Corpus TALN
Version 1 – 28 Avril 2020

1/ Description :

Le corpus TALN rassemble les articles des conférences TALN et RÉCITAL des années 1997 à 2019. Il se compose de 1602 articles scientifiques en qui traitent du domaine du traitement automatique des langues (TAL) pour un total d'environ 5,8 millions de mots.
Le corpus est essentiellement en français : n'ont été retenus que les articles rédigés en français ou pour lesquels une partie au moins du contenu est en français (résumé, mots-clés).

Les articles sont au format XML suivant la norme TEI P5 et leur structure contient les éléments suivants :
- métadonnées : titre ; nom des auteurs ; année ; éditeur ; lieu de la conférence ; résumé (en français et en anglais) ; mots-clés (en français et en anglais). Un identifiant unique est attribué à chaque article, indiquant notamment la conférence (TALN/RECITAL) et le type d'article (long, court, poster, invité, démonstration, etc.)
- corps du texte : sections et sous-sections, numéro, titre, type principal ; figures et tables (numéro et légende), notes de bas de page. Les paragraphes sont marqués à titre indicatif, et correspondent généralement aux segments de texte séparés par une marque structurelle parmi les précédentes ou un saut de page. La liste des références bibliographiques est indiquée mais non analysée (pas de découpage en items).

Le corpus est le fruit d'une conversion effectuée à partir des fichiers PDF servant à diffuser les actes des conférences. Malgré le soin apporté à sa constitution, un ensemble de problèmes résiduels peuvent être rencontrés, notamment des informations manquantes, des segments de texte absents, mal balisés ou présentant des problèmes de codage. Certains articles présents dans les actes des conférences peuvent être absents du corpus pour des raisons purement techniques liées au contenu du fichier PDF original.

Ce corpus a été constitué dans le cadre du projet ANR ADDICTE avec pour objectif l'apprentissage de modèles d'analyse distributionnelle en domaine de spécialité, et notamment pour étudier l'impact de la structure logique des documents. Le consortium CORLI (CNRS) a également apporté un soutien pour sa finalisation.

Pour citer cette ressource, merci d'utiliser la référence bibliographique suivante :

Tanguy, L., Fabre, C. et Bard, Y. (2020). Impact de la structure logique des documents sur les modèles distributionnels : expérimentations sur le corpus TALN. *Actes de TALN*, Nancy.

```
@InProceedings{TanguyFabreBard2020,
  author =    {Ludovic Tanguy and Cécile Fabre and Yoann Bard},
  title =     {Impact de la structure logique des documents sur les modèles
distributionnels~: expérimentations sur le corpus TALN},
  booktitle = {Actes de TALN},
  year =      2020,
  address =         {Nancy}}
```

Pour toute question, merci de contacter Ludovic Tanguy (ludovic.tanguy@univ-tlse2.fr), CLLE : CNRS & Université de Toulouse.

2/ Travaux antérieurs :

