Les VOCAUX

Julie GLIKMAN, Nicolas MAZZIOTTA, Camille FAUTH, Christophe BENZITOUN

https://www.atilf.fr/ressources/corpus-les-vocaux/

Corpus Les Vocaux – version 0.0.2 – Février 2025

Document de présentation de cette version

Table des matières

1 Présentation du corpus

Les sms vocaux inclus dans le corpus Les Vocaux proviennent de deux campagnes de recueil, menées en 2021 et 2022. Les campagnes sont identifiées par l'année de recueil.

1.1 Campagne 2021

La première campagne (dite « martyr ») a eu lieu en 2021 auprès de sept volontaires. Les participantes ont été reçues en entretien (en présentiel ou à distance) par Julie Glikman. Durant cet entretien, elles ont été interrogées sur leurs pratiques en lien avec les sms vocaux. Des questions permettant d'établir un jeu minimal de métadonnées (sexe – en l'occurrence les sept volontaires sont des femmes, lieu de naissance, lieu de vie actuelle) ont également été posées. Durant l'entretien, il leur a été demandé de raconter un évènement sans intervention de l'enquêteur. Ce passage de l'entretien a été ensuite isolé et conservé en tant qu'échantillon de parole en vue de comparaison ultérieure avec leur sms vocaux. Ces passages sont inclus dans le corpus sous l'identifiant « IDlocuteur_extrait ». Les participantes étaient enfin invitées à transmettre le nombre de vocaux de leur choix à J. Glikman. Les extraits des entretiens ainsi que les vocaux recueillis durant cette campagne sont inclus au corpus avec l'indication « 2021 » dans la colonne date des métadonnées et correspondent aux identifiants locuteurs de 01 à 07.

- Données recueillies via la campagne 2021 :
  - Nombre de locutrices : 7
  - Origine : Strasbourg et Paris
  - Nombre de vocaux : 59 et 7 extraits d'entretien
  - Durée totale : 1h00m44.73s

1.2 Campagne 2022

Pour la campagne de recueil 2022, un appel à participation a été lancé sur les réseaux sociaux et relayé sur les listes de diffusion, d'avril à septembre 2022. Les participants devaient répondre à un sondage sur Limesurvey. Le sondage permettait de recueillir leur consentement ainsi que des informations sur leur pratique des vocaux. Le sondage contenait aussi des questions permettant d'établir un jeu de métadonnées minimales (sexe, lieu de naissance, lieu de vie actuelle). Hormis le recueil explicite du consentement, toutes les autres questions du sondage étaient facultatives. Une fois le sondage rempli, les participants avaient accès à un numéro de téléphone auquel ils pouvaient transmettre librement le nombre de vocaux voulus. Les vocaux envoyés devaient avoir été enregistrés par le participant lui-même, et ne pas avoir été enregistré dans le but de l'étude. Les réponses au sondage, quand présentes, ont permis de construire les métadonnées en lien avec les vocaux reçus. Les vocaux reçus dans le cadre de cette campagne ont été inclus au corpus avec l'indication « 2022 » dans la colonne date des métadonnées. Pour plus d'information, voir la page du projet : https://www.atilf.fr/ressources/corpus-les-vocaux/.

- Données recueillies via la campagne 2022 :
  - Nombre de locuteurs : 45
  - Origine des locuteurs : 28 locuteurs de France ; 8 Belgique ; 4 Suisse ; 1 Canada
  - Nombre de vocaux : 1160
  - Durée totale : 19h06m49.2s

2 Contenu de cette version

La version 0.0.2 contient la totalité du corpus, soit **1196 fichiers audio** provenant des deux collectes (dont la totalité de la campagne 2021, y compris les extraits). Ces fichiers totalisent une durée de **19h32min49sec**, correspondant à plus de 240 000 tokens transcrits. Chaque fichier correspond à un vocal authentique (sauf les extraits de la campagne 2021, identifiés comme tels dans le nom de fichier même et dans les métadonnées). Ces 1196 vocaux constituent la version complète finale du corpus (campagne 2021 et 2022). Certains vocaux reçus ont été écartés de l'étude pour diverses raisons (par ex. présence d'un autre locuteur dans l'enregistrement). Les métadonnées ont été également enrichies par rapport à la précédente version (voir section 8).

