LES PRINCIPAUX DOMAINES DE L'ANALYSE LINGUISTIQUE ET QUELQUES EXEMPLES D'APPROCHES D'ANALYSE AUTOMATIQUE DU LANGAGE

Dans ses différentes dimensions, orales et écrites

Domaine	**unité**	**Problématiques**	**Outils / bibliothèques / ressources**
Phonétique	Phone	Extraire corrélats du signal (DURÉE, FRÉQUENCE, formants) Outils : Aligneurs texte/son Annotation (token : alignement, phonétisation, syllabe, Prosodie)	Praat : analyse du spectre praat.org • align data • extract data • analyse data • train model EXAMPLE Adrien Méli's PhD http://mapage.noos.fr/admeli/soutenance/#/ Se pilote avec python via la bibliothèque Parselmouth https://parselmouth.readthedocs.io/en/stable/ https://github.com/YannickJadoul/Parselmouth SPPAS sppas.org (représentation en SAMPA) (se pilote aussi en python) ➔ aligneur, phonétiseur ➔ analyse des syllabes ➔ analyse de la prosodie (MOMEL-INTSINT) ###ALIGNERS https://eleanorchodroff.com/tutorial/index.html

		Reconnaissance vocale (VR) TTS text to speech resynthèse vocale	forced-alignment-tools, A collection of links and notes on forced alignment tools by Alberto Pettarin P2FA (aligneur de Penn) PyTorch-NLP • Audio I/O and Pre-Processing with torchaudio • Speech Command Recognition with torchaudio #CMU Resources (part of the NLTK toolkit) Dedicated python packages: Pronouncing MISC Phonetic conversion (and text to speech) Tacotron
Phonologie	Phonème	faire l'inventaire des systèmes phonologiques, des syllabes possibles, étudier les correspondances (sons/lettres :	PHON (Y. Rose) Algorithmes de découpages en syllabes https://www.phon.ca/phon-manual/getting_started.html Algorithmes de découpages en syllabes dans SPPAS (sppas.org) API en SAMPA API en ARBANET Ressources : le CMU http://www.speech.cs.cmu.edu/cgi-bin/cmudict

		grapho-phonématique) (OCR)	(disponible dans la bibliothèque NLTK) # le CELEX (décomposition en morphèmes, transcription, fréquence, ressource payante pour l'anglais, se méfier du découpage en syllabes) #French: A lexical database for contemporary French : LEXIQUE™ A lexical database Pallier et al. 2001
Morphologie	morphème	Détecter les morphèmes *Singer / stronger* Morphème dérivationnel Morphème flexionnel	Pour l'anglais Linguistica Pour des analyses non-supervisées : Chipmunk {-er} [sing]$_{V}$ / [singer]$_{N\ /\ 2\ types\ (25}$ lemmas) {strong, stronger} [strong]$_{A\ 2\ tokens}$ /1 lemma Pour le français: LEFF:Morphological and syntactic lexicon for French http://pauillac.inria.fr/~sagot/index.html#wolf LEXIQUE (et fréquence) http://www.lexique.org/ http://www.lexique.org/shiny/openlexicon/
Lexicologie	Lexème (forme de mot)	Retrouver les lemmes Tokenisation	BNC (100M) CLAWS BNC2014 COCA (350M) BLP SUBTLEX US

		Analyse des stopword Identifier les collocations	Ressources pour la fréquence http://crr.ugent.be/papers/SUBTLEX-US%20frequency%20list%20with%20PoS%20and%20Zipf%20information.zip l'échelle de Zipf : http://crr.ugent.be/archives/1352 Tokeniser (tokenisation) / lemmatisation (lemma/type) TTR NLTK / wordnet Treetagger PARSEME (fr) Multi-word expression / MWE / MWU
Syntaxe	Syntagme Propositions syntaxiques	Pos-tagging PARSING Universal Dependencies	Treetagger / NLTK coreNLP Spacy Stanford corenlp (JAVA) Démo en ligne : https://corenlp.run/ UDpipe https://github.com/ufal/udpipe UD models

