Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 166–171
Marseille, 11–16 May 2020

Providing semantic knowledge to a set of pictograms for people with disabilities: a set of links between WordNet and Arasaac: Arasaac-WN

D. Schwab, P. Trial, C. Vaschalde, L. Vial, E. Esperança-Rodier, B. Lecouteux
Univ. Grenoble Alpes, LIG, 38000 Grenoble, France
{didier.schwab, pauline.trial, celine.vaschalde, loic.vial,
Emmanuelle.esperanca-rodier, benjamin.lecouteux}
@univ-grenoble-alpes.fr

Abstract

This article presents a resource that links *WordNet*, the widely known lexical and semantic database, and *Arasaac*, the largest freely available database of pictograms. Pictograms are a tool that is more and more used by people with cognitive or communication disabilities. However, they are mainly used manually via workbooks, whereas caregivers and families would like to use more automated tools (use speech to generate pictograms, for example). In order to make it possible to use pictograms automatically in NLP applications, we propose a database that links them to semantic knowledge. This resource is particularly interesting for the creation of applications that help people with cognitive disabilities, such as text-to-picto, speech-to-picto, picto-to-speech...In this article, we explain the needs for this database and the problems that have been identified. Currently, this resource combines approximately 800 pictograms with their corresponding *WordNet* synsets and it is accessible both through a digital collection and via an SQL database. Finally, we propose a method with associated tools to make our resource language-independent: this method was applied to create a first text-to-picto prototype for the French language. Our resource is distributed freely under a Creative Commons license at the following URL: https://github.com/getalp/Arasaac-WN.

Keywords: Assistive Technology for people with disabilities, Pictograms, Augmentative and Alternative Communication

1. Introduction

Communicating when oral and signed language is inaccessible or difficult is a real challenge for individuals with multiple disabilities and their family and friends. The use of an alternative communication code in daily activities can be hindered by learning difficulties, negative representations of non-verbal communication, or costly communication tools, causing significant frustration for those involved.

To enable individuals with language disabilities to communicate, there are several methods of Augmentative and Alternative Communication (AAC) (Nègre, 2017). These methods are called "alternative" when they completely replace the means of oral expression and are called "augmentative" when they make it possible to supplement the communication skills already present, which may even help the emergence of oralization or graphics for certain individuals (Beukelman and Mirenda, 2017).

For many people with multiple disabilities, gestures and speech are impossible and only the communication by pictograms is possible. A pictogram can be defined as a schematic graphic sign whose signifier has a more or less strong similarity with the signified, unlike phonic or graphic linguistic signs whose stimulus form is arbitrary and independent of that of the referent. It allows a more iconic representation of the information and is more easily interpretable. The development of this type of tool also corresponds to an increasingly strong social demand to develop accessibility: it is then a question of presenting a certain image to ask for a drink, to present another to ask for food (sometimes by designating the gaze through an oculometer (Schwab et al., 2018)).

The core of this resource establishes the basis for future tools related to pictograms such as, for example, the automation of pictogram communication.[1] This is indeed a need observed both in the structures for welcoming people with disabilities and among relatives: they cannot communicate with their environment in a traditional way with their voice, and sometimes not even with gestures. Communication via pictograms is thus widely used in institutions and is beginning to be standardised.

We propose to establish a relationship between two resources that are unanimously recognized: on one hand the *Princeton WordNet* lexical database and on the other hand the *Arasaac* pictogram set. Indeed, at the moment, it is difficult to automatically use the *Arasaac* pictogram set in a Natural Language Processing (NLP) pipeline, because there is no associated semantic information available.

One part of this work was presented in the French conference TALN (Schwab et al., 2019) as a demonstration. For visibility and dissemination purposes, we are presenting an extended version to LREC.

Firstly we will explain why and how pictograms are used in communication. Secondly we will successively present the resources we have linked: *Arasaac* and *WordNet* and show potential applications related to our resource. Then we will present our resource and the protocols we used to design it. Finally, we will show the interest of using *WordNet* to project our resource on French language that is not the one used by *WordNet*.