Cette livraison inclut :

- Les fichiers audio anonymisés au format .wav
- Le fichier des métadonnées associées aux fichiers au format tableur (.ods, .xlsx et .csv) (voir section 8 Métadonnées)
- Les transcriptions orthographiques au format .txt encodées en UTF8 (voir section 7 Conventions de transcription)
- Le glossaire des mots ou graphies spécifiques au format tableur
- Une version **TXM** des transcriptions orthographiques associées aux métadonnées et annotées automatiquement en POS via la version Treetager intégrée à l'outil d'importation TXM (**attention** : ces annotations ne sont pas vérifiées et seront amenées à évoluer dans les distributions suivantes du corpus)
- Une version compilant toutes les transcriptions précédées des métadonnées du fichier dans un seul fichier .txt (compatible Libre Office, Word, NotePad…). Les métadonnées ont été encodées de manière à être exploitable dans le logiciel **Lexico** (format < … >)

Les fichiers .wav et .txt portent exactement le même nom, dans lequel le premier nombre correspond à l'identifiant locuteur, le deuxième nombre permet d'identifier de manière unique le vocal par un numéro (IDloc_NumVocal)

Note : le corpus étant toujours en cours d'édition, la version distribuée doit être considérée comme une version « alpha » provisoire. Certains choix éditoriaux sont susceptibles d'être modifiés (outre les corrections d'erreurs manifestes). **Il convient ainsi de veiller à citer explicitement la version du corpus utilisée pour toute recherche**.

3 Objectifs du projet

Le corpus **Les Vocaux** est réalisé dans le cadre du projet ORALIDIA (*Oralité et diachronie : une voie d'accès au changement linguistique*). Malgré le développement des corpus oraux, l'accès à des contextes diversifiés d'oral spontané reste difficile, l'entretien étant de loin la situation la plus représentée. Le projet ORALIDIA vise à la constitution d'un corpus inédit de français parlé spontané : les «sms vocaux» ou «vocaux». Ces données sont spontanément produites en dehors de toute enquête ou entretien linguistique, et constituent une voie d'accès à la parole spontanée non surveillée, nécessaire pour la description de la langue naturelle. Ces données sont ainsi le lieu privilégié pour l'étude de la diffusion des formes émergentes ou de leur disparition. A terme, le corpus comportera les fichiers audio, une transcription

orthographique, un alignement phonétique au signal, une lemmatisation, une annotation morphosyntaxique et une annotation syntaxique de type UD.

4 Membres du projet

4.1 Principaux membres

Julie Glikman (Université de Lorraine, ATILF) https://perso.atilf.fr/jglikman/
Nicolas Mazziotta (Université de Liège, Traverses) http://orbi.ulg.ac.be/ph-search?uid=U185884
Camille Fauth (Université de Strasbourg, LiLPa) https://lilpa.unistra.fr/index.php?id=19800
Christophe Benzitoun (Université de Lorraine, ATILF) https://perso.atilf.fr/benzitoun

4.2 Autres participants et stagiaires

Mélanie Lancien (U. Lorraine, collaboratrice)
Mathilde Hutin (ATILF, collaboratrice)
Thomas Verjans (U. Toulouse, collaborateur)
Auphélie Ferreira (U. Strasbourg, collaboratrice)
Lori Lamel (Limsi, collaboratrice)
Philippe Boula de Mareuil (Limsi, collaborateur)
Thalassio Briand (U. Strasbourg, stagiaire)
Salomé Klein (U. Strasbourg, stagiaire)
Elia Vertueux (U. Strasbourg, stagiaire)
Jonathan Fontaine (U. Strasbourg, stagiaire)
Hanji Kim (U. Strasbourg, stagiaire)
Lou-Anne Gartiser (U. Strasbourg, stagiaire)
Lilian Huther (U. Strasbourg, stagiaire)
Maïwenn Tellosa (U. Lorraine, stagiaire)