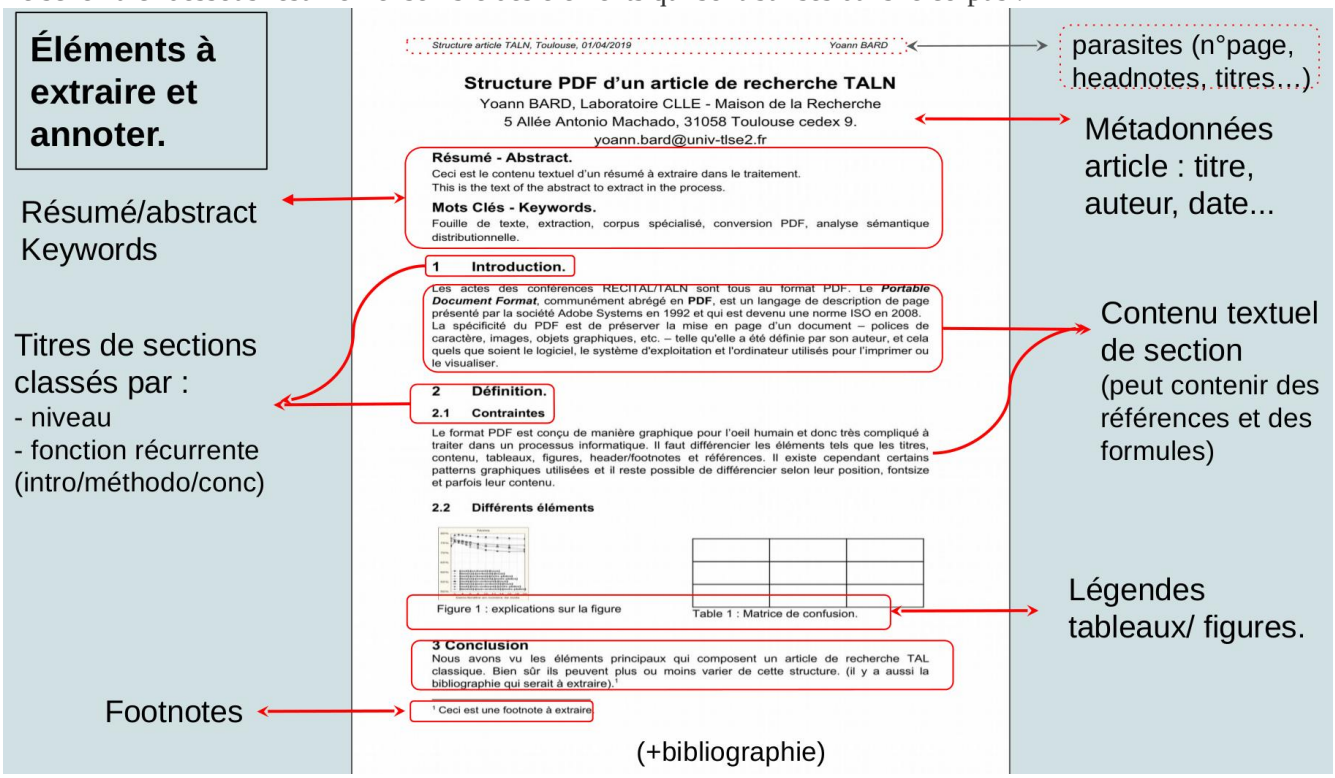
L'état actuel du corpus est l'aboutissement de deux étapes précédentes :

1. La création des Archives TALN par Florian Boudin : compilation des fichiers PDF des articles des conférences TALN/RECITAL de 1997 à 2015 ; extraction et formatage des métadonnées dans une base de données.

2. La création d'une première version du Corpus TALN par Ludovic Tanguy. Sur la base des fichiers des archives TALN, il a été procédé à une extraction grossière du contenu textuel des articles. Le corpus au format texte brut a été conçu avant tout pour ne retenir que les seuls phrases complètes en vue d'une analyse automatique (étiquetage morphosyntaxique, parsing, analyse distributionnelle), en supprimant une bonne partie des contenus (paratexte, titres de section, notes, légendes, bibliographie, etc.). Le corpus est disponible pour les années 2007 à 2013 et contient environ 2 millions de mots. Il est déjà diffusé librement avec l'accord de l'ATALA et est disponible sur http://redac.univ-tlse2.fr/corpus/taln.html

3/ Procédure de constitution :

Le schéma ci-dessous résume l'ensemble des éléments qui sont balisés dans le corpus :



Les traitements effectués à cet effet sont détaillés ci-dessous :

1) Récupération et segmentation des articles PDF de 2016 à 2019.

2) Conversion des fichiers du format PDF aux formats TXT et HTML avec *pdfminer* en respectant au mieux la structure graphique d'un document.

3) Segmentation et annotation des fichiers TXT avec *ParsCit* qui structure un article scientifique en séparant les titres de section, paragraphes, notes de bas de page, figures, tables, références etc.

4) Traitement automatique (scripts python à base de règles et expressions régulières) :

- Conversion au format XML des sorties ParsCit.

- Suppression des scories (caractères spéciaux, doublons symboles, idéogrammes, en-têtes, articles en anglais, numéro de page, noms d'auteurs ou de conférence …).

- Traitement des césures. Nous avons utilisé un lexique de mots composés pour ne pas supprimer les traits d'union correspondants.

- Extraction des métadonnées des articles de 1997 à 2015 à partir des archives constituées par Florian Boudin.

- Correction automatique du contenu textuel mal annoté lors du traitement précédent.

Dans les sorties ParsCit on retrouve un grand nombre de légendes de figures, de titres de sections et de notes de bas de page qui ne sont pas segmentés correctement ou qui se trouvent dans la mauvaise balise, plus fréquemment à l'intérieur d'un paragraphe.
On peut corriger automatiquement la plupart en suivant un schéma précis : les résumés commencent par la séquence de caractères : « RESUME » ; les légendes de figures sont souvent précédées par « FIGURE [NUM]. » en début de ligne. Les titres de sous-sections sont précédés par une suite de numéros telle que « 1.2. ». Il y a aussi des titres uniques comme « 5. Conclusion » ou « 1. Introduction », etc.

