Distribution du corpus BEL-RL-fr
Version 3.1

Sandrine Ollinger
sandrine.ollinger@atilf.fr

26 avril 2024

Table des matières

Introduction

Cette documentation décrit le contenu de la version 3.1 des fichiers de distribution sur la plateforme Ortolang du corpus BEL-RL-fr [1].

Le corpus BEL-RL-fr est une distribution de la base de citations lexicographiques du Réseau Lexical du français (RL-fr)[2], modèle formel du lexique du français contemporain en cours de construction. Sa réalisation s'inscrit dans le cadre des travaux sur les *Systèmes Lexicaux*[3] réalisés au laboratoire ATILF.

Chacune des citations qu'il contient a été sélectionnée avec soin pour servir d'exemple lexicographique en illustrant le sens et la combinatoire d'une unité lexicale et provient initialement de l'une des sources suivantes :

- la base de données textuelles Frantext,
- le corpus Web frWaC,
- le corpus journalistique issu de l'Est Républicain,
- le corpus OrthoCorpus,
- diverses sources hors corpus (Web, publications, supports audiovisuels, conversations, courriers, chansons).

Dans de rares cas, les citations ont été fabriquées par les lexicographes.

[1] `https://hdl.handle.net/11403/examples-ls-fr`
[2] `https://www.ortolang.fr/market/lexicons/lexical-system-fr`
[3] `https://lexical-systems.atilf.fr/`

Pour une compréhension des opérations de sélection des citations par les lexicographes, nous vous recommandons la lecture de Ollinger et Polguère (2020) et Lux-Pogodalla (2014).

1 Statistiques V3.1

La version courante de BEL-RL-fr comporte :

- 31 987 citations
- 53 013 segments textuels annotés en unité lexicale

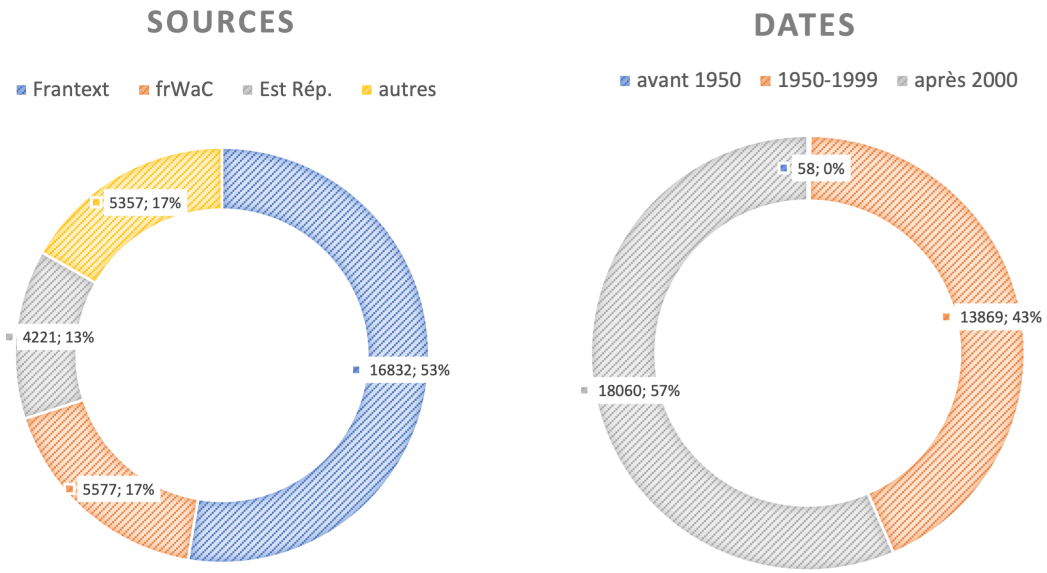

FIG. 1 : Répartition des citations par sources et par dates

2 Historique des versions

2.1 Version 3.1 : changements par rapport à la version 3

- La version 3.1 comporte un enrichissement des données linguistiques de la version 3 : ajout de citations et ajouts de segments textuels annotés.
- De plus, elle intègre une segmentation en phrases réalisée à l'aide de la cascade de graphes CasFin[4] Ollinger et Maurel (2024 à paraître). Cette segmentation et sa réintégration aux fichiers originaux ont bénéficié de l'aide de Denis Maurel et de Bertrand Gaiffe. Merci à eux.

2.2 Version 3 : changements par rapport à la version 2

- La longueur des codes de statuts lexicaux a changé. Elle est passée de 14 caractères à 22 caractères.
- La version 3 comporte un enrichissement des données linguistiques de la version 2 : ajout de citations et ajouts de segments textuels annotés.

