Document des séances de regroupement

# I. Problématique de recherche

Un titre de document scientifique est un emplacement singulier d’une importance cruciale. D’une part, il s’agit d’un texte très court d’une dizaine de mots. D’autre part, il est le premier contact entre le document et ses lecteurs et bien souvent le dernier car seulement 8 % des articles sont lus après la lecture de leur titre (Mabe et Amin, 2002). Le titre est donc lu seul dans un premier temps, indépendamment du document titré, de son résumé et de son introduction, et c’est cette lecture qui est discriminante.

Pour ce travail, notre problématique est de

# II. Résumé de l’état de l’art

# III. Données et méthode employées

## III.1 Origine des données

L’accès aux titres a été grandement facilité par la création de bases de données bibliographiques, dont celles des archives ouvertes. Chaque chercheur, quelle que soit sa discipline, ou documentaliste d’un centre de recherche, est libre de déposer un document sur HAL, avec l’accord des auteurs. Une archive ouverte présente l’avantage de centraliser l’accès aux travaux scientifiques, d’aider à leur diffusion et de les conserver manière pérenne, par rapport au site d’une institution particulière ou le site web personnel d’un chercheur, et de façon gratuite et accessible à tous, au contraire du catalogue des grands éditeurs scientifiques.

Nous avons utilisé un corpus constitué par Tanguy et Rebeyrolle (à paraître). Pour obtenir de grandes quantités de titres français, ils se sont tournés vers l’archive ouverte **Hyper Article en Ligne (HAL)** (Nivard, 2010). Cette archive fonctionne depuis 2001 et est gérée par le Centre pour la Communication Scientifique directe du Centre National pour la Recherche Scientifique (CNRS). Plusieurs institutions, dont le CNRS, encourage le dépôt sur HAL des travaux produits par leurs chercheurs, garantissant un nombre important de titres issus de plusieurs disciplines scientifiques.

## III.2 Description des données

Le corpus de près de 340 000 titres fournit pour chacun des informations supplémentaires relatives au document titré :

* Le nombre d’auteurs
* Le type de document
* L’année de publication
* Le domaine scientifique

HAL permet d’attribuer plusieurs domaines à un document. Les domaines sont organisés en arbre, néanmoins la granularité des branches est très variable : « Sciences de l'Homme et Société » est une des racines de l’arbre, regroupant de nombreuses disciplines scientifiques, tout comme « Science non linéaire » et « Économie et finance quantitative ». Tanguy et Rebeyrolle (à paraître) propose une méthode de recodage des domaines pour n’en garder qu’un seul, le plus important et discriminant, que nous avons également utilisée.

Les titres ont été analysé à l’aide du logiciel Talismane (Urieli et Tanguy, 2013) qui fournit un découpage en différents éléments, nous regroupons par ce terme les mots et les signes de ponctuation, un étiquetage morphosyntaxique des mots et une analyse syntaxique en dépendances des éléments. Pour chaque élément du titre nous avons :

* Sa forme dans le titre
* Son lemme (égale à sa forme pour les signes de ponctuation)
* Sa catégorie morphosyntaxique/classe grammaticale (ou ‘PONCT’ pour les signes)
* Des informations complémentaires dépendantes de la catégorie, comme le genre pour les noms, le mode et le temps pour les verbes
* Son élément régisseur
* Son type de dépendance

## III.3 Méthode employée

Dans un premier temps, nous avons analysé les titres selon trois axes.

Talismane considère certains titres comme ayant plusieurs paragraphes et son analyse de dépendance est effectuée paragraphe par paragraphe. Nous avons écarté ces titres pluri-paragraphes, considérant que la détection de plusieurs paragraphes s’apparentent à une erreur. Nous avons ensuite découpé les titres en segments en reprenant d’Anthony (2001) comme délimiteurs de segments le double-point, le point-virgule et les différents points mais écartant en le tiret car il sert en français dans les mots composés.

pour les titres restant

les titres possédant plusieurs parties, plus complexes et ne comptant que pour X % des titres.

# IV. Premiers résultats obtenus

# V. Échéancier sur la période finale

# Bibliographie

Anthony, L. (2001). Characteristic features of research article titles in computer science. *IEEE Transactions on Professional Communication, 44(3)*, 187-194.

Mabe, M. A. et Amin, M. (2002). Dr. Jekyll and Dr. Hyde: Author-reader asymmetries in scholarly publishing. *Aslib Proceedings, 54(3)*, 149-157.

Nivard, J. (2010). *Les Archives ouvertes de l’EHESS*. Récupéré sur La Lettre de l'École des hautes études en sciences sociales n°34: <http://lettre.ehess.fr/index.php?5883>

Urieli, A. et Tanguy, L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane. *Actes de TALN*, Sables D'Olonne.