Document complémentaire - les indices au niveau des articles Vikidia

M2 LITL, 2 avril 2019

Ce document a pour but de préciser les résultats que nous avons obtenu lors de l'étude de Vikidia et de ses deux catégories : les "super articles" et les articles "à simplifier".

1 Liste des indices

1.1 Liste détaillée des indices étudiés

CODE INDICES	VALEUR
GEN_TITLE	titre de l'article
TYPE	type de l'article (super , à simplifier, ...)
GEN_SENTENCE_LEN	longueur en phrases
SYN_NB_PASSIVE	nb de voix passives par nb de phrases
GEN_WORD_LENGTH	nb de mots
PRONOM_AVG_PAR_PHRASE	moyenne de pronoms par phrase
DIS_PRO	nb de pronoms sur nb de noms propres et communs
VERBECONJ_PROREL_VERBE_CONJ_COUNT	nb de verbes conjugués
VERBECONJ_PROREL_SUB_COUNT	nb de propositions subordonnées
VERBECONJ_PROREL_AVG_V1	position moyenne du premier verbe conjugué
SYN_NB_MODIFIERS_PER_NOUN	moyenne du nb de modifieurs par nom
AVG_SENTENCE_LEN	moyenne de la longueur des phrases
TENSE_COUNT_VINF	nb de verbes à l'infinitif
TENSE_COUNT_VPP	nb de participes passés
NB_CONNECTEUR_COUNT	nb de connecteurs de discours
NB_PATRONS_COUNT	nb de reformulations
META_NB_LINKS_COUNT	nb de liens
META_NB_PICTURES_COUNT	nb d'images
STRUCTURE	présence ou absence d'éléments de structuration
META_HTML_score_COUNT	score du calcul de la ponctuation
TENSE_COUNT_VS	nb de verbes au subjonctif
FLESCH	score du test de Flesch-Kincaid
POS_CONT_X	nombre de X

Dans les indices POS_CONT_X, X peut renvoyer aux classes grammaticales suivantes : déterminants interrogatifs (DETWH), pronoms interrogatifs (PROWH), préposition+pronom (P+PRO), d'interjections (I), d'adverbes interrogatifs (ADVWH), préposition+déterminant (P+D), clitiques réflexifs (CLR), clitiques objets (CLO), conjonctions de coordination (CC), de clitiques sujets (CLS), participes passés (VPP), de déterminants (DET), de noms communs (NC), d'adverbes (ADV), d'adjectifs (ADJ), ponctuation (PONCT), préposition (P), noms propres (NPP), pronoms, pronoms relatifs (PROREL) et conjonctions de subordination (CS).
Tous les indices contenant la chaîne de caractères "COUNT" ont été normalisés au niveau des analyses par le nombre de mots de l'article (_GEN_WORD_LENGTH).

1.2 Indices significatifs

Pour obtenir les indices qui discriminent les trois catégories d'articles étudiées (Vikidia, les super articles et ceux à simplifier) nous avons utilisés plusieurs méthodes statistiques :

- Un test statistique de comparaison entre les moyennes des indices sur les "super" articles et les articles "à simplifier"
- L'entraînement d'une classifieur automatique prédisant le type de l'article (arbre de décision)

En ce qui concerne la première méthode, nous avons effectué un test t de Welch. Ce test pose l'hypothèse nulle d'égalité entre les moyennes des deux échantillons (ici, les échantillons correspondent aux

"super" articles et à ceux "à simplifier"). Le tableau (REF) présente les résultats de ces tests pour les indices pour lesquels ont peut rejeter cette hypothèse, autrement dit, les indices pour lesquels il y a une différence significative entre les moyennes de ces deux types d'article de Vikidia [1].

Indice	Groupes	Moyenne	Test t	P-value	Significativité
GEN_SENTENCE_LEN	VikiSimply Vikibest	62 123.13	-1.9924	0.04941	*
GEN_WORD_LENGTH	VikiSimply Vikibest	942.88 2460.29	-3.0646	0.002885	**
PRONOM_AVG_PAR_PHRASE	VikiSimply Vikibest	0.92 1.51	-4.4996	0.0001248	***
DIS_PRO	VikiSimply Vikibest	0.18 0.25	-3.4211	0.001752	**
AVG_SENTENCE_LEN	VikiSimply Vikibest	20.05 25.33	-2.9426	0.008845	***
NB_CONNECTEUR_COUNT	VikiSimply Vikibest	0.07 0.08	-2.7286	0.01116	*
META_NB_PICTURES_COUNT	VikiSimply Vikibest	0.0012 0.016	-2.1227	0.0371	*
POS_COUNT_NC	VikiSimply Vikibest	0.25 0.22	3.1769	0.005324	**
POS_COUNT_P	VikiSimply Vikibest	0.13 0.14	-2.738	0.01204	*
POS_COUNT_CLS	VikiSimply Vikibest	0.022 0.028	-2.419	0.02349	*
POS_COUNT_CLO	VikiSimply Vikibest	0.0051 0.0085	-3.2522	0.0033	**

L'arbre de décision a permis de dégager deux indices comme les meilleurs prédicteurs du type (super ou à simplifier) des articles : META_NB_PICTURES_COUNT et _GEN_WORD_LENGTH (nombre d'images normalisé et taille de l'article en mots).
De plus, une observation des données textuelles nous a permis de tirer les conclusions présentées par la suite. Nous avons ainsi écarté les indices qui considérés individuellement n'ont pas d'interprétation pertinente en termes de complexité (cf. résultats de l'ACP).

