

Corector

Acquisition semi-automatique de cas de corrections de phrases en français

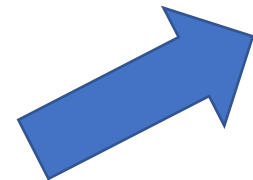
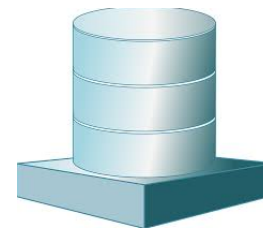
Damien Levy
Tuteur : M. Emmanuel Jeandel
Licence informatique L3

Encadré par :
M. Bruno Guillaume, M. Yves
Lepage, M. Jean Lieber,
M. Emmanuel Nauer

Corrector : logiciel de correction de phrases

Projet divisé en 3 :

- Moteur de raisonnement à partir de cas
- Interface web et base de données
- Acquisition semi-automatique de cas de correction de phrases en français



Base de cas initiale

- Recherche des erreurs courantes
- Création d'un fichier CSV

Exemple

PHRASE AVANT CORRECTION	J'aime pas les pommes.
PHRASE APRÈS CORRECTION	Je n'aime pas les pommes.
STATUT	Correct
GROUPE DE MOTS AVANT MODIFICATION	j'aime
GROUPE DE MOTS APRÈS MODIFICATION	je n'aime
INDICE DE PREMIÈRE DIFFÉRENCE	1
INDICE DE SOURCE	2
LANGUE	fr

WiCoPaCo

- Historique des modifications de pages sous Wikipédia
- WiCoPaCo est réalisé sous licence GNU Free Documentation License (GFDL)

« Aurélien Max and Guillaume Wisniewski, Mining Naturally-occurring Corrections and Paraphrases from Wikipedia's Revision History, LREC 2010. »

Exemple

```
<modif id="27" wp_page_id="9" wp_before_rev_id="33563"  
wp_after_rev_id="110030" wp_user_id="182"  
wp_user_num_modif="255" wp_comment="francisation de la  
phrase">
```

```
<before>Lalgèbre abstraite est une branche des  
mathématiques , qui <m num_words="1">concerne</m>  
principalement des structures et des fonctions entre elles .</before>
```

```
<after>Lalgèbre abstraite est une branche des mathématiques ,  
qui <m num_words="1">porte</m> principalement sur l' étude des  
structures et des fonctions entre elles .</after>
```

```
</modif>
```

Exemple

```
<modif id="27" wp_page_id="9" wp_before_rev_id="33563"
wp_after_rev_id="110030" wp_user_id="182"
wp_user_num_modif="255" wp_comment="francisation de la
phrase">
```

```
<before>Lalgèbre abstraite est une branche des
mathématiques , qui <m num_words="1">concerne</m>
principalement des structures et des fonctions entre elles .</before>
```

```
<after>Lalgèbre abstraite est une branche des mathématiques ,
qui <m num_words="1">porte</m> principalement sur l' étude des
structures et des fonctions entre elles .</after>
```

```
</modif>
```

Exemple

```
<modif id="27" wp_page_id="9" wp_before_rev_id="33563"  
wp_after_rev_id="110030" wp_user_id="182"  
wp_user_num_modif="255" wp_comment="francisation de la  
phrase">
```

```
<before>Lalgèbre abstraite est une branche des  
mathématiques , qui <m num_words="1">concerne</m>  
principalement des structures et des fonctions entre elles .</before>
```

```
<after>Lalgèbre abstraite est une branche des mathématiques ,  
qui <m num_words="1">porte</m> principalement sur l' étude des  
structures et des fonctions entre elles .</after>
```

```
</modif>
```


Extraction de cas

- Étude de la forme du fichier XML
- Extraction des couples phrase avant/après modification
- Mise en place d'un filtre afin d'avoir un maximum de cas pertinents
- Écriture automatique des cas sélectionnés dans un fichier CSV

Exemple de cas à supprimer

- `<modif id="18" wp_page_id="7" wp_before_rev_id="3549368" wp_after_rev_id="3977925" wp_user_id="0" wp_user_num_modif="1096911" wp_comment="">`
 `<before>On nomme <m num_words="2">algèbre linéaire</m> la branche des mathématiques [...] </before>`
 `<after>On nomme <m num_words="1">elbi</m> la branche des mathématiques [...] </after>`
 `</modif>`
 `<modif id="19" wp_page_id="7" wp_before_rev_id="3977925" wp_after_rev_id="3977943" wp_user_id="0" wp_user_num_modif="1096911" wp_comment="">`
 `<before>On nomme <m num_words="1">elbi</m> la branche des mathématiques [...] </before>`
 `<after>On nomme <m num_words="2">algèbre linéaire</m> la branche des mathématiques[...] </after>`
 `</modif>`

Statistiques

Nombre d'éléments	Nombre d'éléments supprimés	Éléments restants	Taux d'éléments supprimés
408 816	108 354	300 462	~26,5 %

Choix du langage : python 2

- Version identique à celui utilisé pour le moteur
- Bibliothèque : `xml.etree.ElementTree`

Conclusion

Problèmes rencontrés

- Extraction des cas complets :
*<before>On nomme <m
num_words="1">elbi</m> la branche des
mathématiques [...] </before>*
- La comparaison des couples est très longue
- Filtrer de manière pertinente afin d'avoir un maximum de cas utiles dans la base

Améliorations possibles

- Amélioration du filtre mis en place
- Ajout de filtres :
 - Détection de modification de chiffres / dates
 - Analyse de phrases