

ORTOLANG

Etienne Petitjean - ATILF - CNRS / Université de Lorraine

Cyril Pestel - ATILF - CNRS / Université de Lorraine

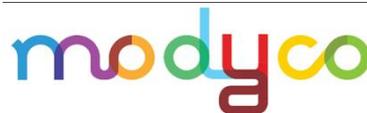
Christophe Parisse – MODYCO – CNRS/Université Paris Nanterre

Historique

- Projet monté en **2011**
- En réponse à l'appel d'offre Equipex dans le cadre des **PIA2**
- Un équipement d'excellence de **mutualisation** de ressources et d'outils sur la langue et son traitement informatique
- Budget total de 2,6M € sur 7 ans
- Dès le départ vision nationale et internationale
- Volonté marquée d'axer le projet sur les **données ouvertes**



un consortium réunissant des
compétences complémentaires



Objectifs

- **Mutualisation** de ressources et d'outils pour des travaux de recherche :
 - les ressources et les outils restent la propriété des laboratoires.
- Sécurité des données :
 - **Authentification** et **droits d'accès**
- **Valorisation** des données de la recherche :
 - FAIR
- **Science Ouverte** :
 - Encourager au maximum les données ouvertes
 - « **Aussi ouvert que possible, aussi fermé que nécessaire** »

