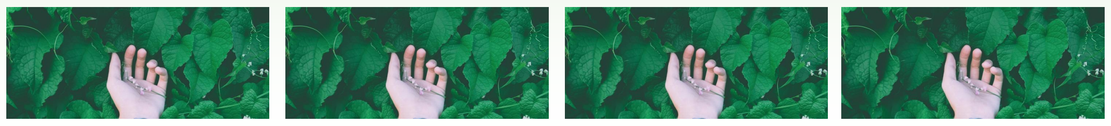
Le corpus COP21

Caroline Rossi, Camille Biros & Achille Falaise

Univ. Grenoble Alpes (ILCEA4) et CNRS (LLF)



LE CORPUS

	Nombre de mots	Nombre de documents	Organisations
Nations-Unies (ONU) EN	4 957 734	26	GIEC, UNFCCC, REDD+, UNDP, UNEP, WB
Nations-Unies (ONU) FR	1 490 661	27	GIEC, UNFCCC, REDD+, UNDP, UNEP, WB
Nations-Unies (ONU) SP	1 661 067	27	GIEC, UNFCCC, REDD+, UNDP, WB
ONG EN	2 786 564	103	350, CAN, CARE, CISDE, Climate Institute, Carbon Trade Watch, EcoEquity, EDJ, EJF, FoE, Climate Justice Alliance, Greenpeace, IBON, IICAT, International Trade Union Confederation, MRF, NRDC, Australian Youth Climate Coalition, Orataio, Oxfam, Rainforest Alliance, WDM, WRI, WWF
ONG FR	592 633	44	Attac, Equiterre, Réseau Action Climat, WWF, Greenpeace, Carbon Trade Watch, Oxfam, Noé21
ONG SP	834,348	59	350, CDKN, Amigos de la tierra, CAN, CISDE, Energia sin frontera, Greenpeace, IBON, ITUC, Oxfam, Terram, WWF
Presse EN	423 444	428	The Telegraph, The Guardian, The New York Times, USA Today
Presse FR	342 163	389	Le Figaro, Le Monde, Le Parisien, Libération
Presse SP	557,119	804	La Vanguardia, El Pais, ABC, El Comercio, La Tercera, El Nuevo Diario, El Tiempo, La Prensa, La Republica, Diario Libre, Clarin, La Razon, La Nacion, El Nacional, El Universal VZ, Juventud Rebelde, Nacion, Hondudiario, Granma, Diario de Cuba, El Nuevo Herald, El Universal MX
Entreprises EN	243 022	43	Acore, Ebico, Ecotricity, Good Energy, IRENA, REA, Clean Choice Energy, Green Energy, Good Energy UK, LoCO2
Entreprises FR	103,605	32	CLER, Enercoop, ENR, IRENA, Direct Energy, Planète OUI, Energem, Engie

A propos du choix des **textes** et des organisations

Objectif : analyse de la variation dans les discours sur la justice climatique

Étapes de constitution :

1. Recherche de sources et identification de la COP21
2. Communautés de discours identifiées pour leur adoption (+/-) du terme
3. Recherche web par mot clé au sein de ces communautés

UN CORPUS MULTILINGUE SUR LA JUSTICE CLIMATIQUE

Questions de recherche et travaux

1. Evolutions diachroniques (COP15 puis COP21) : nuages de mots clés



Figure 1. Corpus ONU COP15 EN



Figure 2. Corpus ONU COP21 EN

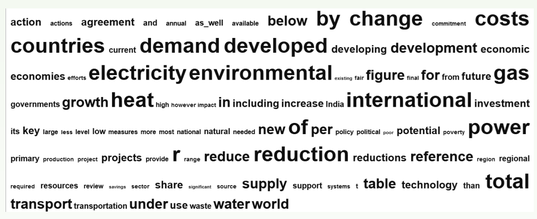


Figure 3. Corpus ONG COP15 EN

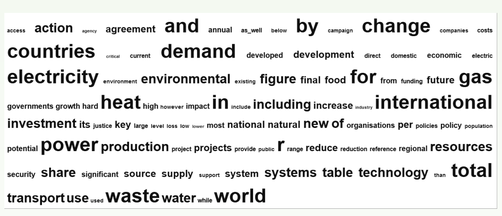


Figure 4. Corpus ONU COP21 EN

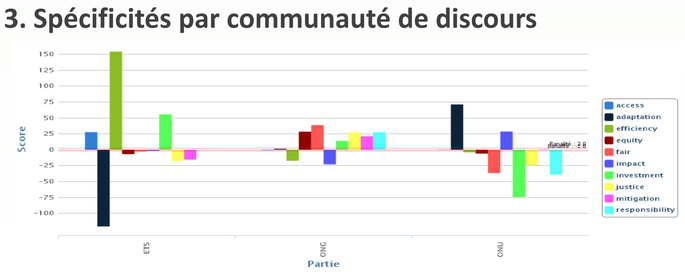
Nuages obtenus dans Wmatrix (Rayson, 2008)

2. Comparaison du corpus maison avec un corpus web BootCat (Baroni & Bernardini 2004).

Permet d'étendre le corpus sur certaines thématiques, utile surtout pour la terminologie

Ne permet pas de travailler sur les contextes temporels ou discursifs (les textes récupérés sont liés à une URL seulement)

