



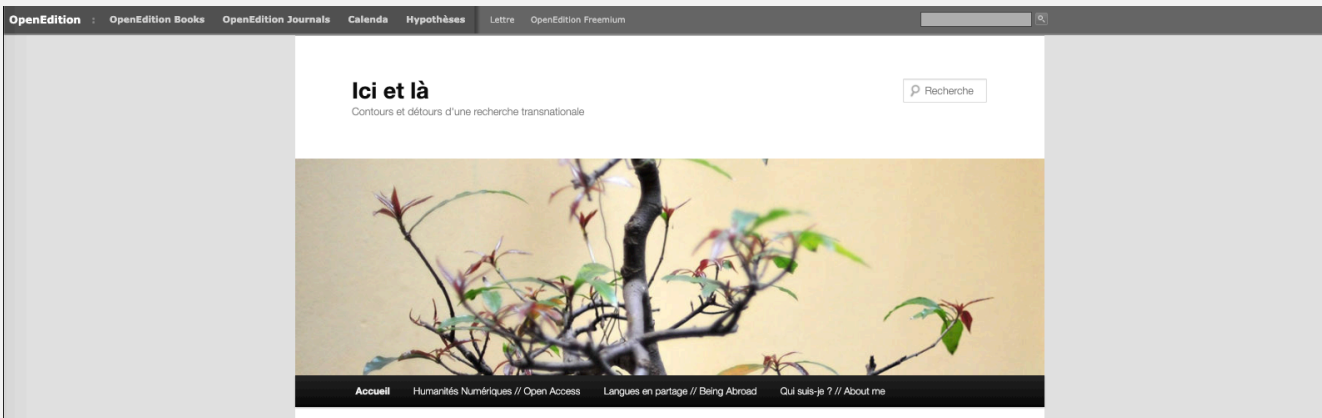
OUTILS NUMÉRIQUES POUR LES SHS

Naomi Truan, Universität Leipzig

Formation à l'Université du Mans, 11 et 12 février 2020

QUELQUES MOTS POUR ME SITUER

- Germaniste de formation, thèse entre la France et l'Allemagne (2019)
- A cheval entre **plusieurs cultures disciplinaires** (linguistique / études germaniques / études anglophones)
- Libre accès pas encore très établi en études germaniques (en France !), plus largement connu et pratiqué en linguistique
- **Pas une défenseuse du libre accès dès le début** : un processus guidé (merci à Laurent Romary !)
- Évolution **de l'indifférence à l'engagement**
- L'accès ouvert : désormais quelque chose que je fais presque automatiquement, mais toujours **à mon propre rythme** – comme pour tout, il faut trouver **ce qui marche pour vous**



Capture d'écran du carnet de recherche <https://icietla.hypotheses.org/> sur *Hypotheses*, portails pour les sciences humaines et sociales (*OpenEditions*)

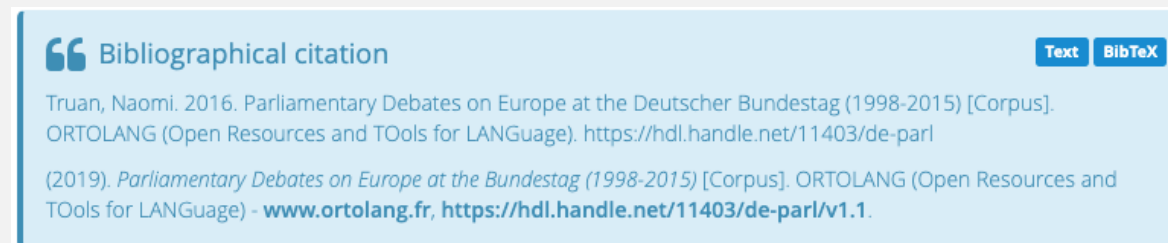


Capture d'écran de mon compte Twitter



Capture d'écran de l'archive ouverte HAL-SHS <https://cv.archives-ouvertes.fr/naomi-truan>