Statistiques

USERS



3,587

FILES



1,357,378

DATA



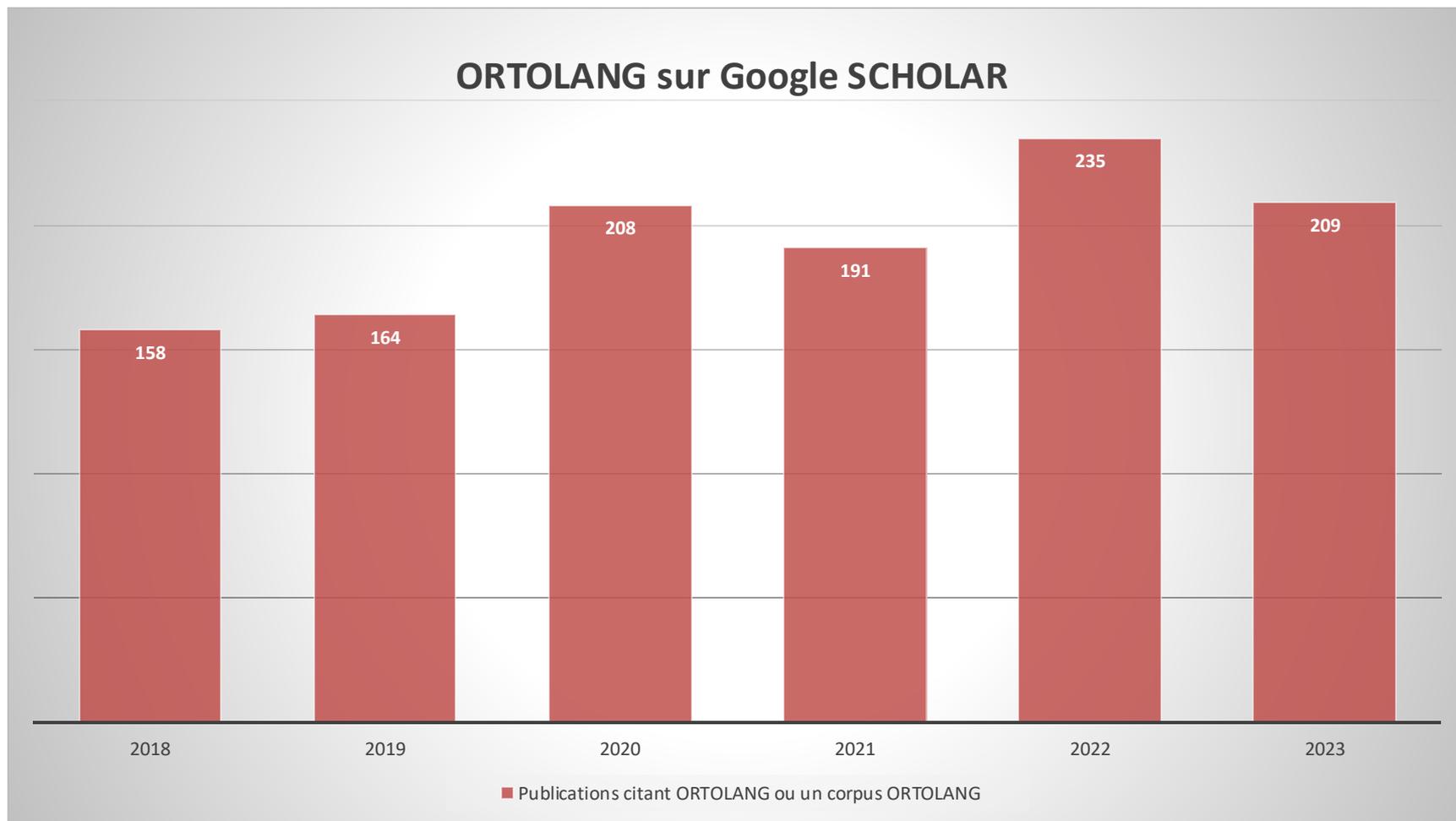
18.8 TB

RESOURCES



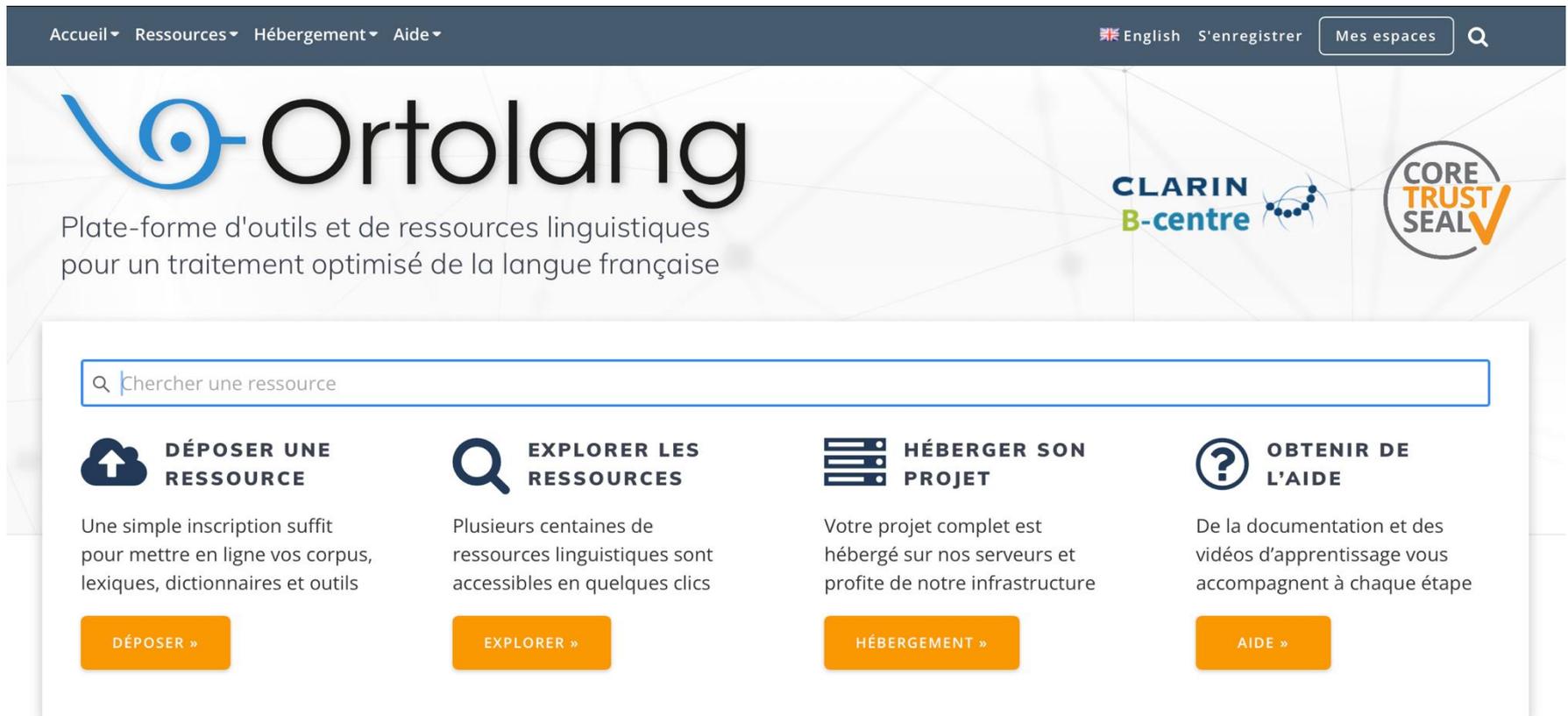
727

Visibilité (Google Scholar)