[1] See for instance the following video: https://www.youtube.com/watch?v=tYVkRsFpwtg

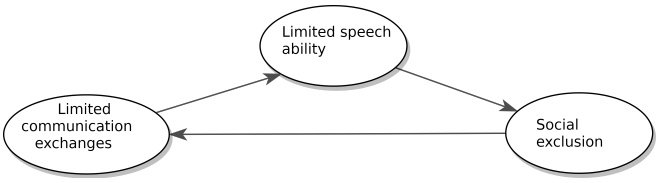

Figure 1: Vicious circle of Vygotskij

2. From pictograms to natural language

Pictograms, thanks to their iconicity, can help people to communicate in a foreign country when they do not speak the local language and do not share any linguistic background with local inhabitants. Mihalcea and Leong (2008) have shown that pictogram translations can help people who do not share the same language to communicate.
Pictograms are also a recognized tool in institutions, particularly for communicating with people with cognitive problems: people who do not speak are poorly or very poorly integrated into society. This Vicious Circle of Vygotskij is demonstrated in the 1930s among deaf children (Petitpierre and Barisnikov, 1994). Pictograms are then used as a vector of communication in order not to socially disconnect people (Figure 1).
Whether in institutions or in everyday life, pictograms are used manually: they are stored in workbooks and caregivers have to find them to show them.
This search is extremely time consuming: finding the required pictogram in a communication workbook is an uneasy task. The caregivers have to learn how to use the communication tool and, if it is a physical communication workbook, they have to spend time to look for the relevant pictogram. Because of this complicated navigation, interaction is not spontaneous and can even be perceived as really negative. A frequent request is to automate this retrieval in order to make it closer to natural language.
Currently, there is no knowledge base that formally link pictograms from *Arasaac* to their semantic representation. Generally, users of pictograms will choose them according to a textual description, a graphic representation, a user manual or following training. This association between a pictogram and a semantic representation is nevertheless an essential element for creating NLP tools using the pictograms. An equivalent work has already been done for another lexical database and another set of pictograms (Vandeghinste and Schuurman, 2014), but unfortunately the lexical inventory is specific for Dutch, and those pictograms are not used by any institution in France. We therefore present a first database making this link based on pictograms that are both free and used in institutions. Our objective is to promote the development of various tools around pictograms, we wish this database to be enriched over the years. The next section presents the *Arasaac* pictogram set that we have chosen to use in this work.

2.1. The open and free pictogram set: *Arasaac*

Arasaac is a blog for sharing free resources from Aragon (Spain) dedicated to augmentative and alternative communication. The *Arasaac* portal offers graphic resources and materials to facilitate communication for people with difficulties in this area. The Aragonese portal provides more than thirteen thousand black and white pictograms and more than fifteen thousand colour pictograms. This project was funded by the Department of Education, Culture and Sports of the Government of Aragon and coordinated by the Department's Directorate General of Innovation, Equality and Participation. This makes it the largest freely available pictogram database. These pictograms have a Creative Commons BY-NC-SA license that allows them to be used for non-commercial purposes, as well as the creation of derivative works, provided they are distributed under the same license. All the available resources are:

- A bank of 8000 downloadable pictograms, in colour or black and white, in constant progression,
- An image bank,
- A video bank LSE (Spanish Sign Language)

In addition to their availability, these resources can be manipulated online using three main tools that are highly relevant:

- A sentence generator in pictograms,
- A symbol generator (image + legend + frame...)
- An animated symbol generator (.gif)

We also chose *Arasaac* because in practice it is one of the most widely used pictograms in France for institutions and families caring for people with disabilities.
The Figure 2 shows examples of pictograms from *Arasaac*. We can see that some pictograms are very easy to understand, while others are much less intuitive. Linking the pictograms to *WordNet* provides a clear semantic representation.
We would like to point out that the sentence generator to pictograms is based only on a mapping word-to-pictogram. This tool therefore shows limitations very quickly in the context of real use, where a word can have several very distinct meanings. It is mainly for this type of application that we propose to link each pictogram to a complete semantic representation. For this purpose, we propose to use the *WordNet* database.