			https://universaldependencies.org/#language-u (Sequoia) https://github.com/UniversalDependencies/UD_French-Sequoia (Partut) Plus fin : https://surfacesyntacticud.github.io/conversions/ SUD Surface Syntactic Universal Dependency https://surfacesyntacticud.github.io/ analyse en coréférence
Sémantique	Sème/ sémantème	WORDNET (anglais) WOLF (français) détection des synsets relations Détecter les noms propres, les références au lieu et au temps NER	 https://wordnet.princeton.edu/ http://wordnetweb.princeton.edu/perl/webwn le « wordnet français », INRIA, équipe de Benoit Sagot : http://blog.onyme.com/etude-de-lontologie-wordnet-libre-du-francais-wolf/ http://alpage.inria.fr/~sagot/wolf.html représentations ontologiques : maison http://blog.onyme.com/wp-content/uploads/2013/09/wolf-0.1.5-jouet.png Encore des difficultés : http://blog.onyme.com/wp-content/uploads/2013/09/wolf-1.0b-maison-wrong.png NER (Named Entity Recognition) NLTK, Spacy

		word embedding..	Hugging-face (BERT, ALBERT, ROBERT .... FlaubERT) CamemBERT,BioBert word similarities:cosine, npmi... Glove,Word2vec
Pragmatique	Acte de langage (Austin , Grice)	Que veut-il dire quand il utilise telle forme/telle phrase ? pertinence Calcul d'implicature Maximes de la conversation Principe de coopération	L'analyse du sentiment (sentana : sentiment analysis) https://corenlp.run/ Chatbot …

Pour aller plus loin:

Emily M. Bender (2013) *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax*, Morgan & Claypool

Emily M. Bender & Alex Lascarides (2020*) Linguistic Fundamentals for Natural Language Processing II. 100 Essentials from Semantics and Pragmatics*, Morgan & Claypool

OUTILS POUR L'ANALYSE AUTOMATIQUE DE LA COMPLEXITE SYNTAXIQUE ET LEXICALE

La page de Xiaofei Lu
http://www.personal.psu.edu/xxl13/download.html
(voir aussi son analyseur D-Level Analyzer)

Pour la complexité lexicale
Lexical Complexity Analyzer

L2SCA : L2 Syntactic Complexity Analyzer
http://www.personal.psu.edu/xxl13/downloads/l2sca.html

La famille d'outils de Kyle & Crossley:

Les outils d'analyse de la sophistication lexicale (TAALES), de la complexité (TAASC), de la cohésion (TAACO) mais aussi pour l'analyse automatique du sentiment (SEANCE)
L'interface est très intuitive et multi-OS (pour des raisons de d

pour l'analyse automatique du sentiment : SEANCE
https://www.linguisticanalysistools.org/seance.html

article de synthèse : https://hal.archives-ouvertes.fr/hal-02768504v3/document

DES RESSOURCES LINGUISTIQUES A L'ASSISTANCE A LA REDACTION EN ANGLAIS (*COMPUTER-AIDED WRITING*)

Votre moteur de recherche préféré peut être utile (utilisez les guillemets), mais vous aurez aussi des surprises

Pour vérifier que le mot, la collocation existe et pour vérifier sa construction : les corpus de références comme le COCA (350M mots, américain) et le BNC (100M mots, britannique)
https://www.english-corpora.org/
https://www.english-corpora.org/coca/
https://www.english-corpora.org/bnc

A partir de votre texte rédigé en français, utiliser un moteur de traduction, https://www.deepl.com/fr/translator
puis vérifier vos doutes éventuels avec votre moteur de recherche préféré

!! Attention au plagiat par traduction
Ne soumettez pas de travaux où vous auriez juste copié- collé des phrases des articles (ou des blogs scientifiques). Il s'agirait de plagiat, ce qui est répréhensible

Pour bien comprendre le degré de reformulation autorisé qui distingue le plagiat d'un véritable travail de rédaction scientifique, étudiez les exemples proposés sur les sites des universités britanniques et américaines.
Voici par exemple le site d'Oxford.
https://www.ox.ac.uk/students/academic/guidance/skills/plagiarism

pour la traduction (corpus alignés)
https://www.linguee.fr/anglais-francais/traduction/go+online.html
(C'est la qualité initiale de ces ressources qui ont assuré le succès foudroyant de DeepL)

La détection automatique d'erreurs grammaticales et orthographiques (songez à votre traitement de texte)
Grammarly : https://www.grammarly.com/

La suggestion collocations plus fréquentes (COLLOCAID)
Pour les collocations : http://www.collocaid.uk/