messenging Service for Individuals with Cognitive Disability," in *IWAAL 1995 – Proceedings of 3rd International Workshop on Ambient Assisted Living*, 1995, pp. 57–64.

[6] T. Takasaki and Y. Mori, "Design and Development of a Pictogram Communication System for Children around the World," in *IWIC 2007 – Proceedings of the 1st International Conference on Intercultural Collaboration*, 2011, pp. 193–206.

[7] J. Munemori, T. Fukada, M. Yatid, T. Nishide, and J. Itou, "Pictograph Chat Communicator III: a Chat System that Embodies Cross-Cultural Communication," in *KES 2010 – Proceedings of the 14th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems: Part III*, 2012, pp. 473–482.

[8] R. Mihalcea and C. Leong, "Toward Communicating Simple Sentences Using Pictorial Representations," *Machine Translation*, vol. 22, no. 3, pp. 153–173, 2009.

[9] R. Patel, S. Pilato, and D. Roy, "Beyond Linear Syntax: an Image-Oriented Communication Aid," *ACM Journal of Assistive Technology: Outcomes and Benefits*, vol. 1, no. 1, pp. 57–66, 2004.

[10] G. Hayes, S. Hirano, G. Marcu, M. Monibi, D. Nguyen, and M. Yeganyan, "Interactive Visual Supports for Children with Autism," *Personal Ubiquitous Computing*, vol. 14, no. 1, pp. 663–680, 2010.

[11] M. Allen, J. McGrenere, and B. Purves, "The Field Evaluation of a Mobile Digital Image Communication Application Designed for People with Aphasia," *ACM Transactions on Accessible Computing*, vol. 1, no. 1, 2008.

[12] A. Goldberg, X. Zhu, C. Dyer, M. Eldawy, and L. Heng, "Easy as ABC? Facilitating Pictorial Communication via Semantically Enhanced Layout," in *CoNLL 2008 – Proceedings of the Twelfth Conference on Computational Natural Language Learning*, 2008.

[13] K. Falck, *The Practical Application of Pictogram*, Lycksele, 2001.

[14] P. Vossen, I. Maks, R. Segers, and H. van der Vliet, "Integrating Lexical Units, Synsets, and Ontology in the Cornetto Database," in *LREC 2008 – Proceedings of the 6th International Conference on Language Resources and Evaluation*, 2008.

[15] H. van der Vliet, I. Maks, P. Vossen, and R. Segers, "The Cornetto Database: Semantic issues in Linking Lexical Units and Synsets," in *EURALEX 2010 – Proceedings of the 14th EURALEX 2010 International Congress*, 2010.

[16] A. G. Agirre, E. Laparra, and G. Rigau, "Multilingual Central Repository Version 3.0: Upgrading a Very Large Lexical Knowledge Base," in *LREC 2012 – Proceedings of the 8th International Conference on Language Resources and Evaluation*, 2012.

[17] V. Vandeghinste and I. Schuurman, "Linking Pictographs to Synsets: Sclera2Cornetto," in *LREC 2012 – Proceedings of the 9th International Conference on Language Resources and Evaluation*, 2014, pp. 3404–3410.

[18] L. Sevens, V. Vandeghinste, and F. Van Eynde, "Improving the Precision of Synset Links Between Cornetto and Princeton WordNet," in *LG-LP 2014 – Proceedings of the COLING Workshop on Lexical and Grammatical Resources for Language Processing*, 2014.

[19] P. Vossen, L. Bloksma, and P. Boersma, *The Dutch Wordnet. EuroWordNet Paper*, Amsterdam, 1999.

[20] P. Halácsy, A. Kornai, and C. Oravecz, "HunPos - an Open Source Trigram Tagger," in *ACL 2007 – Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Companion Volume, Proceedings of the Demo and Poster Sessions*, 2007, pp. 209–212.

[21] H. Schmid, "Improvements in Part-of-speech Tagging with an Application to German," in *SIGDAT 1995 – Proceedings of the ACL SIGDAT-Workshop*, 1995.

[22] K. Papineni, S. Roukos, T. Ward, and W. Zhu, “Evaluation of Machine Translation,” in *ACL 2002 – Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[23] L. Sevens, V. Vandeghinste, I. Schuurman, and F. Van Eynde, “Natural Language Generation from Pictographs,” in *ENLG 2015 – Proceedings of the 15th European Workshop on Natural Language Generation*, 2015.