5 Financement

Le projet a reçu le financement de l'IDEX Exploratoire de l'Université de Strasbourg (oct. 2022-déc 2024 – 18 000 euros), du laboratoire ATILF et de l'Université de Lorraine (2022 – 4 100 euros), ainsi que le soutien du CNRS (accueil de J. Glikman en délégation CNRS à l'ATILF 2021-2023)

6 Publications et communications

Glikman J., C. Fauth (2022) « Un nouvel accès à la parole spontanée : les vocaux » *34e Journées d'Études sur la Parole, JEP2022*, 154 162. ISCA. doi.org/10.21437/JEP.2022-17. https://www.isca-speech.org/archive/pdfs/jep_2022/glikman22_jep.pdf

Mazziotta, N. & Glikman, J. (2023). Emplois discursifs et pragmatiques des formes du verbe *écouter* : Observations sur les corpus *88milSMS* et *Les Vocaux*. In M. Saiz-Sánchez & S. Gómez-Jordana Ferary (Eds.), *Études de sémantique et pragmatique en synchronie et diachronie. Hommage à Amalia Rodríguez Somolinos*. Presses Universitaires de Savoie Mont Blanc. https://hdl.handle.net/2268/304614

Delferrière, F. (2023). *Les marqueurs discursifs comme articulateurs d'énoncés : étude d'un corpus de messages vocaux contemporains*. (Unpublished master's thesis). Université de Liège, Liège, Belgique. http://hdl.handle.net/2268.2/17588

Benzitoun Christophe, Julie Glikman, Nicolas Mazziotta et Camille Fauth, "Les vocaux : de la constitution à l'exploitation d'un corpus de pratiques langagières émergentes", Journée d'études *LangaJE, Données et corpus en linguistique et en didactique à la lumière de la science ouverte : problématiques et enjeux méthodologiques*, Lyon, 15 nov. 2024.

Glikman J., Mazziotta N. (2022) « Projet "Les Vocaux" : Outils et formats », *TraSoGal*, Liège, 24 juin 2022. https://hdl.handle.net/2268/294441

Glikman J., C. Fauth, N. Mazziotta, C. Benzitoun (2022) « Une nouvelle voie d'accès au français populaire : l*es Vocaux* », *13e congrès des francoromanistes*, 21-24 septembre 2022, Université de Vienne. https://hal.science/hal-04312509

Glikman J., Mazziotta N., Fauth C., Benzitoun C. (2022) « Le projet *Les Vocaux* : bilan d'étape. » *Sciences participatives et nouvelles données*, Nancy, 30 sept. 2022. https://hal.science/hal-04312522

Glikman J. (2022) Présentation du projet *Les Vocaux*, entretien pour le magazine *Savoir(s)* de l'Université de Strasbourg : https://savoirs.unistra.fr/eclairage/les-enregistrements-vocaux-passes-a-la-loupe

Glikman, J., N. Mazziotta (2023). « Le projet *Les Vocaux:* premières analyses » Séminaire PRAXILING, Montpellier, France. [Paper presentation]. https://hdl.handle.net/2268/302120

Glikman J. (2023) « Le projet *Les Vocaux* : mise en place et chaîne de traitement », Université Saint-Louis, Bruxelles, Belgique, mars 2023.