Naomi Truan | 02/2020 | CC BY 4.0



Capture d'écran des indications bibliographiques pour le corpus de débats parlementaires allemands en ligne sur ORTOLANG (Open Resources and TOols for LANGuage)

<https://www.ortolang.fr/market/corpora/de-parl>

Les différentes facettes de ma présence en ligne comme enseignante-chercheuse

LE PROGRAMME D'AUJOURD'HUI

CONSTITUER ET ANNOTER UN CORPUS

- Quelles **possibilités et stratégies d'annotation** ?
- Comment et pourquoi mettre à disposition (une partie de) ses données en **accès ouvert** ?
- Jusqu'où aller dans la finesse d'annotation et **comment consigner ses choix** ?
- Quels sont les enjeux d'un **partage des données** avec la communauté scientifique, quelles licences choisir ?

CONCEVOIR ET APPLIQUER UNE STRATÉGIE GLOBALE DE DIFFUSION NUMÉRIQUE

- Comment et où publier sur **quel aspect de ses données** ?
- Comment penser une **stratégie de publication cohérente** ?



CONSTITUER ET ANNOTER UN CORPUS

CORPUS VS. DONNÉES

DONNÉES

- **matérialité langagière**
(oral/écrit/distinction non applicable ?)
- **support** qui véhicule ces paroles en relation avec une situation de communication

(Charaudeau 2009: 37)

CORPUS

- données **assemblées / regroupées**
- en vue d'une **question de recherche**
- éventuellement **transcrites / annotées**
- valeur de **représentativité** du matériel recueilli
- corpus **exhaustif et clos, ou partiel et ouvert ?**

« DIS-MOI QUEL EST TON CORPUS, JE TE DIRAI QUELLE EST TA PROBLÉMATIQUE »

À L'INTÉRIEUR DU MATÉRIAU LANGAGIER

- **catégories qui vont faire l'objet de l'analyse** : **grammaticales** (connecteurs, pronoms, verbes, etc.), **lexicales** (par champs ou de façon aléatoire), **syntaxiques** (selon divers types de construction)
- **variables externes** à la production des actes langagiers, telles que les types de locuteurs et locutrices, les dispositifs de communication, de même que les variables concernant le temps (l'historicité) et l'espace (les cultures)

OUTIL DE TRAITEMENT DES DONNÉES

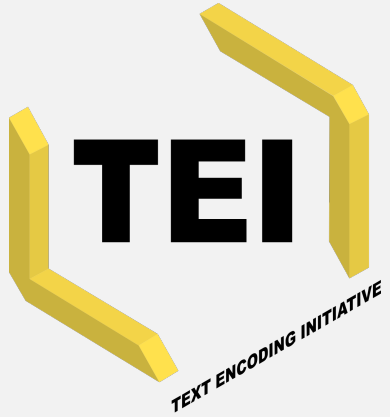
- dépouillements **manuels**
- traitement **informatique** à l'aide de logiciels *ad hoc*
- ...

(Charaudeau 2009: 38)

LA TEI EN QUELQUES MOTS

- a consortium which **collectively** develops and maintains a **standard** for the **representation of texts in digital form**
- a set of **guidelines** which specify encoding methods for machine-readable texts
- “the TEI emphasizes what is **common** to every kind of document” (Burnard 2014)
- In this sense, and despite the fact that the political context changes over time between France, Germany and the United- Kingdom, the TEI allows for a **common technical, practical and methodological framework between the three corpora** and the three languages.
- Furthermore, the TEI annotation enables to fruitfully visualise the **articulation, i.e. the continuum between text and context**. Interpretative data is situated within the corpus, which enables any researcher to see it.

