

Corpus - Textes en Pictogrammes - Adultes avec une Déficience Intellectuelle

Ce corpus a été créé dans le cadre de la thèse de doctorat (2025) de Magali Norré au Centre de Traitement Automatique du Langage (CENTAL), Université catholique de Louvain (UCLouvain) et au Département de Traitement Informatique Multilingue (TIM), Université de Genève (UNIGE) : Traduction automatique du français vers les pictogrammes pour améliorer la communication entre médecins et patients avec une déficience intellectuelle en milieu hospitalier.

Sous la direction du Prof. Thomas François (CENTAL, UCLouvain), de la Prof. Pierrette Bouillon (TIM, UNIGE) et du Dr. Vincent Vandeghinste (Instituut voor de Nederlandse Taal & KULeuven).

Il se compose de trois dossiers contenant les données de plusieurs chapitres.


- Le dossier **text-to-picto**

Il contient les corpus français en pictogrammes utilisés pour l'évaluation automatique du système de traduction automatique Text-to-Picto, ainsi que les tables de la version française de Text-to-Picto (dictionnaire de pictogrammes Arasaac, dictionnaire de pictogrammes Sclera adapté de Vandeghinste et Sevens, dictionnaire de pictogrammes Beta adapté de Vandeghinste et Sevens, WOLF, WoNeF coverage, WoNeF fscore, WoNeF precision, ReSyf, dictionnaire de paraphrases médicales), présentés dans les chapitres 3 et 4 de la thèse.

- Corpus français en pictogrammes

Ce sous-dossier contient les corpus sources (src/Source) et les corpus de référence (ref/Ref) traduits manuellement en pictogrammes Arasaac par nous (Corpus Email, Médical), par une logopède belge (Corpus Médical), par Vaschalde (2018) et post-édité par nous (Corpus Livre). Ils sont parfois divisés en corpus de développement (Dev, 40 %) et corpus de test (Tst, 60 %). Les noms des pictogrammes correspondent aux identifiants numériques dans nos tables Arasaac. Ils permettent d'accéder au lien URL de l'image sur le serveur d'Arasaac. Exemple :

Corpus Médical réalisé par la logopède

100a_Tst_Source_Arasaac.txt	100a_Tst_Ref_Arasaac.txt	
<seg>avez-vous des troubles de l'odorat ?</seg>	<seg>vous nez ?</seg>	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

Certains noms de pictogrammes dans les corpus sont numérotés (avec “_n”), ce sont ceux que nous avons renommé dans notre base de données pour distinguer les homonymes (par exemple, des “chiens” de diverses races, des “crèmes” pour différents usages, etc.).

- ❖ Corpus Email (v1, cf. chapitre 3) : 140 phrases (56 Dev + 84 Tst) | **TXT**
 - ❖ Corpus Livre (v1, cf. chapitre 3) : 254 phrases | **TXT**
 - ❖ Corpus Médical (v1, cf. chapitre 3) : 260 phrases | **TXT HTML**
 - ❖ Corpus Médical traduit par une logopède belge (v2, cf. chapitre 4) : 100 phrases
+ Corpus Médical adapté (v2, cf. chapitre 4) : 150 phrases (60 Dev + 90 Tst) | **TXT**
- Référence : Magali Norré, Vincent Vandeghinste, Pierrette Bouillon et 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 de la version française de Text-to-Picto

Tables dump PostgreSQL (cf. chapitres 3 et 4) | **TBL**

Liens et références vers les ressources utilisées/adaptées, à citer en cas d'utilisation :

- Pictogrammes Arasaac : <https://arasaac.org>
- Pictogrammes Sclera : <https://www.sclera.be>
- Pictogrammes Beta : <https://www.betasymbols.com>
- Wordnet Libre du Français (WOLF, version 1.0b4) :  
[https://almanach.inria.fr/software\\_and\\_resources/WOLF-en.html](https://almanach.inria.fr/software_and_resources/WOLF-en.html) (WordNet par défaut dans Text-to-Picto) -> (Sagot et Fišer, 2008)
- WordNet du Français (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)
- Paraphrases médicales : <http://natalia.grabar.free.fr/resources.php> -> (Koptient et Grabar, 2020)



- Le dossier **simplification\_large\_language\_models**

Il contient les simplifications automatiques générées par 8 grands modèles de langue génératifs et les traductions automatiques en pictogrammes Arasaac correspondantes générées par le système Text-to-Picto français, cf. chapitre 4 de la thèse.