[4]La cascade de graphes CasFin est disponible au téléchargement à l'adresse `https://tln.lifat.univ-tours.fr/version-francaise/ressources/casfin`.'

- De plus, elle intègre une partie supplémentaire du travail de révision réalisé par Camille Kuntz au cours de l'été 2020.

2.3 Version 2 : changements par rapport à la version 1

- La version 2 comporte un enrichissement des données de la version 1 : ajout de citations et ajouts de segments textuels annotés.
- De plus, elle intègre une partie du travail de révision réalisé par Camille Kuntz au cours de l'été 2020.
- Enfin, elle comporte pour la première fois une version TXM du corpus, pour la réalisation de laquelle nous avons bénéficié de l'aide d'Alexey Lavrentev, Matthieu Decorde et Serge Heiden. Merci à eux.

3 Organisation générale des fichiers d'export

Le corpus suit une organisation par sources, dans un format XML. La version courante de l'export[5] comporte le fichier `.rng`, les 12 fichiers `.xml` et le fichier `.txm` ci-dessous, tous encodés en UTF-8.

- **Schéma de données**
  `LSexamples.rng`
- **Fichier global**
  `BEL-RL-fr.xml`
- **Fichiers par source**
  `BEL-RL-fr_Frt.xml`
  `BEL-RL-fr_FrW.xml`
  `BEL-RL-fr_ER.xml`
  `BEL-RL-fr_OC.xml`
  `BEL-RL-fr_HbP.xml`
  `BEL-RL-fr_HbA.xml`
  `BEL-RL-fr_HbCh.xml`
  `BEL-RL-fr_HbConv.xml`
  `BEL-RL-fr_HBW.xml`
  `BEL-RL-fr_HBCour.xml`
  `BEL-RL-fr_HbF.xml`
- **Fichier TXM**
  `BELRLFR-V3-1.txm`

Les fichiers `.xml` par source contiennent les citations lexicographiques, enrichies en sens lexicaux, telles que décrites dans la section 4. La structure de ces fichiers, conforme aux recommendations de la Text Encoding Initiative[6], est précisée sous forme de schéma Relax NG dans le fichier `LSexamples.rng`.

Le fichier `.txm` contient l'ensemble du corpus à importer dans le logiciel de textométrie TXM développé par Heiden et al. (2010).

4 Encodage XML TEI des citations

Le BEL-RL-fr constitue une ressource annotée manuellement en sens lexicaux.

Chaque citation est accompagnée de l'ensemble de métadonnées suivant :

- un identifiant unique,
- un type dont la valeur est ici systématiquement « example »,
- une source, sous forme d'URI, qui permet de retrouver la citation dans le RL-fr.

[5] L'export est réalisé en PHP à l'aide d'un ensemble de scripts Python programmés par M. Schmitt et maintenus par S. Ollinger.

[6] `https://tei-c.org/`

```
<cit xml:id="cit9" type="example" source="ls:fr:ex:21">
  <quote><s>Et il débarquait, royal, dans le jeu de quilles quadrillé de nos amitiés, fonçant
      avec sa Jaguar pour <seg xml:id="seg42" type="st0000000000000000000" ana="ls:fr:gc:23"
        function="active" source="https://spiderlex.atilf.fr/fr/id/39110">kidnapper</seg> l'un
      d'entre nous et l'inviter à dîner dans un grand restaurant, ou déposant avec naturel
      comme offrande sur le pas de la porte où il s'imposait une caisse de mouton-rothschild
      qu'il avait payée quelques millions aux enchères à Drouot.</s></quote>
  <bibl>
    <title>À l'ami qui ne m'a pas sauvé la vie</title>
    <author>Guibert, Hervé</author>
    <date when="1990">1990</date>
    <biblScope>p. 197</biblScope>
  </bibl>
</cit>
```

Fig. 2 : Exemple d'encodage de citation

4.1 Annotation en sens lexicaux

Chacune des annotations renvoie à une unité lexicale du Réseau Lexicale du Français (RL-fr). Des balises <seg/> délimitent les annotations. Chacune d'entre elles est associée à l'ensemble d'informations suivant :