2 Interprétation des résultats

A l'issue de ces analyses statistiques, les seuls indices considérés comme significatifs ne laissent en rien présager de la complexité des articles. En effet, les paramètres retenus sont la longueur des articles (en fonction du nombre de mots ou du nombre de phrases) et le nombre d'images.
Les articles de Vikidia "à simplifier" sont en moyenne plus courts que les "super" articles. Pourtant, ce critère est difficilement transposable à la totalité des articles de Vikidia pour rendre compte de leur complexité. En effet, comme on peut le constater sur la figure 1, au sein des articles Vikidia, nous retrouvons une grande quantité d'articles très courts (10% des articles font moins de 2 phrases de longueur). Cette distribution de la taille des articles ne permet pas de discriminer, sur une grande partition d'articles de Vikidia la complexité de lecture. Ces articles courts concernent surtout des articles à compléter ou des ébauches (comme par exemple Cornelius Fudge).

En ce qui concerne les images, la grande majorité des articles à simplifier ne contiennent que peu d'images. Cette observation est en accord avec les principes d'écriture d'un bon article Vikidia, et est déjà préconisée aux rédacteurs. Pourtant, l'absence d'images n'est pas un gage de complexité linguistique, étant donné qu'on n'a en aucun cas évalué le contenu illustratif ou explicatif de ces images, schémas ou cartes, mais un indice meta-textuel de la rédaction de l'article.

D'autre part, les articles les plus longs du corpus Vikidia observés sont globalement très bien structurés et très pédagogiques, notamment des articles du portail des mathématiques (par exemple Fraction ou encore Intégration (mathématique)) qui portent sur des sujets qui ne sont pas simples. La longueur des articles sur des valeurs très positives semble être un symptôme d'un article très travaillé,

1. Les seuils de significativité utilisés sont : $\alpha < 0.05 \rightarrow *$, $\alpha < 0.01 \rightarrow **$, $\alpha < 0.001 \rightarrow ***$

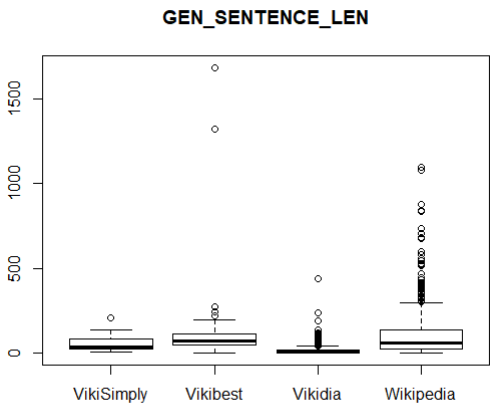

FIGURE 1 – Distribution de la taille des articles sur les 4 corpus étudiés

tout comme un grand nombre d'images, ce qui paraît un bon critère de sa qualité.

Aucun indice de complexité syntaxique parmi ceux que nous avons calculés n'a été retenu comme significatif par la première méthode à part la longueur des phrases. Pourtant, cet indice semble aller à l'encontre des hypothèses de l'état de l'art : des phrases plus longues auraient dû être un indice de complexité, alors qu'elles sont significativement plus longues pour les super articles que pour les articles à simplifier. En comparant les valeurs de cet indice avec les autres corpus, nous avons constaté que les phrases sont en moyenne plus longues pour le corpus de littérature que pour les articles de Vikidia. Ceci nous conforte dans l'hypothèse que ce n'est pas la complexité syntaxique qui détermine un bon ou mauvais article, ou qui en empêche la compréhension, mais plutôt l'organisation et le contenu, des aspects pédagogiques liés à la cohérence et la structure des informations. Cet aspect relève d'un niveau d'analyse pragmatique qui est beaucoup plus difficile à analyser automatiquement.

Pour appréhender l'organisation logique de l'article, nous avons essayé d'approcher le côté pragmatique grâce à l'identification des conncteurs (et de patrons de reformulation). De l'observation du boxplot et des résultats des tests pour les connecteurs, nous avons vu qu'il y a une différence significative entre le VikiBest et VikiSimply. D'un point de vue qualitatif, nous avons pu remarquer que la différence concerne plutôt la variété des connecteurs utilisés et les liens logiques entre propositions. En effet, nous avons remarqué une utilisation plus variée des connecteurs dans VikiBest, avec des connecteurs tels que "car , et, parce que, ainsi" mais aussi "en fait, même si, par exemple". Pour ce qui est des autres articles, on retrouve des connecteurs comme "mais, en fait, donc", sans trop de variation. Cependant, une analyse plus systématique des types de connecteurs employés et de leur contextes serait nécessaire pour en tirer des conclusions plus précises. Toutefois, il pourrait être intéressant de conseiller aux rédacteurs de servir de la page dédiée aux connecteurs logiques pour enrichir leurs articles.

Nous nous sommes aussi intéressés au comptage des différentes parties du discours. Le calcul du ratio du nombre de noms dans les titres par rapport au nombre de total de mots apporte un autre éclairage. Si on regarde les articles dans Vikidia pour lesquels on a un ratio beaucoup plus élevé, on constate que Talismane reconnaît les années comme des noms communs ainsi que '%', ce qu'il faut corriger. Certains articles sont riches en années, pour des résultats sportifs (Juventus Football Club). Les listes d'ingrédients utilisent surtout des noms comme dans des articles sur des produits alimentaires (Fanta). Il est nécessaire pour évaluer la lisibilité de reconnaître certaines structures particulières. De plus, on remarque une prévalence des noms dans les articles courts que l'on pourrait expliquer par l'adoption, à la manière des titres ou de l'oral, de phrases nominales au détriment de phrases ayant un noyau verbal. Les articles dans Vikidia avec un ratio beaucoup plus faible ont une forte présence de noms propres, comme pour le synopsis d'un film (Iron Man 3) ou la description d'une autoroute (Autoroute A86). Les noms propres remplacent les noms communs et ne sont pas comptés pour notre ratio.