- Date de réalisation des expériences réalisées avec les grands modèles de langue génératifs : 29/03/2023
- Lieu de réalisation des expériences : Plateforme Chatbot Arena du LMSys (<https://lmarena.ai>)
- Grands modèles de langue génératifs utilisés : 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 : "Simplifie la phrase en gardant le même sens. Phrase à simplifier : [phrase]"
- Paramètres des grands modèles de langue génératifs : temperature 0, top P 1, max output tokens 1024
- Paramètres de Text-to-Picto : oov 3, wrongnum 4, nonum 9, hyper 15, xpos 8, anto 10, penal 9, dict 2
  
- ❖ Métadonnées des 8 grands modèles de langue génératifs utilisés selon le leaderboard du LMSys (données du 29/03/2023) | **TSV** (#1)
- ❖ Résultats : 100 simplifications par grand modèle de langue génératif (#800) avec les 100 phrases sources et les 100 prompts | **TSV** (#1)
- ❖ Résultats : 100 traductions automatiques en pictogrammes Arasaac générées par Text-to-Picto à partir des simplifications (#800) + 100 traductions des phrases sources sans simplification automatique | **HTML** (#9)
- ❖ Résultats : annotations des traductions automatiques de Text-to-Picto (1 annotateur) par rapport à la traduction automatique sans simplification de Text-to-Picto, où 0 : traduction pire ; 1 : traduction équivalente ; 2 : traduction meilleure | **TSV** (#1)

- Le dossier **user\_tests**

Il contient toutes les données des deux études utilisateurs, incluant les corpus récoltés auprès d'adultes avec une déficience intellectuelle sur la compréhension de phrases médicales traduites en pictogrammes Arasaac, présentés dans le chapitre 6 de la thèse.

- user\_test1

- Date de réalisation des tests : juin 2023
- Lieu de réalisation des tests : Fribourg, Suisse
- Méthode de test : entretien individuel semi-structuré (#9)
- Participants : 1 facilitatrice, 9 adultes avec une déficience intellectuelle
  
- ❖ Documents et questionnaires utilisés dans le cadre des tests | **PDF**

- ❖ Retranscriptions des 9 enregistrements audio (non fournis), manuelles et/ou automatiques depuis Whisper puis post-éditées (par 2 chercheurs en linguistique) | **TXT**
- ❖ Informations sur les participants (anonymisés) | **TSV**
- ❖ Résultats de la compréhension des pictogrammes et de la compréhension des phrases, deux annotateurs (A1 et A2) | **TSV**
- ❖ Code R utilisé pour l'analyse statistique descriptive/inférentielle (modèles linéaires généralisés à effets mixtes, tests de Student pairé, accords inter-annotateurs, etc.) | **R TXT**

→ Référence : Magali Norré, Trang Tran Hanh Pham, Pierrette Bouillon, Vincent Vandeghinste et Thomas François, Évaluation de la traduction en pictogrammes pour la communication médecin-patient par des adultes avec une déficience intellectuelle. *In Actes de la 13e conférence de l'IFRATH sur les technologies d'assistance. Handicap 2024 : Des solutions personnalisées pour des besoins spécifiques*, pages 181–186, Paris, France, 2024. IFRATH.

→ Bibtex :

```
@inproceedings{norre2024,
 title={{É}valuation de la traduction en pictogrammes pour la communication
médecin-patient par des adultes avec une déficience intellectuelle},
 author={Norré, Magali and Ph\{a}m, Trang Tr\{a}n H\{a}nh and Bouillon, Pierrette and
Vandeghinste, Vincent and François, Thomas},
 year={2024},
 booktitle={Actes de la 13e conférence de l'IFRATH sur les technologies d'assistance.
{H}andicap 2024~: {D}es solutions personnalisées pour des besoins spécifiques},
 publisher={IFRATH},
 address={Paris, France},
 pages={181--186}
}
```

- user\_test2

- Date de réalisation des tests : juin 2024
- Lieu de réalisation des tests : Fribourg, Suisse
- Méthode de test : entretien individuel semi-structuré (#10)
- Participants : 1 facilitatrice, 10 adultes avec une déficience intellectuelle

- ❖ Documents et questionnaires utilisés dans le cadre des tests | **PDF**
- ❖ Retranscriptions des 10 enregistrements audio (non fournis), automatiques depuis Whisper puis post-éditées (par 1 chercheur en linguistique) | **TXT**
- ❖ Informations sur les participants (anonymisés) | **TSV**
- ❖ Résultats de la compréhension des pictogrammes et de la compréhension des phrases, deux annotateurs (A1 et A2) | **TSV**

- (Le dossier **desambiguisation\_lexicale**)

Les données (corpus d'évaluation et code Python) du chapitre 5 de la thèse sur la désambiguisation lexicale pour améliorer la traduction automatique du français vers les pictogrammes Arasaac ne se trouvent pas dans ce corpus, mais sur le dépôt Github Text-to-Picto de Vincent Vandeghinste : <https://github.com/VincentCCL/Picto>

→ Référence : Magali Norré, Rémi Cardon, Vincent Vandeghinste et 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={{W}ord {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}
}
```