Linguistique et traitement automatique de corpus
Ambre Dupont
M1 LFC

Université de Lille

Titre de l'article *:*

Simplification syntaxique de phrases pour le français

Auteurs *: Laetitia Brouwers, Delphine Bernhard, Anne-Laure Ligozat, Thomas François*

Année : 2012

Objectif de l'article : présenter une méthode de simplification syntaxique de textes français dans le but de les rendre plus abordables.

Intérêt ?
→ Des textes trop complexes (lexique, syntaxe) peuvent entraîner une mauvaise compréhension. Or savoir lire rapidement et efficacement est un atout majeur...

Exemple :

- **Remplir correctement des demandes d'allocation de chômage** : Richard *et al.* ont montré que sur 92 demandes, la moitié des informations requises étaient absentes à cause d'un manque de compréhension.
- **Prendre correctement un traitement médica**l : Patel *et al.* affirment que la plupart des sujets n'ont pas bien compris les différentes étapes pour l'administration d'un médicament.
- Etc.

SOLUTION : Simplifier ces textes (domaine du TAL) tout en **gardant leur intégrité et leur structure.**

Quelques travaux antérieurs :

- Carroll *et al.* et Inui *et al.* → outils pour simplifier des textes pour les aphasiques ou les sourds

- Belder et Moens → simplification pour des enfants de langue maternelle anglaise

- Siddhartham, Peterson, Ostendorf, Medero → simplification pour apprenants d'une langue seconde

  Remarque : la plupart de ces travaux portent sur l'anglais.

PLAN

1. Présentation du corpus
2. Types et règles de simplification
3. Evaluation du système
4. Conclusion

1. Le corpus

Corpus parallèle construit avec des articles de **Wikipédia** et **Vikidia**.

→ 13 638 fichiers dont 7 460 de Vikidia et 6 178 de Wikipédia.

→ 20 articles sélectionnés soit 72 phrases (Wikipédia) et 80 phrases (Vikidia)

