Premier inventaire (non exhaustif) des corpus avec annotation de haut niveau pour le français écrit contemporain disponibles ou interrogeables en ligne

Assemblée générale de l'Infrastructure de recherche Corpus, Consortium Corpus Ecrits, 23 novembre 2012

Amalia Todirascu (LILPA, Université de Strasbourg), Agnès Tutin (LIDILEM, Université de Grenoble)
N'hésitez à signaler d'autres ressources (écrire à todiras@unistra.fr ou agnes.tutin@u-grenoble3.fr)
Site du groupe : https://listes.cru.fr/wiki/corpus-ecrits/public/groupe-8

Type d'annotation		Nom	Brève description	Références
Annotations syntaxiques	Corpus annotés	**French Treebank**	• Extraits du Monde • correction manuelle • constituants et fonctions (sujets, objets etc.), fonctions de surface • compatibilité avec plusieurs analyseurs Disponible pour la recherche sur demande http://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-fr.php	Abeillé, A., L. Clément, and F. Toussenel. 2003. `Building a treebank for French', in A. Abeillé (ed) *Treebanks* , Kluwer, Dordrecht.
		Séquoia	Multi-domaine, multi-genre, **libre de droit** constituants et dépendances https://www.rocq.inria.fr/alpage-wiki/tiki-index.php?page=CorpusSequoia	Candito M.-H. and Djamé Seddah, 2012, Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical, *Proceedings of TALN'2012*, Grenoble, France.
		Ecritures	corpus constitué de plusieurs états (brouillons) de rapports sociaux rédigés dans le cadre de la protection de l'enfance (300000 mots). Annotation syntaxique ad hoc avec Le Trameur. Projet en cours Disponible sous conditions (corpus sensible) http://www.univ-paris3.fr/anr-ecritures/	**CISLARU Georgeta, SITRI Frédérique**, "De l'émergence à l'impact social des discours : hétérogénéités d'un corpus", **Langages, N°187 (3/2012)**
		CLASSYN	Corpus constitué des articles scientifiques et articles de vulgarisation dans les domaines de la médecine et de l'informatique. Annotation syntaxique automatique (analyseur en dépendances B.Bohnet). Disponible sur demande à todiras@unistra.fr	**Todirascu, A., Pado, S., Krisch, J., Kisselew, M. Heid,** U. French and German Corpora for Audience-based Text Type Classification, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)
	Corpus interrogeables en ligne	**L'arboratoire** (interrogeable en ligne)	Syntaxe de contraintes. Textes français interrogeables en ligne (Europarl) http://corp.hum.sdu.dk/arboratoire.html	Salmon-Alt S., Bick E. , Romary L., Pierrel J.M.,« La FReeBank : vers une base libre de corpus annotés », *Actes de TALN 2004*, 18-23 avril 2004.
		Scientext (interrogeable en ligne)	(écrits scientifiques interrogeables en ligne) (analyse avec Syntex, syntaxe de dépendance, Bourigault) http://scientext.msh-alpes.fr/	Falaise A., Tutin A., Kraif O. (2012), Une interface pour l'exploitation de corpus arborés par des non-informaticiens : la plate-forme ScienQuest du projet Scientext, *Traitement Automatique des Langues* 2011