- un identifiant unique,
- un code qui renseigne le statut lexical de l'unité qu'elle illustre (statut phraséologique, sémantique, pragmatique et langagier),
- un URI qui renvoie à sa partie du discours dans le modèle des caractéristiques grammaticales du RL-fr,
- un attribut function[7] dont la valeur est ici systématiquement « active »,
- une URL qui permet de visualiser la description de l'unité lexicale, dans sa version actualisée quotidiennement[8].

```
<quote><s>Et il débarquait, royal, dans le jeu de quilles quadrillé de nos amitiés, fonçant
    avec sa Jaguar pour <seg xml:id="seg42" type="st0000000000000000000" ana="ls:fr:gc:23"
      function="active" source="https://spiderlex.atilf.fr/fr/id/39110">kidnapper</seg> l'un
    d'entre nous et l'inviter à dîner dans un grand restaurant, ou déposant avec naturel
    comme offrande sur le pas de la porte où il s'imposait une caisse de mouton-rothschild
    qu'il avait payée quelques millions aux enchères à Drouot.</s></quote>
```

Fig. 3 : Exemple d'annotation d'une occurrence de KIDNAPPER II.1

Un glossaire des codes employés pour les statuts lexicaux est disponible dans l'entête de chaque fichier.

Il est possible de reconstruire l'URI de chaque unité lexicale présente à partir des URL fournies, de la manière suivante : https ://spiderlex.atilf.fr/fr/id/**35415** => ls:fr:node:**35415**.

4.2 Occurrences multiples ou disjointes

1 678 citations comportent plusieurs segments textuels annotés avec la même unité lexicale. Il peut s'agir aussi bien d'occurrences multiples que d'occurrences disjointes.

[7]Dans des versions internes du corpus BEL-RL-fr, cet attribut nous permet de différencier les annotations rapides ou incertaines (inactives), des annotations plus mûrement réfléchies (actives).

[8]Pour consulter la description d'une unité lexicale au moment de la création de la version V3.1 du corpus BEL-RL-fr, vous pouvez vous référer à la version V3.1 du lexique RL-fr disponible sur Ortolang.

On doit préconiser le numérotage des générations par des **chiffres** romains, et celui des individus, de la gauche à la droite dans chaque génération, par des **chiffres** arabes.

(**Frantext**, Sans mention d'auteur, *L'Histoire et ses méthodes*, 1961, p. 733)

```
<quote>On doit préconiser le numérotage des générations par des <seg xml:id="seg2277"
    type="st0000000000000000000" ana="ls:fr:gc:20" function="active"
    source="https://spiderlex.atilf.fr/fr/id/26748">chiffres</seg> romains, et celui des
  individus, de la gauche à la droite dans chaque génération, par des <seg xml:id="seg2278"
    type="st0000000000000000000" ana="ls:fr:gc:20" function="active"
    source="https://spiderlex.atilf.fr/fr/id/26748">chiffres</seg> arabes.</quote>
```

FIG. 4 : Exemple d'annotation d'occurrences multiples de CHIFFRES I.1a

Un dimanche après-midi, le banquier m'avait donné rendez-vous au bar du Ritz. J'arrivais en avance. Assise dans le hall d'entrée, je n'**avais** pas **pris** de table, j'attendais.

(**Frantext**, Christine Angot, *Rendez-vous*, 2006, p. 111)

```
<quote>Un dimanche après-midi, le banquier m'avait donné rendez-vous au bar du Ritz.
  J'arrivais en avance. Assise dans le hall d'entrée, je n'<seg xml:id="seg29839"
    type="st0000000000000000000" ana="ls:fr:gc:23" function="active"
    source="https://spiderlex.atilf.fr/fr/id/45088">avais</seg> pas <seg xml:id="seg29840"
    type="st0000000000000000000" ana="ls:fr:gc:23" function="active"
    source="https://spiderlex.atilf.fr/fr/id/45088">pris</seg> de table,
  j'attendais.</quote>
```

FIG. 5 : Exemple d'annotation d'une occurrence disjointe de PRENDRE II.1

4.3 Références bibliographiques

Chaque citation est accompagnée de sa bibliographie qui se compose des éléments optionnels suivants :

- un titre,
- un auteur,
- une date,
- un emplacement, pour lequel le nom de balise varie entre <biblScope/> et <ref/>.

```
<bibl>
  <title>À l'ami qui ne m'a pas sauvé la vie</title>
  <author>Guibert, Hervé</author>
  <date when="1990">1990</date>
  <biblScope>p. 217</biblScope>
</bibl>
```

```
<bibl>
  <title/>
  <author/>
  <date when="2012-10-02">02 10 2012</date>
  <ref type="url">http://claireaunepal.wordpress.com/</ref>
</bibl>
```

FIG. 6 : Exemples de référence bibliographique

Chaque citation, à l'exception des citations fabriquées, est associée à une date :

- date de publication (Frantext, Est Républicain, OrthoCorpus, publications hors corpus, support audiovisuel, chansons),

- date de consultation (Web),
- date de constitution du corpus (frWac),
- date de création.

