# Corpus Saint-Jean

Juillet 2017

Dominique Labbé
Laboratoire Pacte (CNRS – Université de Grenoble)
Dominique.Labbe@umrpacte.fr

Jacques Savoy
Université de Neuchâtel
Jacques.Savoy@unine.ch

Le corpus Saint-Jean est composé de 200 extraits d'œuvres littéraires françaises dont la création s'étend sur une centaine d'années dans un même genre (le roman).

Chaque texte est sous deux formats : le texte original et une version lemmatisée (même nom de fichier précédé du radical CN). Pour chaque fichier, deux codages sont proposés : windows-1252 et UTF-8. Donc nous avons quatre répertoires distincts.

Ce corpus a été constitué dans plusieurs buts :

## 1. Tester les différentes méthodes d'attribution d'auteur.

Dans ce but, les textes sont anonymés. La table des correspondances (numéro, auteur, titre de l'œuvre) sera fournie à l'issue des expériences. Cependant, les indications suivantes sont disponibles.

Tous les extraits ont la même longueur calculée en nombre de "formes graphiques standardisées" (voir ci-dessous).

Ce corpus comprend trente auteurs. Tout auteur a au moins deux extraits (mais par forcément de la même œuvre). Certains n'ont qu'une œuvre, la plupart en ont plusieurs. Aucun ne pèse d'un poids prépondérant dans le corpus. Une grande clarté dans le choix des œuvres a été retenue. Par exemple, pas de réédition avec complément ou modification de l'œuvre originale (comme une édition posthume remanié par une autre personne), ou des textes contenant des passages importants en langue étrangère ou en jargon, ou encore des œuvres dont l'attribution serait douteuse.

Ce corpus devrait donc permettre d'évaluer l'efficacité des méthodes d'attribution d'auteur par ordinateur. On notera à ce propos que, dans ces expériences, le but principal est de ne pas commettre d'erreur dans les attributions. En effet, une réponse erronée proposée par la machine sera toujours vue comme « stupide » par l'usager final. Il en résulte un manque de confiance envers la solution informatique. Dès lors, un effort doit être entrepris afin d'éviter des erreurs d'attribution et de proposer de répondre « je ne sais pas » plutôt que d'exiger une attribution intégrale.

## 2. Tester les méthodes de classification

Comment classer tous ces textes de manière efficace ? Selon quels facteurs ces classifications sont opérées ? Comment les représenter et en mesurer la qualité ? Comment justifer une attribution ? Ainsi, une classification devrait regrouper les œuvres écrites par le même auteur ou selon le même genre littéraire. Cependant, là encore le principal critère sera de ne pas commettre d'erreur dans les attributions.

## 3. Tester les normes de saisie et de dépouillement des textes

- fichiers .txt : l'orthographe est corrigée, la graphie des mots standardisée c'est-à-dire que le même vocable aura toujours une seule orthographe (par exemple, yaourt, yoghurt, yoghourt, …).

- fichiers CN : il comprend la version lemmatisée. Dans ce fichier, on retrouve une ligne pour chaque mot du fichier txt.

En première position, la forme standard correspondant à chaque mot du texte (dans la graphie figurant dans les dictionnaires de langue).

En seconde position, l'entrée de dictionnaire (le lemme) de ce mot. Par exemple, les deux formes « puis/peux » se retrouvent sous le lemme du verbe « pouvoir ».

En troisième position, sa catégorie grammaticale (voir tableau ci-dessous). Ces trois informations sont séparées par une virgule.

Par exemple, une ligne peut comprendre l'entrée suivante :

```
parents,parent,21
```

Avec la forme standard (parents), suivi de son lemme (parent) et de l'étiquette morpho-syntaxique (21 = substantif masculin).

La toute première ligne indique le nombre de lignes du fichier CN et la seconde comprend une balise encadrée des signes < et >. Toutes les autres balises seront précédées par le signe < et se terminent par le signe >.

Dans ce format, les nombres apparaissant dans le texte sont convertis en toutes lettres. Ainsi, si le nombre « 38 » apparaît dans le texte, le fichier CN indiquera :

```
<Nombre 38>,<>,<>
trente,trente,72
huit,huit,72
<Fin nombre>,<>,<>
```

La virgule et le point requièrent un traitement spécial. En effet, la virgule indique la séparation des champs et le point possède parfois une fonction spéciale (abréviations, suspension, etc.). Dès lors, la virgule et le point seront présentés comme suit :

```
mais,mais,82
",",",",p
une,un,71
fin,fin,22
/.,/.,P
```

Le codage de ces fichiers CN correspond à la norme Windows-1252. Une version dont le codage est UTF-8 est disponible dans le répertoire CN-UTF8.

Pour les outils d'exploitation de ces fichiers lemmatisés, s'adresser à Dominique Labbé.

Cette lemmatisation a été effectuée par des automates. Pour certains textes, il peut rester quelques erreurs. N'hésitez pas à nous les signaler.

Pour le calcul du nombre de mots par texte, on compte tous les lemmes sans les signes de ponctuation. Pour les nombres, on compte le nombre de lemmes (par exemple, « trente deux » qui contient deux mots et non « 32 » qui compterait pour une seule unité).

## 4. Les auteurs du corpus

Sur ces 200 extraits, on retrouve 31 auteurs différents, avec la répartition suivante :

| Nombre extraits | Noms |
|---|---|
| 13 | Balzac |
| 11 | Flaubert |
| 10 | Dumas (père), Maupassant, Sand, Sue, Zola |
| 8 | Barbey d'Aurevilly, France, Hugo, Régnier |
| 7 | Lamartine, Loti |
| 6 | Bourget, Daudet, Sainte-Beuve, Stael, Stendhal |
| 5 | Fromentin, Gautier, Goncourt, Nerval, Vigny |
| 4 | Erkmann-Chatrian, Huysmans, Vallès, Verne |
| 3 | Châteaubriand, Musset, Proust |

La nomenclature des catégories grammaticales dans les fichiers CN

| |
|---|
| **1. Verbe :**<br>12, participe passé<br>13, participe présent<br>14, infinitif<br>15, futur<br>16, conditionnel<br>17, présent (indicatif, subjonctif, impératif)<br>18, imparfait (indicatif et subjonctif)<br>19, passé simple |
| **2. Substantif :**<br>20, « nom propre » (mot à majuscule initiale)<br>21, substantif masculin<br>22, substantif féminin |
| **3. Adjectif :**<br>30, adjectif « pur »<br>31, participe passé dans un emploi « adjectivé » |
| **5. Pronom :**<br>51, personnel (1, 2, 3)<br>52, relatifs, réfléchis, interrogatifs, possessifs, etc. |
| **60. Adverbes :** |
| **7. Déterminant :**<br>71, articles (défini et indéfini)<br>72, numéraux et cardinaux<br>73, possessifs<br>74, démonstratifs<br>75, adjectifs indéfinis |
| **81. Préposition** |
| **82, 83 Conjonction**<br>82, Conjonction de coordination<br>83, Conjonction de subordination |
| **91** Locution<br>**92** Expression & mot étranger<br>**93.** Interjection |
| **Ponctuation**<br>- p, ponctuation mineure (interne à la phrase)<br>- P, ponctuation majeure (délimitant la phrase) |
| **Commentaires et paratexte : <>** |