Approximation de la complexité perçue – méthode d'analyse

Adrien Barbaresi
ICAR / ENS Lyon

Enjeux

- **Assistance à la compréhension** : déterminer le niveau de complexité d'un texte pour un lecteur potentiel ou dans la perspective d'un traitement automatique, le calibrer en fonction.
- **Thème très riche** : linguistique, études sur la lisibilité, sciences cognitives et théorie de l'information.
- **Outillage de la langue** : réfléchir à l'intégration de techniques en sciences humaines.

Méthode

« **Croiser les approches** »
→ La complexité regroupe différents phénomènes dont il peut s'agir de modéliser le rapport.
→ Articulation entre appréhension théorique du phénomène et approximation acceptable.

Évaluation prévue
→ Banc d'essai constitué de textes que l'on sait simplifiés (pour enfants ou apprenants).
→ Étalonnage avec un panel de locuteurs.

⇒ L'étude porte originairement sur l'**allemand**, un **élargissement** à l'anglais, au français et éventuellement à une langue « exotique » est envisagé.

Corpus

⇒ Opposer différents types de registres sur lesquels on a formulé des hypothèses de simplicité.

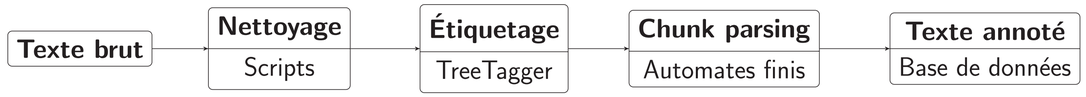
Registre	Source	Volume	Remarques
Journal	Die Zeit	50.000+ articles	étude comparative de la complexité respective des différentes thématiques
Journal	Bild	100.000+ articles	
Magazine	Geo	2.000 articles	
Journal simplifié	News4Kids	289 articles	langue simplifiée
Journal simplifié	Deutsch Perfekt	87 articles	pour apprenants, niveaux de difficulté
Magazine simplifié	Geolino	740 articles	langue simplifiée
Discours politiques	Chancellerie	∼ 1700 discours	langue élaborée (argumentation)
Discours politiques	Présidence	∼ 1300 discours	
Romans	Projet Gutenberg	?	niveau de langue, textes pour enfants
Textes philosophiques	Projet Gutenberg	?	étude du registre
Articles scientifiques	internet	?	langue spécialisée
Romans simplifiés / de gare	Format papier	peu	scan + OCR ?

⇒ Créer un corpus distribuable sur ce phénomène ?

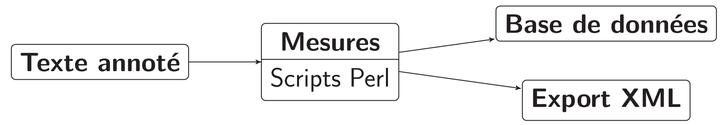
Traits déjà observés

1. Mots
   - **longueur** seuil à 17 caractères (souvent moins de 5 % des mots)
   - **fréquence relative** mots absents des X % les plus fréquents du vocabulaire total du corpus
   - **lemme** reconnu ou non par le TreeTagger, et par un dictionnaire de mots communs
2. Groupes
   - **composition** groupes nominaux et prépositionnels atypiques (voire erreurs de l'étiqueteur)
   - **taille et nombre** difficultés de rattachement des composants et au sein de la phrase
   - **groupe verbal** rection et complémentation
3. Phrases
   - **longueur** critère souvent retenu dans les études de lisibilité (à l'usage, seuils vers 130 et 190 caractères)
   - **subordonnées** type et nombre par phrase (repérées par les virgules et les subordonnants)
   - **attaque d'énoncé** ou avant-première position : phénomènes de linéarisation propres à l'allemand

Architecture de traitement du texte brut au texte enrichi



Mesure et export des résultats



Problèmes à résoudre

- **Repérage et apport éventuel** des **traits sémantiques** de la complexité (densité conceptuelle) et des phénomènes situés au **niveau du discours** (segmentation thématique, cohésion et cohérence, organisation et style).
- **Sélection** des observables : interdépendances, **liens de corrélation** à établir
  ⇒Au final : **critères isolés ou reliés** ?
- **Ajustement** : Quelle complexité pour quel type de lecteur / programme ?

École Normale Supérieure de Lyon – 15, parvis René Descartes F-69342 Lyon BP 7000 Cedex 07 http://perso.ens-lyon.fr/adrien.barbaresi/ adrien.barbaresi@ens-lyon.fr