Seules 60 citations sont antérieurs à 1950. Elles servent à illustrer des unités lexicales aujourd'hui inactives en tant que telles, mais formellement incluses dans des unités polylexicales.

4.4 Segmentation en phrases

Chaque citation est segmentée en phrases selon les principes décrits dans Ollinger et Maurel (2024 à paraître). Cette segmentation autorise la présence de phrases incluses dans d'autres phrases, dans le but de limiter les ruptures syntaxiques. La figure 7 illustre une telle inclusion.

```
<cit xml:id="cit8198" type="example" source="ls:fr:ex:8291">
    <quote><s>À Toulon, le 4 mai, on colle des aigles sur les fleurs de lys des affiches
            administratives ; à Dole, le 9 juin, on <seg xml:id="seg18291"
                type="st00000000000000000000" ana="ls:fr:gc:23" function="active"
                source="https://spiderlex.atilf.fr/fr/id/42321">appose</seg> ce <seg
                xml:id="seg18292" type="st00000000000000000000" ana="ls:fr:gc:20"
                function="active" source="https://spiderlex.atilf.fr/fr/id/41481"
                >placard</seg> : « <s>Vive le roi pour trois jours ! vive Bonaparte pour
                toujours !</s> » </s></quote>
    <bibl>
        <title>1815</title>
        <author>Houssaye, Henry</author>
        <date when="1893">1893</date>
        <biblScope>p. 8</biblScope>
    </bibl>
</cit>
```

Fig. 7 : Exemple de phrase incluse dans une phrase

5 Vérifications des données

Comme le détaillent Ollinger et Polguère (2020), les lexicographes sont attentifs à différents aspects lors de l'intégration d'une nouvelle citation au RL-fr. Pour renforcer cette vigilance, un ensemble de vérifications sont réalisées de manière régulière.

5.1 Systématiques

Les points suivants font l'objet d'une vérification hebdomadaire.

Dans le texte des citations :

- Y a-t-il des guillemets doubles droits ou des apostrophes droites ?
- Y a-t-il des mentions [SIC] collées à ce qui précède ?
- Y a-t-il des suites d'espaces, insécables ou non ?
- Y a-t-il un nombre impaire de guillemets ?

Dans les références bibliographiques :

- Y a-t-il des guillemets doubles droits ou des apostrophes droites ?

- Y a-t-il des espaces en début de titres ou de noms d'auteurs ?
- Y a-t-il des espaces en fin de titres ou de noms d'auteurs ?
- Y a-t-il des virgules à la fin de titres ?
- Y a-t-il des espaces superflues entre les noms d'auteurs en cas d'auteurs multiples ?
- Y a-t-il des citations tirées du FrWaC dont la date n'est pas février 2008 ?
- Y a-t-il des citations dont la date contient une année qui ne compte pas quatre chiffres ?
- Y a-t-il des citations qui devraient être associées à une date, mais qui ne le sont pas ?

Dans les associations entre unités lexicales et citations :

- Y a-t-il des cas d'associations pour lesquels aucun segment textuel n'est identifié comme étant une occurrence de la lexie associée ?
- Y a-t-il des cas d'association où le segment textuel identifié dépasse la longueur de la citation ?
- Y a-t-il des cas où le segment identifié semble coupé en deux ?
- Y a-t-il des segments identifiés qui commencent par une espace ?
- Y a-t-il des segments identifiés qui se terminent par une espace ?
- Y a-t-il des cas où le segment identifié semble incomplet ?

5.2 Récurrentes

D'autres points sont vérifiés avant chaque diffusion sur Ortolang :

- Y a-t-il des segments textuels associés à deux lexies distinctes ? Et si oui, est-ce qu'il s'agit d'erreurs ?
- Y a-t-il des erreurs dans les cas de lexies dont le statut lexical est composé de plusieurs éléments ?
- Y a-t-il des citations de type « HorsBases temporaires » qui ne sont pas inactives ?
- Y a-t-il des citations en double (strictement identiques) ?

5.3 Exceptionnelles

Enfin, quelques points sont vérifiés lors de campagnes spécifiques :

- Y a-t-il des erreurs dans les noms d'auteurs qui provoquent des doublons ?
- Y a-t-il des erreurs dans les titres qui provoquent des doublons ?
- Y a-t-il des citations en partie identiques qui peuvent être regroupées ?
- Y a-t-il des erreurs de typographie ou d'orthographe ?
- Y a-t-il des citations qui sont trop longues ou difficiles à comprendre ?