Glikman J. (2023) « Les "vocaux" constituent-ils un "genre" ? », Séminaire CLLE, Toulouse, avril 2023. https://hal.science/hal-04312595

Glikman J. (2023) « Retour sur les verbes parenthétiques », Séminaire Lattice, Paris, avril 2023. https://hal.science/hal-04312574

Glikman J., C. Benzitoun, C. Fauth, N. Mazziotta (2023) « Étudier la variation intra-individuelle : exploitation à partir du corpus *Les Vocaux*. » *JTTR L'ORATEUR & L'ORATRICE IN SITU : pluriphonie, agentivité et identités*. ATILF, 20 novembre 2023. Vidéo de la conférence : https://ultv.univ-lorraine.fr/atilf-en-video/video/16229-journee-thematique-transversale-de-latilf-lorateur-loratrice-in-situ-pluriphonie-agentivite-et-identites/

7 Conventions de transcription

7.1 Principes d'anonymisation

Les vocaux sont anonymisés selon les principes suivants :

- le signal sonore n'est pas modifié (fréquence fondamentale, etc.) ;
- sont cependant masqués dans le signal sonore par un bip ainsi que dans les transcriptions et annotations correspondantes :
  - les prénoms, surnoms et noms de famille lorsqu'il ne s'agit pas d'une personne célèbre ou d'un personnage historique, politique, etc. ;
    - balise **#anon** ou **#name** dans la transcription (version 0.0.2)
  - les noms de lieux sont conservés sauf s'ils permettent de retrouver des informations personnelles (adresses complètes), et, dans ce cas, ils sont anonymisés ;
    - balise **#anon** ou **#city** ou **#location** (pour une localisation autre) dans la transcription (version 0.0.2)
  - les noms de ville sont conservés sauf s'ils sont liés à une information permettant d'identifier le locuteur (ex. *je travaille à l'hôpital de Mulhouse*) ;
    - balise **#anon** ou **#city** dans la transcription (version 0.0.2)
  - les surnoms et appellatifs qui ne sont pas particuliers tels que « maman », « ma chérie », « Darling » ne sont pas anonymisés parce que jugés courants et non personnels.

- les interventions d'autres locuteurs clairement audibles sont également masqués par un bip
    - balise **#anon** dans la transcription, (version 0.0.2)

Après anonymisation, la transcription orthographique est effectuée dans un premier temps automatiquement sur la plateforme YobiYoba (https://www.yobiyoba.com/fr/). Cette première transcription automatique est ensuite révisée et corrigée manuellement (en général un premier passage est effectué par un stagiaire et validée une seconde fois par un collaborateur expert) avant d'être exportée sous la forme d'un fichier .txt.

7.2 Principes généraux de transcription

Le corpus est transcrit en orthographe standard traditionnelle avec application des rectifications orthographiques (y compris dans la résolution des amalgames ex. « je suis » et élisions ex. « tu as » prononcé [ta]). Cependant, pour les formes *il* et *ne* : ils sont transcrits uniquement quand ils sont clairement audibles ou interprétables dans le phone (ex. [õnɛ] transcrit « on n'est »). Ils ne sont pas transcrits quand ils ne sont pas récupérables dans le phone (ex. « on veut »).

- Pause :
    - On considère qu'il y a pause au-delà de 80 milliseconde d'absence de parole. Elles sont indiquées par « … » dans la transcription.
    - **Important** : Cette version du corpus n'inclut pas systématiquement les pauses, qui sont définies selon leur durée minimale lors de l'alignement texte – parole. La présence des pauses connait donc un traitement différent selon l'état d'avancement du traitement des fichiers du corpus.
- Majuscules :
    - On applique des majuscules aux noms (y compris noms de magasin)
    - On met des majuscules aux sigles et aux acronymes (tout, sans point, quelle que soit la prononciation)
- Ponctuation :
    - La transcription orthographique n'est pas ponctuée à ce stade (excepté les pauses marquées par « … »)
- etc.
    - On transcrit en toutes lettres « et caetera »
- Chiffres et nombres :
    - Ils sont transcrits en toutes lettres selon la nouvelle orthographe (« deux-mille-vingt-deux » sauf usage spécifique (type « 3D »)
- Acronymes
    - Les acronymes sont transcrits en majuscules et sans point (OK, BU, SDF…)
- Amalgames (type *chuis chais, m'fin*…)
    - On décompose systématiquement : *je suis / je sais / mais enfin*
- Élision (type *j'vais, t'as, i'vient, 'fin*) :
    - On rétablit l'orthographe standard : *je vais, tu as, il vient, enfin*
    - Mais : s'il s'agit d'un usage spécifique lexicalisé, la forme est conservée : ex. *je me suis fait plaiz*
- Il :