2.2. The *WordNet* database

Princeton WordNet (Miller et al., 1990), or simply *WordNet*, is a lexical database developed by linguists at Princeton University's Cognitive Sciences Laboratory for the past 20 years. Its purpose is to identify, classify and relate the semantic and lexical content of the English language in various ways. *WordNet* versions for other languages exist, but the English version is the most complete to date.
The database and tools are available free of charge. In addition to the tools provided, a developer can also access the database from the interfaces available for several programming languages (Java, PHP, Python...).
WordNet is distributed under a free license, allowing it to be used commercially or for research purposes.
The atomic component on which the entire system is based is the synset (synonym set), a group of interchangeable

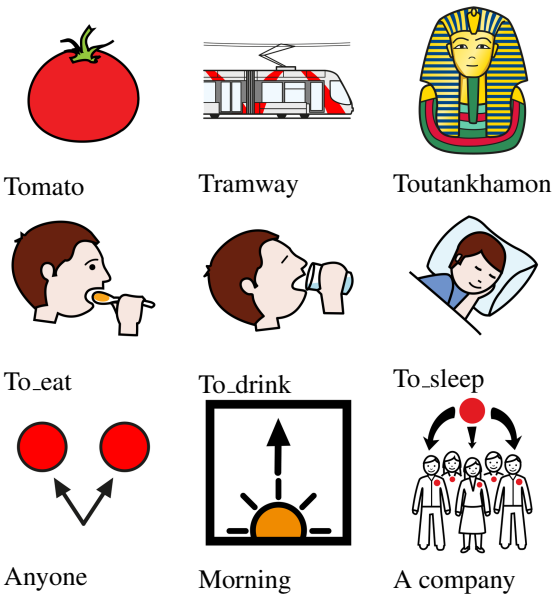

Figure 2: Example of pictograms: they are not always explicit at a first glance.

word senses, denoting a particular meaning or use. Each synset denotes a specific meaning of a word, described by a short definition. A particular occurrence of this word denoting for example the first (most common) meaning, in the context of a sentence or statement, would thus be characterized by the fact that the polysemous word could be replaced by one or the other of the words in the synset without altering the meaning of the whole.
Currently, our work focuses on the link between pictograms and *WordNet*, for further uses on any target language such as French. We have chosen to use *WordNet*, and not a specific French lexical database, because, from our point of view, it is the most complete and reliable database and it is consistently maintained. We will detail in this article our method to project *WordNet* in other languages in a relatively easy way.
In the next section, we show the applications that can result from a link between *WordNet* and *Arasaac*.

2.3. Examples of applications

Once the link between a pictogram database and *WordNet* has been established, it is possible to consider many applications. We present two of them wich we would like to develop:

1. Voice generation from pictograms that would allow a person with a disability to compose messages in the form of a synthesized voice. The problem thus returns to the search for the pictograms necessary for the composition of the message; this search is essential because the speed of such a composition is reputed to be ten times slower than writing. The organization of ideas is a problem discussed since the Greek philosophers. Its transposition to pictograms is therefore also one of them. *Arasaac-WN* would thus allow a search according to the ideas associated with the images, passing through the text of the definitions and/or the lexical network, for example in the manner of Zock and Schwab (2011);

2. Generation of pictograms from natural language (Vaschalde et al., 2018b; Vaschalde et al., 2018a; Sevens et al., 2017). It is thus a question of associating the pictograms with the corresponding discourse. An automatic pictogram generation tool working with everyday speech is a good way to solve this problem. Such a tool allows people close to the user of an AAC method to speak with their own language without necessarily having to learn how to encode pictogram sentences and without losing time to find pictogram in a communication board. It gives a better access to school for AAC users. If a text-to-speech tool is also used, it becomes possible for students using AAC to communicate with teachers or other students. A social bond can be created, with possibilities of mutual help, which leads to a positive learning environment. Pictogram generation allows to overcome the language barrier between people, and can allow people to join a school or a training course more easily than before.

   The idea is to immerse a person who needs to learn or relearn from scratch in language. Indeed, it is by multiplying the immersion when he or she is confronted with the association of a pictogram with a particular word, a particular concept that the person will manage to associate them in a natural way just as typical children learn words. For example, he must associate the term “pool” with its common meaning with the image of a swimming pool, the term “go” with the pictogram of moving, etc. This can obviously be done through games, or pictogram workbooks, but implementation in a real situation is essential.