					Volume 52 Numéro 3, http://www.atala.org/IMG/pdf/Falaise-TAL52-3.pdf
Annotations sémantiques	Corpus annotés	Entités nommées	**Etape** **Ester2**	Corpus utilisé dans des campagnes d'évaluation des campagnes d'évaluation, annotation des personnes, lieux, organisation (transcription d'émissions de radios) **Diffusé par ELRA**	Sylvain Galliano, Edouard Geoffrois, Djamel Mostefa, Khalid Choukri, Jean-François Bonastre, Guillaume Gravier: The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. *INTERSPEECH 2005*: 1149-1152 http://www.afcp-parole.org/etape/docs/interspeech05-ester.pdf
		Sentiments et opinions	**Blogoscopie**	Blogoscopie (ANR) (piloté par le LINA) : annotation des blogs avec les concepts et les évaluations. **Téléchargeable après accord** http://www.lina.univ-nantes.fr/?Blogoscopie,762.html	DAILLE B., DUBREIL E., MONCEAUX L. ET VERNIER M. Annotating opinion—evaluation of blogs : the Blogoscopy corpus. *Language Resources and Evaluation* Springer 2011.
		Expressions temporelles	**French TimeBank**	Repérage des entités temporelles et des relations entre entités dans un corpus de textes journalistiques (Est Republicain) **Librement téléchargeable** https://gforge.inria.fr/projects/fr-timebank/	Bittar A., Amsili P., Denis P. (2011), French TimeBank : un corpus de référence sur la temporalité en français, *TALN 2011 - Traitement Automatique des Langues Naturelles*, (2011) 259-270 http://www.linguist.univ-paris-diderot.fr/~amsili/papers/TALN-2011-French-TimeBank.pdf
		Désambiguïsation	**Romanseval**	Corpus de la campagne ROMANSEVAL (60 mots désambiguïsés). **Diffusé par ELRA**.	http://aune.lpl.univ-aix.fr/projects/romanseval/
	Corpus interrogea-ble en ligne	Nominalisation	**Nomage**	Corpus avec annotation des nominalisations (French Treebank) **http://nomage.recherche.univ-lille3.fr/nomage/**	Balvet A., Barque L., Condette M.-H., Haas P. , Huyghe R., Marín R., Merlo A. (2012). La ressource Nomage, Confronter les attentes théoriques aux observations du comportement linguistique des nominalisations en corpus, *TAL*, Volume 52 – n° 3/2012, pp. 1-24. http://www.atala.org/IMG/pdf/Balvet-TAL52-3.pdf
Annotations textuelles et discursives	Corpus annotés	Annotations textuelles	**Annodis**	Relations de discours entre les unités minimales de discours, macro-structures, notamment les structures énumératives, chaînes topicalisées. Corpus variés. **Librement disponible** : http://redac.univ-tlse2.fr/corpus/annodis/	Péry-Woodley M.-P., Afantenos S. D., Ho-Dac L.-M., Asher N. (2011). La ressource ANNODIS, un corpus enrichi d'annotations discursives. *TAL 52(3)*, pp 71-101. http://redac.univ-tlse2.fr/corpus/annodis/biblio/peryEtAl2011-TA52-3.pdf

			Géopo	Corpus de 270 000 mots d'articles expositifs autour des relations internationales. Structure textuelles, sections, énumérations. **Librement disponible** : http://redac.univ-tlse2.fr/corpus/geopo.html	Lydia-Mai Ho-Dac : *La position initiale dans l'organisation du discours : une exploration en corpus*, thèse de doctorat, Université de Toulouse-Le Mirail, novembre 2007.
		Annotations discursives	**Lelie**	Corpus de textes du genre procédural (procédures, exigences, textes didactiques) 8000 textes procéduraux en français et environ 1500 textes procéduraux en anglais, 2000 exigences français et 1200 en anglais. Annotations des relations du discours, structure des verbes Adresse du projet : http://www.irit.fr/recherches/ILPL/lelie/accueil.html	
			French Discourse Treebank	Objectif : construire un corpus similaire à Penn Discourse Treebank. Analyse des connecteurs inspirée de SDRT et RST (en cours). Projet porté par ALPAGE	
			Projet COMTIS	Extraits d'Europarl annotés avec connecteurs de discours www.idiap.ch/dataset/europarl-direct/	Popescu-Belis A., Meyer T., Liyanapathirana J., Cartoni B. & Zufferey S. 2012. *Discourse-level Annotation over Europarl for Machine Translation: Connectives and Pronouns*. Proceedings of LREC 2012, May 23-25 2012, Istanbul, Turkey.
		Coréférence et anaphores	**Corpus d'expressions anaphoriques grammaticales**	Expressions anaphoriques grammaticales (presse, écrits scientifiques, ouvrages scientifiques). 1 million de mots **Diffusé par ELRA**	Tutin A., Trouilleux F., Clouzot C., Gaussier E., Zaenen A., Rayot S., Antoniadis G. (2000). Annotating a large corpus with anaphoric links. *Proceedings of DAARC 2000*. Lancaster 16-18 November 2000 http://halshs.archives-uvertes.fr/docs/00/37/33/27/PDF/tutin_daarc2000.pdf
			Dédé	Corpus de descriptions définies. **Librement disponible** : http://www.cnrtl.fr/corpus/dede/	Gardent C., Manuélian H. (2005). Création d'un corpus annoté pour le traitement des descriptions définies. *TAL*. 2, pp 1-15 http://www.loria.fr/~gardent/publis/tal05-dede.pdf
			EvalRefGen	Corpus annoté en chaînes de référence (multigenre) **Librement disponible** sur demande à l'auteur todiras@unistra.fr	Laurence Longo, Amalia Todirascu: RefGen: a Tool for Reference Chains Identification. IMCSIT 2010: 447-454
			Projet MC4	Etude de chaines de coréférence dans des textes du français médiéval et contemporain, dans les textes narratifs et non-narratifs. Projet en cours.	