<https://tei-c.org/>



LA TEI À LA SAUCE 'NAOMI' 😊

Quelques réflexions à partir de mon retour d'expérience

Pas de règles absolues, pas de recettes

CORPUS EN LIGNE

DÉBATS PARLEMENTAIRES (FR, DE, UK)

- Truan, Naomi. 2016a. Parliamentary Debates on Europe at the Assemblée nationale (2002-2012) [Corpus]. *ORTOLANG (Open Resources and TOols for LANGuage)*. hdl.handle.net/11403/fr-parl.
- Truan, Naomi. 2016b. Parliamentary Debates on Europe at the Deutscher Bundestag (1998-2015) [Corpus]. *ORTOLANG (Open Resources and TOols for LANGuage)*. hdl.handle.net/11403/de-parl.
- Truan, Naomi. 2016c. Parliamentary Debates on Europe at the House of Commons (1998-2015) [Corpus]. *ORTOLANG (Open Resources and TOols for LANGuage)*. hdl.handle.net/11403/uk-parl.

RAPPORTS ADMINISTRATIFS (FR, DE)

- Truan Naomi & Léa Renard. 2016a. Rapports allemands sur l'intégration (1991-2014) [Corpus]. *ORTOLANG (Open Resources and TOols for LANGuage)*.
- Truan, Naomi & Léa Renard. 2016b. Rapports français sur l'intégration (1991-2012) [Corpus]. *ORTOLANG (Open Resources and TOols for LANGuage)*.

I) UN CORPUS POUR QUOI, POUR QUI ?

- Important : **faire état de la question de recherche et des enjeux sous-tendant la composition et annotation du corpus**
- Dans mon cas : chacun des corpus mis en ligne peut être étudié indépendamment des autres, mais les corpus fonctionnent aussi en trio (FR, DE, UK) > **dimension contrastive a eu des conséquences non négligeables sur les choix d'annotation** (ex. le codage des interventions non autorisées au parlement)

Idéalement, la dimension contrastive serait plus explicite (je vais aller corriger sur ORTOLANG !)

This annotation has been realised as part of my doctoral work entitled “*Who Are You Talking About?*”. *The Pragmatics of Third-Person Referring Expressions. A Contrastive Corpus-Based Study of British, German, and French Parliamentary Debates* (2018).

<occupation> **Occupation**. Contains an informal description of a person’s trade, profession or occupation.

In this corpus, `occupation` mainly corresponds to the role “Member of Parliament” (abbreviation: “MP” in the three corpora).

When ~~the speaker also endorses another role which differs~~ from “MP”, it is always written in the language of the corresponding parliament since it corresponds to official functions (e.g. “Président.e de la commission des affaires européennes” or “VizepräsidentIn des Deutschen Bundestages”, for instance).

II) ETRE TRANSPARENT ·E

- Dire clairement ce que l'on pourrait s'attendre à trouver, ce qui « manque », pourquoi ça n'y est pas
- Par exemple, je n'ai pas assez thématisé l'aspect contrastif de mon corpus, et ce que ça implique en terme de (non) codage (je le fais maintenant dans le *data paper* en cours d'évaluation), et cela a pu susciter des critiques légitimes

Important notice:

The French corpus has been added after the German and British corpora had been collected and annotated. I followed the same protocol to look for the national plenary debates in France, but many debates are missing. Nevertheless, the corresponding European Councils are still in the table (see document "Description of the French Corpus"), so that the comparison with the German and British corpora remains possible.

Descriptor: speech type (debate, interruption, vote explanation, etc.)