- Restructuration des balises et attribution de numéros et type (introduction, conclusion…) aux titres de section.

5) Traitement semi-automatique :

- Extraction des métadonnées des articles de 2016 à 2019 (titres, résumés, mots-clés, abstract, keywords) à partir des sites des conférences.

- Repérage et correction des cas d'ambiguïtés
On distingue un certain nombre de situations résultant d'ambiguïtés lors de la segmentation. Une note de bas de page peut être confondue avec un titre de section ou encore des légendes de tableaux qui sont coupées en deux par une fin de ligne.
Pour ces cas-là, il est facile de les repérer dans le texte mais plus compliqué de les corriger de manière automatique.
Deux approches ont été abordées :
Premièrement on effectue un diagnostic avec une liste non exhaustive des titres de section extraits des fichiers TALN convertis au format HTML. Si le schéma retrouvé dans le texte correspond avec un des titres de la liste, on peut procéder à une correction automatique.
La deuxième approche consiste à insérer un commentaire XML pour souligner l'ambiguïté et permettre une vérification manuelle plus rapide.

6) Traitement manuel : correction de la segmentation des articles et récupération du contenu textuel oublié lors du traitement.

7) Conversion au format TEI P5 : rédaction de l'en-tête et structuration des éléments repérés en respectant la norme correspondante. Vérification de la bonne formation et de la validité du corpus en utilisant la DTD TEI (complète).

4/ Structure du corpus

L’arborescence du corpus est la suivante :

```
<teiCorpus xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
      (en-tête du corpus : méta données, contributeurs, droits, licence, etc.)
  </teiHeader>
  <TEI>
      (contenu d’un article)
  </TEI
 […]
</teiCorpus>
```

Voici la structure générale d’un article !

```
<TEI>
      <teiHeader>
            <fileDesc xml:id="IDENTIFIANT INTERNE (voir note)">
                  <titleStmt>
                        <title xml:lang="fr|en">Titre principal (avec langue indiquée)</title>
                        <author>
                              <persName>
                                    <name>Prénom et nom de l’auteur</name>
                                    <email>Adresse mail de l’auteur</email>
                              </persName>
                        </author>
                    </titleStmt>
                  <publicationStmt>
                        <publisher>Nom de l’éditeur des actes</publisher>
                        <pubPlace>Lieu de la conférence</pubPlace>
                        <date>Année de la conférence</date>
                  </publicationStmt>
                  <sourceDesc>
                        Référence bibliographique de l’article (non détaillée ici)
                  </sourceDesc>
            </fileDesc>
      </teiHeader>
      <text>
            <front>
                  <div type="abstract" xml:lang="fr|en">
                        <p>Résumé en français ou en anglais</p>
                  </div>
                  <div type="keywords" xml:lang="fr|en">
                        <p>mots-clés en français ou en anglais (séparés par des virgules)</p>
                  </div>
            </front>
            <body>
                  <div n="1" type="section" subtype="introduction|conclusion">
                        <head>Titre de la section</head>
                        <p>Texte de la section</p>

                        <div n="1.1" type="subsection">
                              <head>Titre de la sous-section</head>
                              <p>Texte de la sous-section</p>

                              […]
                              <table n="n° de la table">
                                    <head>Légende de la table</head>
                                    <row><cell/></row>
                              </table>
```

```
                        <figure n="" n° de la figure">
                              <head>Légende de la figure</head>
                        </figure>

                        <note n="" n° de la note de bas de page" place="bottom">Texte de la
                  note de bas de page
                        </note>
                  </div>
            </div>
      </body>
      <back>
            <listBibl>
                  <bibl>références</bibl>
            </listBibl>
      </back>
</text>
</TEI>
```

NOTES :