3. The resource *Arasaac-WN*

In order to best present the resource, we have opted for the use of specialized tools for the distribution of digital collections. We focused on Omeka (Scheinfeldt, 2008) which is a free digital library management software made available under the GPL license.
The tool is developed by the Center for History and New Media (CHNM) at George Mason University, which also developed the Zotero bibliographic management software. Currently, the Omeka site lists several dozen projects around the world using it. This tool is dedicated to the organization, exposure and online publication of iconographic data, with their metadata, which makes it very easy to publish them on the Web. Modular in design, the tool allows each site to adapt the features offered using plugins and themes. For example, Omeka allows standardized metadata to be combined with the Dublin Core format (semantic web vocabulary used to express data in a model *Resource Description Framework*, RDF) to perform advanced searches within the digital collection. Omeka is an easy to use tool for people who are not computer specialists; while allowing to query it via more specialized tools such as SQL.

3.1. Resource structure

The resource is available in RDF format, so it is possible to use classic Semantic Web tools and query it in SPARQL

```
<dcterms:title><![Darling]></dcterms:title>
<dcterms:alternative><![Chéri]></dcterms:alternative>
<dcterms:description><![A {01465246}]></dcterms:description>
<dcterms:description><![darling#1]></dcterms:description>
<dcterms:description><!(darling%3:00:00:loved:00)]></dcterms:description>
<dcterms:creator><![CDATA[LIG GETALP]></dcterms:creator>
<dcterms:contributor><![Pauline Trial]></dcterms:contributor>
<dcterms:relation><![http://www.arasaac.org/...?Submit=Search..</dcterms:relation>
<dcterms:relation><![http://wordnetweb.princeton.edu/...s=darling...</dcterms:relation>
<dcterms:language><![Anglais]></dcterms:language>
```

Figure 3: Example of resource structure using RDF format

(*SPARQL Protocol and RDF Query Language*). We decided to use the initial fields of the Dublin Core to represent the different pictograms. The main attributes of *Arasaac-WN* are as follows:

1. *Title*: name of the pictogram as it was indicated on *Arasaac*;
2. Description: *WordNet* identifiers that allow the unique identification of a term in the lexical database and thus to remove any ambiguity (Database locations (1), sense numbers (2) and sense key (3));
3. *Creator*: the organization that created the link;
4. *Contributor*: name of the person who added the entry to the database;
5. *Relation*: as indicated in the Dublin Core specifications, this field allows the use of a related resource and can be used to provide links to them. This element is therefore used to highlight the links from which the data are extracted;
6. *Language*: this element is used to specify that the information provided is in English. The *WordNet* being in English, all the words in the two lists in French were included in the collection under their English name on *Arasaac*;
7. *Alternative title*: this element of the Dublin Core qualified allows, as its name indicates, to give an alternative title to the resource. This field was therefore used to fill in the French name of the pictogram and thus facilitate the search in the database.

As of Nov 29, 2019, the resource contains about 800 links, i. e. 800 pictograms of *Arasaac* that have been linked to *WordNet*. Is is freely available at the following URL: `https://github.com/getalp/Arasaac-WN`.

3.2. Examples, difficulties and annotation protocols

For some pictograms (Figure 4), it is difficult to extract a single meaning from the represented content. There are several cases:

- The first example shows two mouse (computer and animal): it is a trivial case. However, in the *Arasaac* database, pictograms are identified with the same filename, except for a different digit. Unfortunately, this filename does not contain any semantic information that would allow them to be distinguished. This simple example illustrates the need to link *Arasaac* with *WordNet*.