"From a linguistic point of view, this descriptor, which is not included in the data model of the corpus provided by (Truan, 2017), is particularly important when it comes to differentiate effects of register variation ranging from highly formulaic to less formal speech (as in the case of e.g. interruptions)." (Diwersy, Frontini & Luxardo 2018)

III) UN PRINCIPE : CODER PLUS QUE CE DONT ON A BESOIN

```
<teiHeader>
[...]
```

```
<listPerson type="parliamentarians">

<person xml:id="SCHAEFER">
<persName>Axel Schäfer</persName>
<sex>male</sex>
<occupation>MP</occupation>
<affiliation>SPD</affiliation>
<trait type="party">|
<desc>Left</desc>
</trait>
```

```
floruit>no opposition</floruit>
floruit>direkt gewählt</floruit>
```

```
<residence>Böckmann 17, Nordhofen, Westfalen</residence>
<nationality>German</nationality>
</person>
```

Balise TEI codée manuellement, premier (et seul ?) corpus (à ma connaissance) à prendre en compte la variable majorité / opposition dans les débats parlementaires

<floruit> Position. Contains information about a person's period of activity.

In this corpus, `floruit` tags enable to take the variable majority/opposition into consideration. Following `floruit` tags have been used: `opposition / no opposition` (independent being considered in the opposition).

In the German corpus, the specificity of the electoral system (first vote/second vote) was added when available⁵: `direkt gewählt / gewählt über Landesliste`

Balise TEI valable seulement pour le corpus allemand (pas d'équivalent pour UK-PARL et FR-PARL) et qui s'est révélée ne jouer aucun rôle dans l'analyse (mais on n'aurait pas pu le savoir avant...)

IV) VIVE L'IMPERFECTION !

- Accepter de ne publier qu'une **version provisoire** du corpus
- Accepter d'avoir fait des erreurs
- Idéalement, **compter sur la communauté** pour nous signaler les erreurs

Corpus Annotation

Naomi Truan

(Sorbonne Université / Freie Universität Berlin)

*This document is a **slightly revised version** from May 2019 of the document released on ORTOLANG in November 2016. The title of my doctoral work (submitted in October 2018, defended in January 2019) **has been updated**. The date format has also been unified (from two tags to only one, featuring both the format “year-month-day” and the date in letters). Spelling mistakes and faulty formulations **have been corrected**.*

V) FACILITER LA RÉUTILISATION (ET CITATION) 1/2

- **Indiquer l'auteur** + la date + la licence **dans le corpus** (TEI Header)

```
<fileDesc>
  <titleStmt>
    <title/>
    <author>Naomi Truan</author>
  </titleStmt>
  <editionStmt>
    <edition>
      <date>2016</date>
    </edition>
  </editionStmt>
  <publicationStmt>
    <p>CC BY 4.0</p>
  </publicationStmt>
  <sourceDesc>
    <p>Available from <ref target="http://dipbt.bundestag.de/doc/btp/14/14014.pdf"/></p>
  </sourceDesc>
</fileDesc>
```

V) FACILITER LA RÉUTILISATION (ET CITATION) 2/2

- **Choisir une licence** (de préférence la plus ouverte possible) et l'indiquer clairement, i.e. dire : « vous avez le droit d'utiliser le corpus comme ci et comme ça »

- **Mauvaise pratique** : se contenter de renvoyer au corpus avec un **lien** (même si ORTOLANG attribue des URL pérennes)
- Reconnaître le travail de collecte et d'annotation de données : un immense chantier en SHS

The French Corpus is at your disposal under the terms of the Licence Creative Commons BY (Attribution) 4.0.

Naomi Truan 2016 – [CC BY 4.0](#)

How to quote this corpus?

Truan, Naomi. 2016. Parliamentary Debates on Europe at the Assemblée nationale (2002-2012) [Corpus]. *ORTOLANG (Open Resources and TOols for LANGuage)*. <https://hdl.handle.net/11403/fr-parl>

LA PUBLICATION DU CORPUS EN LIGNE – UN BILAN D'ÉTAPE

- **Corpus cité**, même sans 'publication' au sens strict : *data paper* soumis en juillet 2019 à un journal, toujours en cours d'évaluation (!)
- **Corpus utilisé** pour des projets très divers
- **Corpus servant de 'modèle' pour d'autres projets d'annotation** (au sens d'inspiration ou de première tentative, pas de corpus 'parfait', ce qui contredirait ma vision de la recherche comme processus)

access to parliamentary debates corpus

2 messages

To: naomi.truan@uni-leipzig.de

Dear Dr. Naomi Truan,

I am a PhD student at [redacted] University.
I am deeply interested in the corpus of parliamentary debates 1998-2015 I found on your publications.