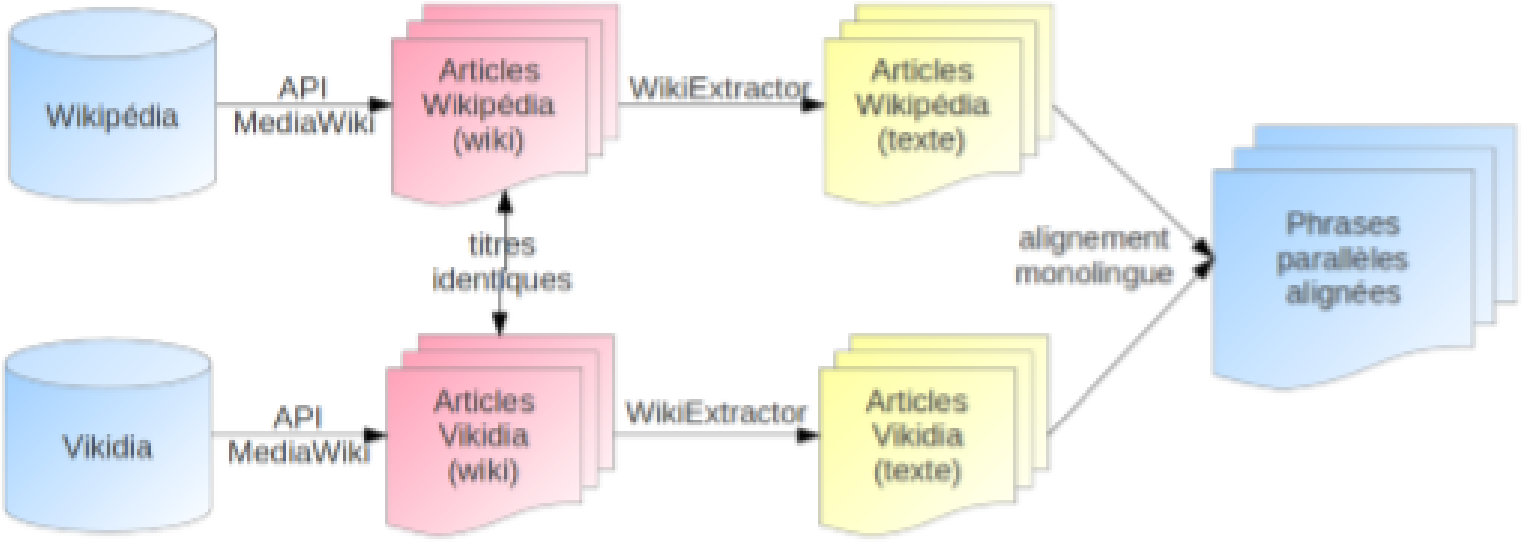

FIGURE 1 – Constitution du corpus de phrases parallèles

Exemple des extraits sélectionnés pour l'entrée « **archipel** » :

Wikipédia : Un archipel est un ensemble d'îles relativement proches les unes des autres. Le terme «archipel» vient du grec ancien "Archipelagos", littéralement «mer principale» **(de "archi" : «principal» et "pélagos" : «la haute mer»)**. En effet, ce mot désignait originellement la mer Égée, caractérisée par son grand nombre d'îles **(les Cyclades, les Sporades, Salamine, Eubée, Samothrace, Lemnos, Samos, Lesbos, Chios, Rhodes, etc.).**

Vikidia : Un archipel est un ensemble de plusieurs îles, proches les unes des autres. Le mot «archipel» vient du grec "archipelagos", qui signifie littéralement «mer principale» et désignait à l'origine la mer Égée, caractérisée par son grand nombre d'îles.

2. Types et règles de simplification :

3 niveaux : lexical , sémantique et syntaxique

Lexique	Sémantique	Syntaxe
Synonyme ou hyperonyme Traduction	Réorganisation Suppression Ajout	Temps Suppression Modification Division Regroupement

TABLE 1 – Typologie

Problèmes :

- certaines modifications nécessitent de recourir à la sémantique → difficulté pour le faire de manière automatique.
- Phrase négative > phrase affirmative : cela peut entraîner un changement de verbe.
- Modification du temps d'un verbe → veiller au respect de la concordance des temps.

Système de simplification syntaxique :

2 étapes :

a) Sur-génération de simplifications possibles pour chaque phrase du texte : 19 règles (suppression (12), modification (3), division (4))

Les règles de regroupement et les changemements de temps ont été laissés de côté.

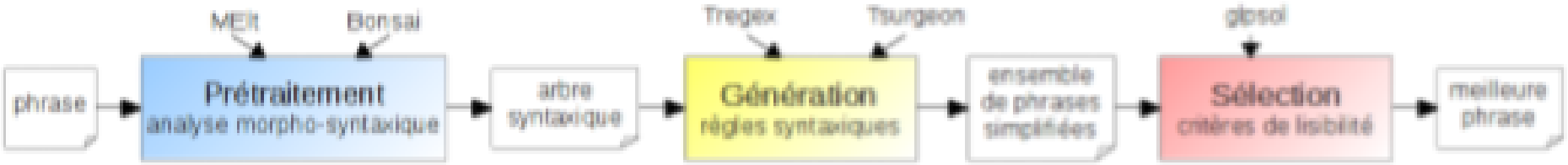

FIGURE 2 – Organisation du système de simplification syntaxique

- Repérer les structures que l'ont veut changer grâce aux expressions régulières et *Tregex*

- Utilisation de *Tsurgeon* pour modifier les arbres syntaxiques (opérations Tsurgeon delete, Tsurgeon move, Tsurgeon insert)
  Ex : *supprimer une coordonnée introduite par « soit »*
  Repérage (Tregex) : COORD=Pcoord < (CC < /soit/)
  Opération (Tsurgeon) : delete Pcoord

Typiquement, on va supprimer :

- Les compléments circonstanciels
- Les phrases entre parenthèses
- Certaines propositions subordonnées
- Les propositions entre virgules, ou introduites par *comme, voire, soit*
- Les adverbes
- Les compléments d'agent

Pour la division :

- La proposition secondaire est supprimée et la principale est enregistrée
- La phrase de base est reprise pour ne garder que la proposition secondaire. Cette dernière est modifiée dans le but de devenir indépendante.