Utilisation d'ORTOLANG

- Accès par l'adresse **ortolang.fr**



The screenshot shows the homepage of the Ortolang website. At the top, there is a dark navigation bar with links for 'Accueil', 'Ressources', 'Hébergement', and 'Aide'. On the right side of the navigation bar, there are links for 'English', 'S'enregistrer', and 'Mes espaces' with a search icon. Below the navigation bar, the main header features the Ortolang logo, a tagline 'Plate-forme d'outils et de ressources linguistiques pour un traitement optimisé de la langue française', and logos for 'CLARIN B-centre' and 'CORE TRUST SEAL'. A search bar is positioned below the header with the placeholder text 'Chercher une ressource'. The main content area is divided into four columns, each with an icon, a title, a description, and an orange button with a right-pointing arrow. The columns are: 1. 'DÉPOSER UNE RESSOURCE' with a cloud upload icon, describing the ease of uploading corpora, lexicons, and tools. 2. 'EXPLORER LES RESSOURCES' with a magnifying glass icon, stating that hundreds of linguistic resources are accessible with a few clicks. 3. 'HÉBERGER SON PROJET' with a server rack icon, explaining that complete projects are hosted on their infrastructure. 4. 'OBTENIR DE L'AIDE' with a question mark icon, mentioning that documentation and learning videos accompany each step.

Accueil ▾ Ressources ▾ Hébergement ▾ Aide ▾

English S'enregistrer Mes espaces 🔍

Ortolang

Plate-forme d'outils et de ressources linguistiques pour un traitement optimisé de la langue française

CLARIN B-centre CORE TRUST SEAL

🔍 Chercher une ressource

- DÉPOSER UNE RESSOURCE**
Une simple inscription suffit pour mettre en ligne vos corpus, lexiques, dictionnaires et outils
[DÉPOSER »](#)
- EXPLORER LES RESSOURCES**
Plusieurs centaines de ressources linguistiques sont accessibles en quelques clics
[EXPLORER »](#)
- HÉBERGER SON PROJET**
Votre projet complet est hébergé sur nos serveurs et profite de notre infrastructure
[HÉBERGEMENT »](#)
- OBTENIR DE L'AIDE**
De la documentation et des vidéos d'apprentissage vous accompagnent à chaque étape
[AIDE »](#)

Explorer



Explorer

Rechercher une ressource :

Types de ressources disponibles sur la plateforme ORTOLANG :



CORPUS



LEXIQUES



TERMINOLOGIES



OUTILS

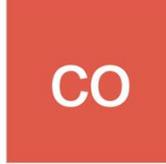
Les corpus

ORTOLANG Catalogue

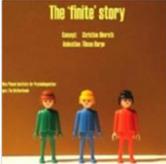
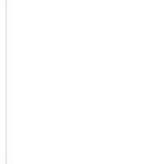
Aide Langue Se connecter S'inscrire

Rechercher un corpus

Corpus écrits (77 ressources)

 <p>Corpus lexical du farsi (thèse Jaferian 2024)</p> <p>2 oct. 2024</p>	 <p>Climate discourses</p> <p>9 sept. 2024</p>	 <p>Dictionnaire Étymologique Roman (DÉRom)</p> <p>22 août 2024</p>	 <p>corpus_textuel_these_soni</p> <p>24 juil. 2024</p>	 <p>Corpus Philéduc</p> <p>23 juil. 2024</p>	 <p>EFG_WikiCorpus -- discussions en coulisse de Wikipedia (anglais, français, allemand)</p> <p>21 juil. 2024</p>	 <p>MoNoPoli - Mots construits sur Noms propres de personnalités Politiques</p> <p>4 juil. 2024</p>	 <p>partie d'un texte exemple de Y utilisé par X</p> <p>Real, @.Accompiliter p=ac</p> <p>Il citait toujours un exemple ou une phrase</p> <p>BEL-RL-fr</p> <p>19 juin 2024</p>
---	---	--	---	--	--	--	--

Corpus oraux (193 ressources)