I would be grateful if you could assist me with a reference to download the corpus. Is it please please possible?

I tried to look at HAL archives but found only the paper itself: *On the Pragmatics of Interjections in Parliamentary Interruptions*. I could not find the data :(

Thank you so much in advance,

Best

Mail d'un chercheur en Israël me demandant l'accès à
'mon' corpus de débats parlementaires britanniques, 2019

[redacted] A thematically specialized corpus such as the one prepared by Naomi Truan on the parliamentary discourse on Europe may offer significantly more detailed metadata and annotation (Truan, 2017).

Citation avec renvoi vers la citation ORTOLANG (avec une erreur de date...), in : Blätte & Blessing 2018

A user would like some additional informations about the resource uk-parl

Hi,

A user let you a message in relation with a resource that you manage:

[Parliamentary Debates on Europe at the House of Commons (1998-2015)]

Hi Naomi I would like to access the corpus for my research. We have compiled the Malaysia Hansard Corpus and is currently working on the mark-up. thank you. Prof. [redacted]

You can answer directly by answering this email sender address.

Sincerely,
The ORTOLANG Team.

Mail d'un chercheur en Malaisie me demandant l'accès à
'mon' corpus de débats parlementaires britanniques, 2019

Title of the project: Doing 'Interrupting' in parliamentary debates in British English, Finnish French, and German: A cross-linguistic perspective

The data for this project stem from the **parliamentary corpora** in **British English, Finnish, French, and German** taken from open access resources made available by several contributors and listed on the CLARIN infrastructure. The data consist of four mostly manually annotated corpora with TEI-XML markup, offering a ready-to-analyse dataset.

Un début de collaboration avec une chercheuse en Finlande qui travaille sur des débats parlementaires et utilise 'mes' corpus pour de nouvelles questions de recherche

Les retours sur mon corpus trois ans après sa publication en accès ouvert



Je te donne, tu me donnes, nous nous donnons... nos données

J'ai eu l'occasion, le 1^{er} juin 2017, de présenter en une minute mes réflexions au cours de la table ronde « **Publier à l'ère numérique** » co-organisée par le Centre Marc Bloch, DARIAH-EU, le bureau de la coopération universitaire de l'Institut français d'Allemagne à Berlin et l'Université franco-allemande (UFA) – merci à Anne Baillot et à Laurent Romary pour l'invitation ! Bien que nos interventions aient été filmées [1], je voudrais ici approfondir les trois grands points que j'ai développés à cette occasion afin de souligner, une fois encore, à quel point le mouvement de l'accès ouvert (*open access*) représente une avancée certaine pour l'état de nos connaissances.

<https://icietla.hypotheses.org/53>

A photograph of two white-framed, double-hung windows set into a textured, terracotta-colored wall. The windows are open, with the shutters swung outwards. The scene is brightly lit, suggesting a sunny day. The windows look out onto a dark, possibly wooded area.

Image de Martin Pyško
sur Pixabay

CONCEVOIR ET APPLIQUER UNE STRATÉGIE GLOBALE DE DIFFUSION NUMÉRIQUE

... et (s')ouvrir des portes et des fenêtres

UN REGARD INCLUSIF (HOLISTIQUE) SUR NOS TRAVAUX ET NOTRE TRAVAIL

NB : Certaines des réflexions présentées ici ont été développées pour la première fois en anglais en janvier 2019 : <https://icietla.hypotheses.org/994>



- Penser une stratégie *globale* de diffusion numérique, c'est...
 - prendre en compte **tous les produits d'un projet de recherche**, c'est-à-dire **pas seulement les résultats** (qui donnent lieu à « publication » au sens étroit)
 - considérer la recherche comme un **processus**
 - rendre la recherche **reproductible**
 - une position plus personnelle : **diminuer la pression (*publish or perish*)** en reconnaissant que l'on fait **déjà** de la recherche lors de la collecte de données, peu importe ce que l'on en fait ensuite, si les données peuvent être intéressantes pour notre question de recherche ou non à ce stade

QU'EST-CE QUI « COMPTE » COMME PUBLICATION ?