    - Le *il* est transcrit quand un doute est possible sur sa réalisation effective : ex. *il y a* : peut être interprété dans le phone [i]
    - Non transcrit quand aucun doute sur son absence : *faut*
- Négation :
    - On transcrit le *ne* de négation s'il est audible
    - Si clairement non audible, il n'est pas transcrit : *tu veux pas, faut pas …*
    - En cas de doute possible sur sa prononciation, notamment en cas de liaison : *on N a pas…* => elle est transcrite quand peut être interprétée dans le phone
    - Le *non* est systématiquement transcrit *NON*, y compris quand la prononciation tend vers *nan*
- Liaison pataquès (*il n'est point Z arrivé, les chemins de fer Z américains*) :
    - On conserve l'orthographe standard : *il n'est point arrivé, les chemins de fer américains*
- Onomatopées, interjections et marqueurs d'hésitation :
    - Les *euh hum* etc. sont autant que possible transcrits
    - Nous avons suivi la liste du guide de transcription du corpus TCOF ci-dessous pour la graphie :
        - ah, aïe, areu, atchoum, badaboum, baf, bah, bam, bang, bé, bêêê, beurk, ben, bing, bon, boum, broum, cataclop, clap clap, coa coa, cocorico, coin coin, crac, croa croa, cuicui, ding, ding deng dong, ding dong, dring, hé, hé ben, eh bien, euh, flic flac, flip flop, frou frou, glouglou, glou glou, gni, groin groin, grr, hé, hep, hi han, hip hip hip hourra, houla, hourra, hum, mêêê, meuh, miam, miam miam, miaou, oh, ouah, ouah ouah, ouais, ouf, ouh, paf, pan, patatras, pchhh, pchit, pff, pif-paf, pin pon, pioupiou, plouf, pof, pouet, pouet pouet, pouf, psst, ron ron, schlaf, snif, splaf, splatch, sss, tacatac, tagada, tchac, teuf teuf, tic tac, toc, tut tut, vlan, vroum, vrrr, wouah, zip
    - Certaines transcriptions ont été ajoutés quand nécessaire, leur recensement est en cours et sera documenté dans les prochaines versions
- Mots abrégés :
    - On utilise quand c'est possible la graphie attestée et usuelle dans un dictionnaire (y compris le wiktionnaire) (ex. *prof, fac, plaiz*). En cas de graphie multiple, la graphie choisie est mentionnée dans le glossaire (par exemple *resto* / *restau*) (voir section 7.3)
    - On transcrit cependant systématiquement sans apostrophe pour uniformiser le traitement
- Préfixe *re-*
    - On transcrit toutes les préfixations en *re-* soudées au mot, y compris quand elles ne sont pas lexicalisées (*reregardé*).
- Amorces :
    - On transcrit le phonème entendu suivi d'un tilde (~) : j'av~ ; l'amorce choisi est un mot existant dans la langue, correspondant a priori à l'intention ; en cas de doute, on a opté dans tous les cas pour un mot existant (ces t~/ c'ét~ (= c'était) vs sét~)

- Bruits physiologiques (raclement de gorge, respiration, aspiration…)
  - balise **#noise** dans la transcription (version 0.0.2)
- Rires
  - balise **#laugh** (pour *laughter*) **ou #noise** dans la transcription (version 0.0.2)

Les segments de paroles inaudibles, incompréhensibles ou indistincts ont été notés par la balise **#unk** (pour *unknown*) dans la transcription (version 0.0.2).