→ **il faut appliquer ces règles de manière récursive pour obtenir toutes les variantes possibles**

b) Sélection du meilleur ensemble de phrases simplifiées grâce à la programmation linéaire de nombres entiers

4 critères :

- La longueur de la phrase (nbre de mots) → 10
- La longueur des mots (nbre de caractères) → 5
- La familiarité du vocabulaire (liste de Catach) →mots absents ≤ 2
- La présence de termes-clés (mots qui apparaissent au moins deux fois ou plus dans un texte)

	Longueur de la phrase	Longueur des mots	Familiarité des mots	Termes clés
Valeurs souhaitées	10 mots	5 caractères	2 mots absents	
Phrase 1	19 mots	6,1 caractères	11 mots absents	5 termes
Phrase 2	11 mots	4,3 caractères	5 mots absents	5 termes
Phrases 3	5 mots	4,6 caractères	2 mots absents	5 termes
Phrase 4	6 mots	4,5 caractères	3 mots absents	3 termes
Phrase 5	4 mots	4,7 caractères	2 mots absents	2 termes
Phrases 6	9 mots	6,3 caractères	5 mots absents	5 termes
Phrase 7	12 mots	7,3 caractères	8 mots absents	2 termes

TABLE 2 – Valeurs des paramètres pour les phrases de l'exemple (4)

3. Evaluation

- Evaluation manuelle

- Nouveau corpus de 202 phrases de Wikipédia

- 113 phrases ont une ou plusieurs simplifications proposées soit 333 variantes → 71 sont problématiques (21,32%).

- 2 types d'erreurs : pré-traitement (89 % des erreurs)
système de simplication (11 % des erreurs)

- Seulement 2,4 % des phrases produites par ce système sont problématiques (erreurs de contenu ou de forme).

Erreurs de pré-traitement :

- Erreurs d'étiquette à cause de cas ambigus
  Ex : *ainsi que* → connecteur logique ? Adverbe + *que* de la clivée ?

  *Les mélodies sont accrocheuses et les arrangements très soignés; c'est **ainsi que** "Mamma Mia" et "Fernando" (malgré quelques erreurs de grammaire anglaise) occupèrent la première place des palmarès mondiaux dans le premier semestre de cette même année.*

  → *C'est.*

  Ainsi que a été considéré comme un connecteur, la règle de suppression de coordonées a donc donné la phrase précédente.

- Les ponctuations peuvent poser problème : l'analyseur ne distingue pas les points dans les citations des points de fin de phrase.

Erreurs de simplification :

- Suppression des infinitives
  Ex :
  C'est aussi depuis le XVIIIe siècle le terme en usage ***pour désigner un clerc séculier ayant au moins reçu la tonsure.***
  → *C'est aussi depuis le XVIIIe siècle le terme en usage*.

- Supression du complément d'agent
  Ex :
  Ils ne sont pas caractérisés ***par leur profession comme dans la Bible*** *: l'un pasteur, l'autre agriculteur.*
  → *Ils ne sont pas caractérisés : l'un pasteur, l'autre agriculteur.*

- Suppression du référent d'un pronom

- Etc.

Conclusion :

- Technique de simplification pour les enfants. Il serait intéressant de l'étendre à **d'autres publics.**

- Utilise des critères de lisibilité, dont certains n'avaient pas été considérés auparavant.

- les résultats sont satisfaisants : **80 % de phrases correctes**.

- Il serait intéressant d'établir une **simplification lexicale** en plus de la simplification syntaxique en s'inspirant d'études déjà effectuées sur l'anglais (Woodsend et Lapata, 2011).

- Nécessité d'ajouter ou de répéter des termes quand on a une division de phrase.