- les **données brutes**
- les **données annotées**, qui doivent être complétées par des documents expliquant le cadre d'annotation et/ou un *data paper* (déjà une publication au sens strict)
- les **réflexions intermédiaires** que vous développez sur vos données (ou sur le processus de recherche en général), qui peuvent être formalisées sous forme de billets de blog
- les **articles** (évalués par des pairs ou non) menant une analyse sur vos données
- les **corrections ou réflexions supplémentaires** que vous souhaitez peut-être ajouter après les résultats finaux, qui pourront être publiés sous forme d'articles, mais aussi, une fois encore, dans des billets de blog

QU'EST-CE QUI « COMPTE » COMME PUBLICATION ?

- En allant plus loin, on peut considérer que **tout ce qui est lié à votre recherche** et que vous dites (!) ou écrivez est déjà une 'publication' :
 - considérer la pratique consistant à citer quelqu'un 'e comme une **« communication personnelle »** dans les articles de revues de linguistique
 - les **tweets/threads** comme nouvelle forme de publication ?
 - les **enseignements** en tant que recherche en cours de réalisation ? Ils peuvent se transformer en manuels (= publications) et/ou influencer plus ou moins directement votre recherche

Quel type de données ?	Quel type de plateforme ?
<p>Données ‘brutes’</p> <ul style="list-style-type: none"> ✓ Choisir une licence (la plus ouverte possible) 	<p>Archive ouverte dédiée aux dépôts de données (par ex. ORTOLANG : https://www.ortolang.fr/)</p>
<p>Données annotées / constituées en corpus :</p> <ul style="list-style-type: none"> ✓ Choisir une licence ✓ Expliquer comment et pourquoi (pour répondre à quel projet de recherche) les données ont été annotées ✓ Attribution : citer toutes les personnes impliquées (y compris si elles ne font pas partie de la citation à proprement parler) 	<p>Archive ouverte dédiée aux dépôts de données</p> <ul style="list-style-type: none"> ✓ Pas sur une plateforme privée (blog) ✓ Vos données doivent être archivées avec des mots clés et être facilement accessibles à la communauté universitaire (voire au-delà) <div data-bbox="1467 668 2170 843" style="border: 1px solid #ccc; padding: 10px; margin-top: 20px;"> <p>Contributors</p> <div style="display: flex; justify-content: space-around; align-items: center;"> <div style="text-align: center;">  <p>Naomi Truan Annotator</p> </div> <div style="text-align: center;">  <p>Laurent Romary (INRIA) Consultant</p> </div> </div> </div>

- Faites-le dès que possible pour garantir **reconnaissance** et **p/maternité** de vos travaux
- N'ayez pas peur que quelqu'un vous 'vole' quelque chose : la probabilité qu'une autre personne ait la ou les mêmes questions de recherche sur le même corpus est extrêmement faible...
- Il est plus probable que vous développiez des **projets de coopération** !

Quel type de données ?

Réflexions intermédiaires
Réflexions méthodologiques
Corrections après coup
Conférences / Séminaires

Articles

- ✓ Mettre en ligne **tous** les articles publiés, même s'il ne s'agit pas du texte intégral de l'éditeur/-trice (*preprint*)
- ✓ Sur HAL-SHS, il est possible d'indiquer quand la version complète d'un document doit être publiée (pour respecter l'embargo de max. 12 mois en SHS)
- ✓ En tant qu'auteur/-trice, vous avez toujours tous les droits sur un document, la SEULE chose que vous ne pouvez pas mettre en ligne est la version finale éditée, souvent avec la pagination

Quel type de plateforme ?