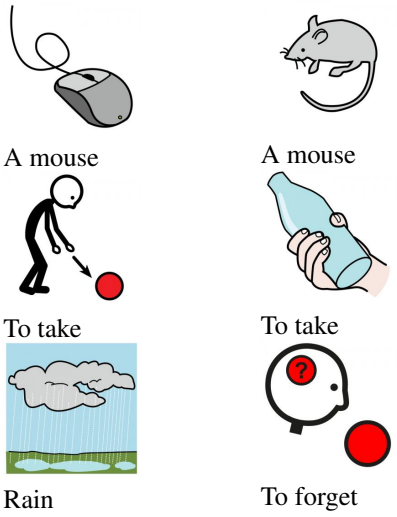

Figure 4: Example of "problematic" pictograms.

- For the "forget" pictogram, for example, the image does not make it possible to establish with certainty the meaning intended by its creator (is it "forget an object" or "have a memory blackout" ?) and make a clear link with *WordNet*. In this type of case, the entries for the same pictogram can be multiplied to correspond to the different possible senses when they are commonly used.
- In *WordNet*, words, and more frequently verbs, have sometimes a lot of senses. The word "take", for example, includes forty-four possible meanings. It is therefore necessary to proceed by elimination, first determining the grammatical class, in the case of taking a verb, then refining the analysis using the indications provided by *WordNet*. In particular, it is necessary to use lexical field information (e.g. cognition, action, communication, etc.) to remove all cases that cannot correspond to the pictogram and then determine the

most relevant meaning using the definitions and example sentences. This operation therefore depends in part on the subjectivity of the contributor and can therefore be subject to misinterpretation.

- *WordNet* uses what is called a "fine-grained" sense inventory, which means that relatively few differences are present between some senses. For the word "rain", for example, *WordNet* proposes the following two definitions: "water falling in drops from vapor condensed in the atmosphere" and "drops of fresh water that fall as precipitation from clouds". When this type of case arises, it is necessary to select the right synset. Other mapping problems may have been identified during the creation of the resource, such as multi-word pictogram names that do not find an equivalent on *WordNet* (ex: to grow larger).

For each new problem encountered, a rule has been recorded. In this way, the annotation of pictograms with *WordNet* should be uniform. All problems and solutions are associated with the resource in its instructions for use. Although we use *WordNet*, an English resource, in the next section we show how this resource can be adapted for other languages. In our case, we take the example of French, which concerns us.

4. How to use our resource for languages other than English ?

Arasaac is language-independent, and most pictograms are labeled with multilingual captions (English, French, Spanish...). In order to assist the process of linking a pictogram to WordNet, we used a simple Word Sense Disambiguation (WSD) system in our native language, which gives the most probable WordNet sense to each word in a given French sentence.

For the construction of this system, we exploited the method proposed by (Hadj Salah et al., 2018), which consists in automatically translating and aligning English sense annotated corpora into another language (in our case French), in order to have French sense annotated corpora. Indeed, manually sense annotated data are rare and almost nonexistent in non-English languages, but they are useful for building a good-quality WSD system.

We then relied on the state-of-the-art WSD system proposed by (Vial et al., 2019), which is implemented in the open-source tool *disambiguate*[2]. In their work, the authors only provide a model able to disambiguate English text, but their tool can be used to train a new disambiguation model in any language, given a set of sense annotated data used for training.

We used the state of the art English to French Machine Translation system of the tool *fairseq*[3] which also provide alignment between the source and the target words, and we translated the two corpora used by (Vial et al., 2019): the SemCor (Miller et al., 1993) and the *WordNet* Gloss Corpus[4]. We then used these two corpora as training data for the WSD system.

Finally, once the WSD system trained, we used its predictions to facilitate the work of the mapping pictogram-to-*WordNet*. For instance, if the annotator wanted to map a pictogram for "take a shower" (in French "prendre une douche"), the sense predicted by the WSD system on the word "take" could help to take a decision.

We present in (Vaschalde et al., 2018b; Vaschalde et al., 2018a; Sevens et al., 2017), some experiments using the French WSD system in order to associate the pictograms with the corresponding discourse.

