

Corpus - Texts in Pictograms - Adults with intellectual disabilities

This corpus was created as part of Magali Norré's doctoral thesis (2025) at the Centre de Traitement Automatique du Langage (CENTAL), Université catholique de Louvain (UCLouvain) and the Département de Traitement Informatique Multilingue (TIM), Université de Genève (UNIGE): translation Automatic from French into pictograms to improve communication between doctors and patients with intellectual disabilities in hospital settings.

Under the supervision of Prof. Thomas François (CENTAL, UCLouvain), Prof. Pierrette Bouillon (TIM, UNIGE) and Dr. Vincent Vandeghinste (Instituut voor de Nederlandse Taal & KULeuven).

It consists of three folders containing data for several chapters.


- The **text-to-picto** file

It contains the French pictogram corpora used for the evaluation automatic of the systemText-to-Picto , machine translation as as wellthe tables for the French version of Text-to-Picto (Arasaac , pictogram dictionarySclera adapted pictogram dictionary from Vandeghinste and Sevens, Beta pictogram dictionary adapted from Vandeghinste and Sevens, WOLF, WoNeF coverage, WoNeF fscore, WoNeF precision, ReSyf, medical paraphrase), presented in Chapters 3 and 4 of the thesis.dictionary

- French pictogram corpus

This subfolder contains the source corpora (src/Source) and reference corpora (ref/Ref) translated manually into Arasaac pictograms by us (corporaEmail, Medical), by a Belgian speech therapist (Medical corpora), by Vaschalde (2018) and post-edited by us (Book corpora). They are sometimes divided into development corpus (Dev, 40%) and test corpus (Tst, 60%). names Pictogram correspond to numerical identifiers in our Arasaac . They provide access to the image's URL link on the Arasaac server. Example:tables

Medical corpus produced by the speech therapist

100a_Tst_Source_Arasaac.txt	100a_Tst_Ref_Arasaac.txt	
<seg>Do you have problems with your sense of smell?</seg>	<Your nose?	2608 2887 3418 ¹

¹ https://static.arasaac.org/pictograms/2608/2608_500.png https://static.arasaac.org/pictograms/2887/2887_500.png https://static.arasaac.org/pictograms/3418/3418_500.png

Some pictogram names in the corpora are numbered (with "_n"); these are the ones we've renamed in our database to distinguish homonyms (e.g. "dogs" of various breeds, "creams" for different uses, etc.).

- ❖ Email corpus (v1, see chapter 3): 140 sentences (56 Dev+ 84 Tst)| **TXT**
 - ❖ Corpus Book (v1, see chapter 3): 254 sentences| **TXT**
 - ❖ Corpus Médical (v1, see chapter 3): 260 sentences| **TXT HTML**
 - ❖ Corpus Médical translated by a Belgian speech therapist (v2, see chapter 4): 100 sentences
+ Adapted Medical Corpus (v2, see chapter 4): 150 sentences (60 Dev+ 90 Tst)| **TXT**
- Reference: Magali Norré, Vincent Vandeghinste, Pierrette Bouillon and Thomas François, Extending a Text-to-Pictograph System to French and to Arasaac. *In Proceedings of the 13th International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1050-1059. INCOMA Ltd, 2021.

→ Bibtex:

```
@inproceedings{norre2021,
  title={Extending a {T}ext-to-{P}ictograph {S}ystem to {F}rench and to {A}rasaac},
  author={Norré, Magali and Vandeghinste, Vincent and Bouillon, Pierrette and François, Thomas},
  year={2021},
  booktitle={Proceedings of the 13th International Conference on Recent Advances in Natural Language Processing (RANLP 2021)},
  publisher={INCOMA Ltd},
  pages={1050--1059}
}
```

- Tables from the French version of Text-to-

Picto Tables dump PostgreSQL (see chapters 3 and 4) TBL|

Links and references to resources used/adapted, to be quoted if used :

- Arasaac pictograms: <https://arasaac.org>
- Sclera pictograms: <https://www.sclera.be>
- Beta pictograms: <https://www.betasymbols.com>
- Wordnet Libre du Français (WOLF, version 1.0b4) :
https://almanach.inria.fr/software_and_resources/WOLF-en.html (WordNet default in Text-to-Picto) -> (Sagot and Fišer, 2008)
- French WordNet (WoNeF coverage/fscore/precision): <https://wonef.pradet.me> -> (Pradet et al., 2014)
- ReSyf: <https://cental.uclouvain.be/resyf> -> (Billami et al., 2018; François et al., 2016)
- Medical paraphrases: <http://natalia.grabar.free.fr/resources.php> -> (Koptient and Grabar, 2020)



- The **simplification_large_language_models** folder

It contains the automatic simplifications generated by 8 large generative and language models the corresponding automatic translations into Arasaac pictograms generated by the French Text-to-Picto system, cf. chapter 4 of the thesis.