Blog de recherche

Hypothèses a une fonction « comment citer »

Cite this article as: Naomi Truan, "Some Useful Free Online Resources When You're Writing a PhD," in *Ici et là*, 07/10/2018, <https://icieta.hypotheses.org/383>.

Archive ouverte du type HAL-SHS

- ✓ Toujours demander à l'éditeur/-trice s'il l'autorise :
 - montre aux éditeurs/-trices qu'il est nécessaire d'avoir un accès libre/ouvert (politique de la demande)
 - certains e-s peuvent vous donner des réponses ambiguës, utilisez-les à votre avantage



The Day I Removed my Publications from Academia & Research Gate

Well, this is it. **I have decided to remove all my publications from Academia and Research Gate** (and no, I will not insert a link to those sites).

Sometimes, there are things you've been wanting to do for a long time, but you didn't feel like it, or you were a bit afraid of the consequences [1]. In my case, I thought that I didn't have the choice as a PhD student; I've been told that I need to be visible online and that scholars love those platforms for their simplicity and their ergonomics. I thought that if my community uses Academia and Research Gate, I had to follow their rules in order to be (accepted as / regarded as) a full member of this community. It is true that for many young scholars who do not have access to a page of the websites of their universities, Academia and Research Gate offer simple ways to publicly present your research.

<https://icietla.hypotheses.org/114>

L'ACCÈS OUVERT EN PRATIQUE (1/2)

- montrer comment la/une recherche **évolue dans le temps** (tout en ayant publié en ligne, et donc en étant tenu·e pour responsable si vous changez d'avis, si vous dites les choses différemment, etc.)
- la recherche comme un **cycle/processus, plutôt qu'un produit fini/absolu**
- s'adresser à **différentes communautés** en fonction de la partie de votre recherche que vous décidez de mettre à disposition

L'ACCÈS OUVERT EN PRATIQUE (2/2)

- Vous disposez d'outils pour **parler de vos corpus dans d'autres publications (plus 'classiques')** :
 - Vous n'avez **pas besoin de présenter vos données de manière détaillée** dans chaque article de revue si vous avez une référence à citer (par exemple, un *data paper*, un lien vers un dépôt ouvert)
 - Cela vous permet de **gagner de la place** et les articles ne deviennent qu'une des étapes, celle où vous présentez **une interprétation de vos données**
 - Vous donnez à vos lecteurs/-trices la possibilité de **« regarder derrière les rideaux »** s'ils ne sont pas d'accord avec certaines de vos conclusions
 - Tout le monde peut **comparer votre article à l'ensemble du corpus**
 - Vous soulignez le fait que chaque article repose sur **une sélection de données**, mais que la vue d'ensemble est également facilement accessible

RÉFÉRENCES

- Blätte, Andreas & Andre Blessing. 2018. The GermaParl Corpus of Parliamentary Protocols. *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference*, 810–816. Miyazaki. <http://www.lrec-conf.org/proceedings/lrec2018/pdf/1024.pdf> (8 February, 2018).
- Burnard, Lou. 2014. The TEI and XML. What is the Text Encoding Initiative? How to add intelligent markup to digital resources. (Encyclopédie numérique). Marseille: OpenEdition Press. <http://books.openedition.org/oep/680> (1 July, 2016).
- Charaudeau, Patrick. 2009. Dis-moi quel est ton corpus, je te dirai quelle est ta problématique. *Corpus* (8). 37–66.
- Diwersy, Sascha, Francesca Frontini & Giancarlo Luxardo. 2018. The Parliamentary Debates as a Resource for the Textometric Study of the French Political Discourse. *Proceedings of the ParlaCLARIN@ LREC2018 workshop*, 6. Miyazaki.