 <p>The CLES corpus of spontaneous L2 English</p> <p>30 juil. 2024</p>	 <p>Corpus Finite Story: structure informationnelle et organisation du discours dans différentes langues et @p.L2 2024</p> <p>27 mai 2024</p>	 <p>Interactions didactiques dans les disciplines scolaires</p> <p>1 mai 2024</p>	 <p>Corpus GEACC - Lab- FEACC</p> <p>10 avr. 2024</p>	 <p>TCOF : Traitement de Corpus Oraux en Français</p> <p>25 mars 2024</p>	 <p>TAM-MOUV</p> <p>13 mars 2024</p>	 <p>Corpus ANCOR Centre Version TEI</p>
---	--	--	--	---	---	--

Corpus multimodaux (53 ressources)

 <p>VintAge - Videos to study Interaction in AGeing</p> <p>16 oct. 2024</p>	 <p>CleLIPC</p> <p>31 juil. 2024</p>	 <p>Dinlang_Doing_Atting</p> <p>1 juil. 2024</p>	 <p>40 brèves</p> <p>28 mai 2024</p>	 <p>Mediapi-RGB</p> <p>18 avr. 2024</p>	 <p>CIENSFO (Corpus d'Interrogatives Enchâssées Non-Standards du Français Oral)</p> <p>10 avr. 2024</p>	 <p>Propicto</p> <p>9 avr. 2024</p>	 <p>SMYLE</p> <p>15 août 2023</p>
--	---	---	---	---	--	--	--

Recherche pour « Modyco »

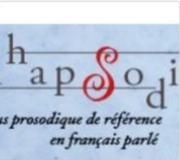
MoDyCo | Modèles, Dynamiques, Corpus - UMR 7114

📍 Nanterre, FR

🌐 <http://www.modyco.fr/fr/>



Liste des productions du laboratoire

 teicorpo 19 juil. 2016	 PFC - Phonologie du Français Contemporain 2 mars 2017	 NARRACOM 3 avr. 2018	 TREEBANK RHAPSODIE 26 juin 2018	 CEFC-GOLD 24 oct. 2019	 VerNom 27 mai 2020	 Shrug 24 juin 2020	 Embodying Language Complexity: Co-speech gestures between 3 and 4 13 avr. 2021
 Signes en famille 18 oct. 2021	 Dictionnaire Informatisé des Mots d'Affect 19 janv. 2023	 TREMolo_Tweets_Corpus 21 mars 2023	 DinLang - GESPIN 2023 24 mars 2023	 Colaje 24 oct. 2023	 Dinlang_Doing_Attending 1 juil. 2024	 VintAge - Videos to study Interaction in AGEing 16 oct. 2024	

Recherche pour « Enfant »

ORTOLANG Catalogue Aide Lingue Se connecter

enfant Q ↑↓

- ALIFE (Acquisition de la Liaison et Interactions Parents **Enfants**)
- Impact de l'amorçage rythmique sur la production de la parole chez l'**enfant** sourd prélingual

Type de ressource

- Corpus
- Lexique
- Terminologie
- Outil
- Projet intégré

Droits d'utilisation

- Libre sans utilisation commerciale
- Négociation nécessaire
- Libre

Producteur

- Langues, Textes, Arts et Cultures du Monde Anglophone - EA 4398 (PRISMES, Paris FR)
- Modèles, Dynamiques, Corpus - UMR 7114 (MoDyCo, Nanterre FR)
- Structures Formelles du Langage - UMR 7023 (SFL, Paris FR)
- Laboratoire parole et langage - UMR 7309 (LPL, Aix-en-Provence FR)
- Département de linguistique et phonétique générales, Université d'Aix-Marseille (Aix-en-Provence FR)
- Département de sciences du langage, Université d'Aix-Marseille (Aix-en-Provence FR)
- Savoirs, textes et langage - UMR 8163 (STL, Lille FR)
- Communication langagière et interaction personne-système (CLIPS-IMAG, Grenoble FR)

Format



Corpus Philéduc

Corpus Écrit Libre sans utilisation commerciale



Dinlang_Doing_Atending

Extraits vidéos du corpus Dinlang, tiré du projet Dinlang: <https://dinlang.ortolang.fr/2022/03/02/le-projet-dinlang-acc...>

Corpus Multimodal Libre sans utilisation commerciale



Corpus Finite Story: structure informationnelle et organisation du discours dans différentes langues et en L2

Corpus Oral Libre sans utilisation commerciale



TCOF : Traitement de Corpus Oraux en Français

Le projet « Traitement de Corpus Oraux en Français » (TCOF) est né de la volonté de conserver des corpus oraux collectés dans les années 80-90 à des fins de recherches personnelles. L'équipe consti...