- Completion date for experiments with large generative language : 29/03/2023
- Location of experiments: LMSys Chatbot Arena platform (<https://lmarena.ai>)
- Major generative language models used: gemini-pro-dev-api, claude-2.1, command-r, gpt-3.5-turbo-0613, llama-2-70b-chat, mistral-medium, qwen1.5-72b-chat, vicuna-33b
- Prompt: "Simplify the sentence while keeping the same meaning. Sentence to simplify: [sentence]".
- Parameters of large generative language models: temperature 0, top P 1, max output tokens 1024
- Text-to-Picto parameters: oov 3, wrongnum 4, nonum 9, hyper 15, xpos 8, anto 10, penal 9, dict 2
- ❖ Metadata for the 8 major generative language models used according to LMSys (data as of 29/03/2023) | **TSV** (#1)
- ❖ Results: 100 simplifications by large generative language model (#800) with 100 source sentences and 100 prompts | **TSV** (#1)
- ❖ Results: 100 automatic translations into Arasaac pictograms generated by Text-to-Picto from the simplifications (#800) + 100 translations of source without automatic simplifications | **HTML** (#9)
- ❖ Results: annotations of Text-to-Picto machine translations (1 annotator) vs. machine translation without Text-to-Picto , simplification where 0: worse translation; 1: equivalent translation; 2: better translation | **TSV** (#1)

- The **user_tests** folder

It contains all the data from the two user , studies including the corpora collected from adults with intellectual disabilities on the comprehension of medical translated into Arasaac pictograms, presented in Chapter 6 of the thesis.

- user_test1

- Test completion date: June 2023
- Test location: Fribourg, Switzerland
- Test method: semi-structured individual interview (#9)
- Participants: 1 facilitator, 9 adults with intellectual disabilities
- ❖ Documents and questionnaires used in tests | **PDF**

- ❖ Transcriptions from 9 recordings audio (not supplied), manual and/or automatically from Whisper, then post-edited (by 2 linguistics researchers) | **TXT**
- ❖ Participant information (anonymized) | **TSV**
- ❖ Results for pictogram comprehension and sentence comprehension two annotators (A1 and A2) | **TSV**
- ❖ R code used for descriptive/inferential statistical analysis (models generalized linear mixed-effects, paired, Student's t-tests inter-annotator agreement, etc.) | **R TXT**

→ Reference: Magali Norré, Trang Tran Hanh Pham, Pierrette Bouillon, Vincent Vandeghinste and Thomas François, Evaluation of pictogram translation doctor-patient communication by adults with intellectual disabilities *In Proceedings of 13th the IFRATH conference on assistive technologies Handicap 2024: Des solutions personnalisées pour des besoins spécifiques*, pages 181-186, Paris, France, 2024. IFRATH.

→ Bibtext:

```
@inproceedings{norre2024,
  title={Évaluation de la traduction en pictogrammes pour la communication par adultes avec déficience intellectuelle},
  author={Norré, Magali and Pham, Trang Tran Hanh and Bouillon, Pierrette and Vandeghinste, Vincent and François, Thomas},
  year={2024},
  booktitle={Acts of the 13th IFRATH conference on assistive technologies Handicap 2024: Customized solutions for specific needs}, publisher={IFRATH},
  address={Paris, France},
  pages={181--186}
}
```

- user_test2

- Test completion date: June 2024
- Test location: Fribourg, Switzerland
- Test method: semi-structured individual interview (#10)
- Participants: 1 facilitator, 10 adults with intellectual disabilities

- ❖ Documents and questionnaires used in tests | **PDF**
- ❖ Transcriptions of the 10 audio recordings (not included), automatically from Whisper then post-edited (by 1 linguistics researcher) | **TXT**
- ❖ Participant information (anonymized) | **TSV**
- ❖ Results for pictogram comprehension and sentence comprehension two annotators (A1 and A2) | **TSV**

- (Le dossier **desambiguation_lexicale**)

The data (corpus evaluation and Python) code for chapter 5 of the thesis on lexical disambiguation to improve machine translation from French into Arasaac pictograms are not to be found in this corpus, but on Vincent Vandeghinste's Text-to-Picto Github <https://github.com/VincentCCL/Picto>

- Reference: Magali Norré, Rémi Cardon, Vincent Vandeghinste and Thomas François, Word Sense Disambiguation for Automatic Translation of Medical Dialogues into Pictographs. *In Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing (RANLP 2023)*, pages 803-812, Varna, Bulgaria, 2023. INCOMA Ltd.
- Bibtex:


```
@inproceedings{norre2023,
  title={Word {S}ense {D}isambiguation      for {A}utomatic      {T}ranslation of {M}edical {D}ialogues into {P}ictographs},
  author={Norré, Magali and Cardon, Rémi and Vandeghinste, Vincent and François, Thomas},
  year={2023},
  booktitle={Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing (RANLP 2023)},
  publisher={INCOMA Ltd},
  address={Varna, Bulgaria},
  pages={803--812}
}
```