5. Conclusion and prospects

In this article we have presented a first version of *Arasaac-WN*, a resource that links the *WordNet* semantic inventory to the *Arasaac* database: the largest freely available database of pictograms for people with disabilities. To this date, about 800 pictograms have been linked to access the *WordNet* ecosystem and thus provide access to a range of automatic language processing tools to assist people with disabilities. The realization of this type of base is quite tedious due to the difficulties encountered and the annotation protocol. Originally, we started to develop this database for our research, but we would like to share it with the community so that tools and research on alternative communication via pictograms can be developed quickly and collaboratively. We have already identified some direct applications. In the longer term, this resource will allow the use of up-to-date NLP tools. We are also receptive to the addition of other sets of pictograms to provide the most universal database possible. As we have shown, there are several problems related to equivalencies with *WordNet* senses that do not exist. In future versions, we will propose solutions to overcome these problems. In addition, we propose a method and associated tools that allow to use our database for languages other than English. Our resource is available on Github at the following URL: `https://github.com/getalp/Arasaac-WN`.

Bibliographical References

Beukelman, D. and Mirenda, P. (2017). *Communication alternative et augmentée : Aider les enfants et les adultes avec des difficultés de communication*.

Hadj Salah, M., Vial, L., Blanchon, H., Zrigui, M., Lecouteux, B., and Schwab, D. (2018). Traduction automatique de corpus en anglais annotés en sens pour la désambiguïsation lexicale d'une langue moins bien dotée, l'exemple de l'arabe. In *25e conférence sur le Traitement Automatique des Langues Naturelles*, Rennes, France, May.

Mihalcea, R. and Leong, C. (2008). Toward communicating simple sentences using pictorial representations. *Machine Translation, vol. 22, no. 3, 2008, pp. 153-173*.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244.

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T. (1993). A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, HLT '93,

[2] `https://github.com/getalp/disambiguate`

[3] `https://github.com/pytorch/fairseq`

[4] `https://wordnetcode.princeton.edu/glosstag-files/glosstag.shtml`

pages 303–308, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nègre, E. C. (2017). *Communiquer autrement: Accompagner les personnes avec des troubles de la parole ou du langage*.

Petitpierre, G. and Barisnikov, K. (1994). *Le Vygotskij que nous (ne) connaissons (pas). Les principaux travaux de Vygotskij et la chronologie de leur composition*. Delachaux et Niestlé.

Scheinfeldt, T. (2008). Omeka: Open source web publishing for research, collections and exhibitions. *Open Source Business Resource*, 12/2008.

Schwab, D., Fejza, A., Vial, L., and Robert, Y. (2018). The GazePlay Project: Open and Free Eye-trackers Games and a Community for People with Multiple Disabilities. In *ICCHP 2018 - 16th International Conference on Computers Helping People with Special Needs*, volume 10896 of *LNCS*, pages 254–261, Linz, Austria, July. Springer.

Schwab, D., Trial, P., Vaschalde, C., Vial, L., and Lecouteux, B. (2019). Apporter des connaissances sémantiques à un jeu de pictogrammes destiné à des personnes en situation de handicap : un ensemble de liens entre wordnet et arasaac, arasaac-wn. *TALN démonstration*.

Sevens, L., Vandeghinste, V., Schuurman, I., and Eynde, F. V. (2017). Simplified Text-to-Pictograph Translation for People with Intellectual Disabilities. *NLDB*.

Vandeghinste, V. and Schuurman, I. (2014). Linking Pictographs to Synsets: Sclera2Cornetto. *LREC*.

Vaschalde, C., Lecouteux, B., and Schwab, D. (2018a). Génération de pictogrammes à partir de la parole spontanée pour la mise en place d'une communication médiée. *50 ans de linguistique sur corpus oraux : Apports à l'étude de la variation*.

Vaschalde, C., Trial, P., Esperança-Rodier, E., Schwab, D., and Lecouteux, B. (2018b). Automatic pictogram generation from speech to help the implementation of a mediated communication. *Conference on Barrier-free Communication*.

Vial, L., Lecouteux, B., and Schwab, D. (2019). Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. In *Proceedings of the 10th Global Wordnet Conference*, Wroclaw, Poland.

Zock, M. and Schwab, D. (2011). Storage does not Guarantee Access: The Problem of Organizing and Accessing Words in a Speaker's Lexicon. *Journal of Cognitive Science*, 12:233–258. (Impact-F 3.52 estim. in 2012).