3. Spécificités par communauté de discours



Les indices de spécificité calculés dans TXM (Heiden et al. 2010) correspondent bien aux indices de répartition par document et par organisation (R^2 = 0,8521)

LES PROBLÈMES

Contrôle, nettoyage et formatage du corpus

Le corpus est constitué de documents PDF ou de données web (articles de presse) convertis automatiquement en texte. Nous n'avons pas pu recruter la personne pressentie pour contrôler, nettoyer et formater ce corpus. Par conséquent, nous avons eu recours à des méthodes automatiques pour effectuer ce travail. Les traitement ont consisté à:

- sélectionner des documents identifiables (29% des documents ont été écartés pour des erreurs de format des métadonnées),
- identifier le contenu textuel (par opposition au paratexte: sommaires, en-têtes, pieds de page, etc.)
- normaliser grossièrement ce contenu textuel (codage des caractères),
- appliquer des heuristiques simples (unités commençant par une majuscule et finissant par un point, absence de caractères aberrants) pour détecter les unités textuelles comportant peu ou pas d'erreurs (66% du texte a ainsi dû être écarté).

Droits d'auteur

Pour partager les données, qui sont toutes librement accessibles sur le web mais sous droit d'auteur, nous avons besoin de l'accord de chacune des organisations listées dans le tableau. Les demandes sont en cours, certaines ont déjà accepté.

LA SUITE

Prise en compte de la dimension multilingue

Standardisation terminologique et traduction des rapports du GIEC (5 résumés à l'intention des décideurs, corpus parallèle)

Variation terminologique dans le corpus comparable

Amélioration du nettoyage et du formatage du corpus

Nous prévoyons d'améliorer le formatage des méta-données afin de pouvoir utiliser tous les textes, et d'améliorer la normalisation, en particulier en ce qui concerne les mots coupés, collés, et les erreurs d'OCR classiques, en réutilisant certains outils développés pour le projet ANR Presto.

Partage sur Ortolang et dans Scienquest.

BIBLIOGRAPHIE

I. Travaux sur le corpus COP21

Biros Camille, Rossi Caroline et Inesa Sahakyan (sous presse). "Discourse on climate and energy justice: a comparative study of Do It Yourself and Bootstrapped corpora" *Corpus.* Numéro special "Petits corpus" (dir. C. Danino).

Rossi Caroline, Biros Camille et Hélène Schmutz (accepté). The case for environmental justice: tracking variation in do-it-yourself corpora. *Studia Neophilologica.*

Biros Camille et Rossi Caroline (soumis). Nommer les Victimes du Changement Climatique Approche variationniste sur un corpus bilingue. *Mots, les langages du politique.*

Biros Camille & Peynaud Caroline (soumis). "Quoting and Evaluating Climate Change Knowledge, Representation of the IPCC in three types of discourse". *Lingua e Linguaggi*.

Biros Camille, Caroline Rossi & Aurélien Talbot, "Translating the International Panel on Climate Change Reports", TRANSIUS Conference, 18th & 19th of June 2018, Geneva.

II. Principales références utilisées pour construire et analyser le corpus

Baker P. (2012) Acceptable bias? Using corpus linguistics methods with critical discourse analysis, *Critical Discourse Studies*, 9:3, 247-256.

Baker P. (2006) *Using Corpora in Discourse Analysis*, London : Bloomsbury Academic.

Baroni, M., & Bernardini, S. (2004) BootCaT: Bootstrapping Corpora and Terms from the Web. In *LREC* (p. 1313).

Heiden S., Magué J-P. et Pincemin B. (2010) TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement , in *JADT 2010 : 10th International Conference on the Statistical Analysis of Textual Data*, Rome, Italie : ENS-Lyon, p. 12 p. http://halshs.archives-ouvertes.fr/docs/00/54/97/79/PDF/Heiden_al_jadt2010.pdf.

Haider, A. S. (2017) Using Corpus Linguistic Techniques in Critical Discourse Studies: Some Comments on the Combination. In *A Corpus-assisted Critical Discourse Analysis of the Arab Uprisings: Evidence from the Libyan Case* : A thesis submitted in partial fulfilment of the requirements for the degree of doctor of philosophy in linguistics. Unpublished PhD. Department of Linguistics. University of Canterbury.

Grundmann R. & Krishnamurthy R. (2010) The discourse of climate change: A corpus-based approach, *Critical Approaches to Discourse Analysis across Disciplines* Vol 4 (2): 125 – 146.

Rayson, P. (2008). From key words to key semantic domains, *International Journal of Corpus Linguistics* 13(4) : 519-549.

REMERCIEMENTS

- Les travaux d'ingénierie d'Achille Falaise ont été finances par le **consortium CORLI,** qui nous a également permis de rémunérer quatre stagiaires qui ont complété la première collecte
- L'hébergement temporaire et le partage des données se fait sur ShareDocs (**TGIR HumaNum**)
- Les premières analyses ont été faites grâce au logiciel **TXM** et nous remercions Serge Heiden qui a formé l'équipe
- **DémarreSHS!** a également financé des vacations pour nous aider à préparer les données
- Thierry Nallet et Sandrine Rol (ILCEA4) ont participé activement à la constitution du corpus espagnol et Caroline Peynaud à la constitution du corpus presse anglais et français.