- L'identifiant de chaque article (associé à <**fileDesc**>) est unique et a une structure du type :
  - **taln-1997-long-001**
    - premier article long de la conférence TALN 1997
  - **recital-2001-poster-003**
    - 3e poster de la conférence RECITAL 2001
- Les marques de paragraphe (**<p>**) sont indicatives et correspondent généralement aux segments de texte séparés par une marque structurelle parmi les précédentes ou par un saut de page. Les véritables paragraphes du texte initial peuvent dans certains cas être identifiés en se basant sur les fins de ligne.
- Nous n'avons retenu que deux niveaux de sections (*section* et *subsection*), mais le numéro de chaque élément permet de reconstituer la hiérarchie complète si nécessaire.
- Nous avons identifié par des règles deux types particuliers de section (*introduction* et *conclusion*), à des fins expérimentales (voir Tanguy et al. 2020).
- Le contenu des tables n'a pas été préservé, même lorsqu'il était textuel. La norme TEI impose néanmoins qu'une table contienne au moins une ligne avec une cellule, ce qui explique les éléments vides.
- Les notes de bas de page sont placées dans le flux du texte là où leur corps apparaît (donc en bas de page), et non pas là où l'appel de note est inséré dans le texte.
- La liste des références bibliographiques est indiquée (**<listBibl>**) mais non analysée (pas de découpage en items). Les éléments **<bibl>** présents sont obligatoires selon la TEI mais, comme les marques de paragraphes, ils ne signalent que les segments connexes de texte contenant des références bibliographiques, et non des items bibliographiques séparés.

5/ Le corpus TALN en quelques chiffres :

Articles : **1602**
Dont :
Articles complets (avec corps du texte) : **1321**
Articles sans corps car rédigés en anglais : **100**
Articles sans corps car contenu non récupérable : **181**

Nombre total de mots ≈ **5,8 millions de mots**.
Nombre total de mots (sans bibliographie) : **4,9 millions de mots**

Nombre d'éléments identifiés

Sections et sous-sections : **14 026**
Dont Introductions : **1200**
Dont Conclusions : **1186**
Figures : **2801**
Tables : **2653**
Notes de bas de page : **4438**

Contributeurs

Ludovic Tanguy : coordination scientifique
Yoann Bard : coordination technique, développement des programmes de conversion automatique, vérification et correction
Alice Adnot-Albinet : vérification et correction
Charline Fabre : vérification et correction (année 2019)
Clémentaine Mailly : vérification et correction (année 2019)

6/ Licence d'utilisation

RÉCITAL, du 19 au 23 juillet 2010 à Montréal,
TALN, du 27 juin au 1er juillet 2011 à Montpellier,
RÉCITAL, du 27 juin au 1er juillet 2011 à Montpellier,
TALN, du 4 au 8 juin 2012 à Grenoble,
RÉCITAL, du 4 au 8 juin 2012 à Grenoble,
TALN, du 17 au 21 juin 2013 aux Sables d'Olonne,
RÉCITAL, du 17 au 21 juin 2013 aux Sables d'Olonne.
TALN, du 1er au 4 juillet 2014 à Marseille
RÉCITAL, du 1er au 4 juillet 2014 à Marseille
TALN, du 22 au 25 juin 2015 à Caen
RÉCITAL, du 22 au 25 juin 2015 à Caen
TALN, du 4 au 8 juillet 2016 à Paris
RÉCITAL, du 4 au 8 juillet 2016 à Paris
TALN, du 26 au 30 juin 2017 à Orléans
RECITAL, du 26 au 30 juin 2017 à Orléans
TALN, du 14 au 18 mai 2018 à Rennes
RECITAL, du 14 au 18 mai 2018 à Rennes
TALN, du 1er au 5 juillet 2019 à Toulouse
RECITAL, du 1er au 5 juillet 2019 à Toulouse

3) L'ATALA cède gratuitement et sans limitation de durée une licence et les droits non exclusifs de reproduction, de distribution de reproductions, de diffusion, de présentation publique, d'adaptation et d'utilisation à des fins de recherche ou d'enseignement des ACTES pour une utilisation non-commerciale.

4) L'ATALA conserve tous ses droits, copyrights, droits d'auteur, et autres droits de propriété intellectuelle ou patrimoniale afférents aux ACTES.

5) Les ACTES peuvent être téléchargés, archivés, envoyés par email ou imprimés. Il peut en être fait des copies des citations, des résumés, des textes intégraux ou partiels, à condition que les informations soient exploitées uniquement pour une utilisation non-commerciale de recherche ou d'enseignement. En aucun cas, les ACTES ne pourront être utilisé dans des produits en tant que composant ou que base de tout autre objet destiné à la vente.

6) Toute diffusion, copie, ou archive des ACTES, qu'elle soit complète ou partielle, doit contenir le présent document.

7) Toute diffusion, copie, ou archive des ACTES, qu'elle soit complète ou partielle, doit contenir des références bibliographiques au document original.

8) L'ATALA n'est en aucun cas responsable des modifications qui pourraient être apportées au contenu original des ACTES.

9) Le présent document sera appliqué et interprété conformément à la loi de l'État Français.

10) Le présent document comprend 10 articles.