Corpus Oral Libre sans utilisation commerciale



Colaje

L'objectif du Projet ANR CoLaJE est de reconstituer l'émergence et le développement de la communication langagière chez le jeune enfant, avec une approche pluridisciplinaire et multimodale. L'analy...

Corpus Oral Libre sans utilisation commerciale



CUP-Morgenstern

Extraits utilisés pour le chapitre du livre *Gesture and first language development: the multimodal child* dans le volume édité par Alan Cienki (Cambridge University Press).

Déposer

- S'identifier
 - Créer un utilisateur
 - Utiliser les identifiants de Renater
 - Identification comme Enseignement et Recherche (ESR)
- Créer et remplir un espace
- Publier cet espace
 - Un dépôt pérenne et un versionnage
- Revenir modifier et republier cet espace
 - Les anciennes versions restent accessibles

S'identifier

- 4 niveaux de protection des données
- Libre (pas d'identification) – données totalement ouvertes
- Utilisateurs identifiés → doivent avoir un compte sur ORTOLANG
 - Ouvert à toute personne
- Utilisateurs ESR → doivent appartenir à une institution d'enseignement supérieur ou de recherche (y compris doctorants)
- Privé → réservé aux données en cours de préparation ou aux données privées
 - L'utilisation de données privées doit être justifiée

Déposer et décrire des données

- Description complète dans <https://www.ortolang.fr/fr/deposer/>
- Créer un espace de travail
 - Le nommer (nom du corpus)
 - Décrire les métadonnées du corpus
 - Déposer les données
 - Glisser déposer, téléchargement (données isolées ou fichiers zip) et dépôt FTP pour les très grosses données
 - Pas de limite de format mais les données ouvertes sont recommandées
 - Donner les droits d'utilisation

Curation des données (modération)

- Une vérification des données est faite avant publication
 - Vérifier que les métadonnées sont correctes
 - Propriétaire et déposants
 - Laboratoires
 - Licence
 - Formats (formats ouverts demandés)
 - Se conformer à la charte d'ORTOLANG
 - Données ouvertes dans des formats connus et ouverts
 - Spécifier explicitement les limitations éventuelles
 - Données sous quarantaine, droits limités, condition d'utilisation

Accès aux données

- Données moissonnables par les moteurs de recherche comme Isidore, Clarin (VLO, Content search)
- Seules les pages principales sont accessibles par Google
- Téléchargement possible de tout un corpus, d'une partie de corpus, ou d'un fichier particulier.
- Tous les documents ont une adresse pérenne
 - Les fichiers peuvent être utilisés pour être accédés depuis un site web
 - https://hdl.handle.net/11403/cefc-orfeo/v1/oral/fleuron/accueil_etudiants_nancy2_partie2.orfeo
 - https://hdl.handle.net/11403/cefc-orfeo/v1/oral/fleuron/accueil_etudiants_nancy2_partie2.wav

Déposez vos données !!!

Évolutions récentes

- Certification [CoreTrustSeal](#)
- Certification [Centre-B CLARIN](#)
- Plateforme recherche.data.gouv.fr

CoreTrustSeal

- Certification **internationale** pour les entrepôts de données numériques
- « Self assessment »
- Formulaire de **16 questions**:
 - Questions **institutionnelles**
 - Positionnement par rapport à **une communauté**
 - Qualité & sécurité des **données**
 - Qualité & sécurité des **métadonnées**
 - Documentation des procédures
 - Licences
 - Etc.

CoreTrustSeal (2)

- Plusieurs **semaines** de travail:
 - Rédaction en anglais (ATILF, INIST, Modyco)
 - Réflexion sur notre **fonctionnement** et nos **procédures**
 - Traduction de toute la documentation en anglais
 - Création d'un site **multilingue** de documentation
- Certification obtenue en aout 2023 pour 3 ans



CLARIN

- Infrastructure de recherche européenne sur les données langagières
- Participation de 24 pays
- Réseau de centres « nœuds »
 - B, C, K
- Notre objectif:
 - Devenir le **premier** Centre B français (22 au total)
 - Renforcer la **visibilité internationale** d'ORTOLANG

CLARIN (2)

- Certification technique:
 - CoreTrustSeal
 - Intégrer la fédération d'identité EduGain
 - Gérer des identifiants pérennes (PIDs)
 - Etc.
- Certification obtenue en mars 2024

CLARIN
B-centre