7.3 Principes de tokenisation (version 0.0.2)

Outre les mots composés clairement identifiés (chiffres systématiquement avec tirets conformément à la nouvelle orthographe, *aujourd'hui*, *après-midi*…), on a inséré un espace pour des fins de traitement :

- Après l'apostrophe (j' ai)
- Avant le tiret (est -ce que)
- A ce stade, les mots comme *parce que* sont considérés comme deux tokens

7.4 Tokens non lexicaux

Ainsi, outre les tokens lexicaux (y compris les amorces), la version 0.0.2 du corpus contient les tokens non lexicaux suivants :

- **…** = pauses (traitement hétérogène à homogénéiser)
- **#unk** (unknown) = segment incompréhensible, indistinct, inaudible
- **#laugh** (laughter) = rire
- **#anon** = segment anonymisé (à trier name/city)
- **#city**, **#name** = segment anonymisé trié par nom (name) ou ville (city)

7.5 Glossaire

Le fichier glossaire contient les mots spéciaux avec identification de la graphie retenue (argot, troncations, abréviations, dérivations non lexicalisées…). Le glossaire est en cours de constitution, la version alpha distribuée avec la version 0.0.2 du corpus est donc une version non aboutie du glossaire, qui est voué à évoluer également.

8 Métadonnées

Le fichier des métadonnées comporte les informations suivantes, renseignées pour chaque vocal (si fourni par le locuteur) :

- SUBCORPUS = vocal ou interview (permet de distinguer les vocaux authentiques des extraits d'entretien)
- VOCAL_ID = identifiant unique du vocal
- SPEAKER_ID = identifiant unique du locuteur ayant produit le vocal
- SPEAKER_GENDER = genre du locuteur
  - F = femme
  - H = homme
  - Autre

    - Non binaire
  - SPEAKER_AGE = âge du locuteur
  - SPEAKER_CURRENT_LOCATION = ville actuelle de résidence du locuteur
  - SPEAKER_CURRENT_COUNTRY = pays actuel de résidence du locuteur
  - SPEAKER_HOME_LOCATION = ville de naissance du locuteur
  - SPEAKER_HOME_COUNTRY = pays de naissance du locuteur
  - SPEAKER_FRENCH_L1 = le locuteur a le français comme langue maternelle
    - OUI
    - NON
  - SPEAKER_MESSAGE_USE = habitudes d'utilisation des vocaux du locuteur
    - De temps en temps
    - Rarement
    - Souvent
    - Tout le temps
  - VOCAL_PRODUCTION_YEAR = année de recueil du vocal
    - 2021 (campagne 2021)
    - 2022 (campagne 2022)
  - SPEAKER_TOTAL_VOCAL_COUNT = nombre total de vocaux fourni par le locuteur
  - SPEAKER_TOTAL_VOCAL_DURATION = durée totale des vocaux fournis par le locuteur
  - VOCAL_DURATION_SECONDS = durée du vocal (exprimé en secondes)
  - VOCAL_DURATION_READABLE = durée du vocal (retranscrite au format HEURES:MINUTES:SECONDE)

9 Changements depuis la version 0.0.1

La version V.0.0.1 ne contenait que 10h de données. Cette nouvelle version contient la totalité du corpus. Les métadonnées ont été enrichies de plusieurs informations complémentaires (voir section 8). En outre, un certain nombre de problèmes ont été corrigés pour l'ensemble du corpus (y compris les fichiers déjà présents dans la version précédente) :

- Encodage unique du corpus en UTF8
- Uniformisation des apostrophes en un même caractère unique
- Uniformisation des conventions de transcription pour les apocopes : systématiquement transcrites sans apostrophe
- Modification des balises (voir section 7.4)
- Modification des conventions de transcription des amorces : notées par un tilde dans cette version (anciennement notées par un tiret dans V.0.0.1)
- Uniformisation des conventions orthographiques en faveur de la nouvelle orthographe, avec corrections systématiques (en particulier pour les chiffres)