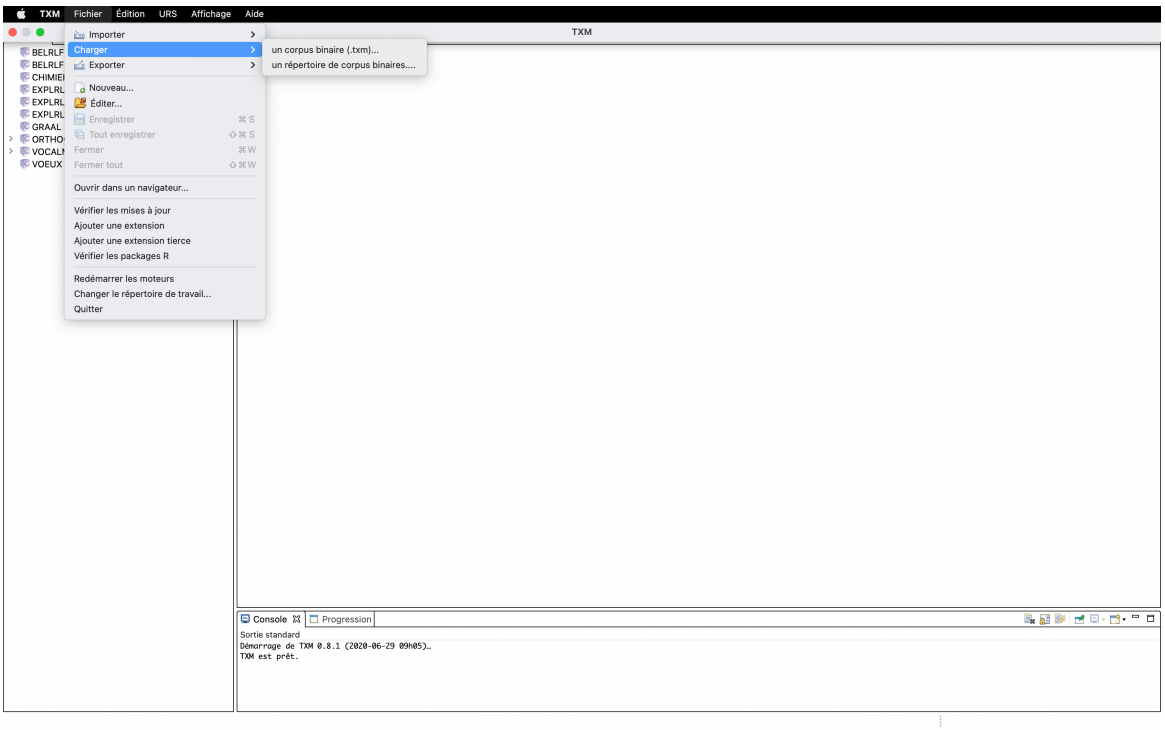

FIG. 8 : Fichier > Charger > un corpus binaire (.txm)...

6 Chargement du corpus dans TXM

Afin de faciliter la consultation du BEL-RL-fr et son interrogation plus avancée à l'aide du logiciel de textométrie TXM, nous diffusons le fichier BEL-RL-frV3-1.txm.

Ce fichier peut être directement chargé dans le logiciel, sans passer par une procédure d'import.

Commencez par sélectionner le fichier BEL-RL-fr.txm depuis le menu Fichiers > Charger > un corpus binaire (.txm)...

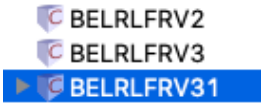

FIG. 9 : Liste de corpus

À la fin de l'opération, la Console indique que le corpus est chargé et le corpus apparaît dans la liste des corpus dans la partie gauche de la fenêtre.

Pour de plus amples informations sur l'utilisation du logiciel TXM, nous vous invitons à consulter le site `https://txm.gitpages.huma-num.fr/textometrie/`.

Références

Serge Heiden, Jean-Philippe Magué et Bénédicte Pincemin. TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. volume 2, page 1021. Edizioni Universitarie di Lettere Economia Diritto, juin 2010. URL `https://halshs.archives-ouvertes.fr/halshs-00549779`. Issue : 3.

Veronika Lux-Pogodalla. Intégration relationnelle des exemples lexicographiques dans un réseau lexical. Dans Brigitte Bigi, dir., *Actes de TALN 2014*, pages 586–591, Marseille, 2014. Laboratoire Parole et Langage, Aix-en-Provence.

Sandrine Ollinger et Denis Maurel. Segmentation en phrases : ouvrez les guillemets sans perdre le fil. Dans *Actes des 17e Journées internationales d'analyse statistique des Données Textuelles*, 2024 à paraître.

Sandrine Ollinger et Alain Polguère. Mémo Systèmes Lexicaux. exemples lexicographiques, 2020.