Proposition de méthodologie de classification des textes spécialisés.

Tantely H. Ravelonjatovo[1]

[1]CIRAM et DIFP – Université d'Antananarivo (Antananarivo) – Madagascar

Abstract

This work is a proposal of a methodology of classification for the specialized texts. The methodology, which is inductive, was elaborated for linguistic analysis of the terms identified by a concordancer. It is a consensus between the terminological point of view and the linguistic point of view. The studied case was a corpus written in Malagasy on the environment and allowed us to have three levels of typology : typology of languages, typology of speech, typology of documents.

Résumé

Ce travail représente une proposition de méthodologie de classification des textes spécialisés. La méthodologie, qui est inductive, a été élaborée pour l'identification des termes à l'aide d'un concordancier, et pour la collecte d'informations linguistiques y afférentes. Elle constitue un consensus entre le point de vue terminologique et le point de vue linguistique. La classification a été conçue pour la structuration d'un corpus écrit en langue malgache sur l'environnement mais dispose de caractères adaptables. Elle s'opère sur trois niveaux à savoir la typologie de langues, la typologie de discours et la typologie de documents

Mots-clés : typologie des textes, terminologie textuelle, classification linguistique, classification terminologique, niveaux de typologies.

1. Introduction

L'essor de la technologie informatique a amplement retentit sur plusieurs disciplines scientifiques. Entre autres, cela a occasionné la prolifération de la linguistique de corpus. Son influence sur la terminologie est également considérable. La terminologie, qui a toujours fait bon ménage avec l'informatique, a connu récemment une nouvelle approche, la terminologie textuelle. A la différence de la terminologie viennoise, cette approche requiert l'utilisation d'outils automatique tout au long du processus méthodologique. En effet, le terminologue textuel en utilise dès la collecte des textes pour la constitution de son corpus jusqu'à l'organisation des termes pour leur diffusion. Le recours aux textes spécialisés constitue un apanage de l'approche. Or, les textes, en tant que produits langagiers des sujets parlant sont naturellement hétérogènes. Ce qui risque de limiter la qualité des résultats après le traitement automatique. D'où la nécessité d'un classement préalable des textes qui est l'objet principal de cet article. Ainsi se posent-elles les questions suivantes. Quels sont les critères à prendre en compte pour la classification des textes spécialisés ? Quelle est la méthodologie adaptée à la classification des textes spécialisés pour un traitement automatique donné ? Pour y répondre, nous nous proposons de partir des objectifs de notre classification dans le premier paragraphe avant d'aborder successivement certaines méthodologies de classification existantes et la méthodologie adoptée pour notre besoin.

2. Les objectifs de la classification

Dans ce paragraphe, nous allons expliciter les objectifs théoriques de la classification pour pouvoir argumenter notre besoin en traitement automatique.

La classification s'inscrit dans le cadre de l'étude des termes environnementaux en langue malgache et s'opère pour la catégorisation des textes spécialisés constitutifs de notre corpus. Ce corpus est composé des articles des revues scientifiques, des articles des journaux, des mémoires de maîtrise es-lettres (spécialité : valorisation du patrimoine naturel), des textes de droit sur l'environnement, des contrats de transfert de gestion et des chartes sur l'environnement et sur le développement durable. Le thème central en est l'environnement et il s'agit d'un corpus monolingue écrit en langue malgache. Cette catégorisation préalable est relative à la forme de publication des textes.

La classification doit également tenir compte du fait que le travail se fait dans le cadre de la terminologie textuelle. C'est une approche qui s'inscrit elle-même dans la linguistique de corpus. Le recours à l'utilisation des textes y est primordial parce que c'est une approche descriptiviste qui se base sur des exemples attestés. Ainsi, la terminologie, en tant que science n'appartient pas aux domaines spécialisés mais à la linguistique appliquée. Les preuves en sont données par L'homme (L'Homme : 2004). L'analyse y est foncièrement linguistique comme ce qui se passe dans la lexicologie. Pourtant, l'unité lexicale, objet d'analyse de la lexicologie y sera qualifiée en unité lexicale spécialisée ou unité terminologique. Nous tenons à signaler que ce dernier diffère de l'unité lexicale par le fait qu'elle s'associe à des sens spécifiques au domaine en cause. De plus, sa combinaison avec les autres signes peut être régie par des règles qui y sont particulières. Ces critères, qui sont loin d'être complets, sont à utiliser pour la détection semi-automatique des termes (première étape du projet). Ils ne relèvent pas du domaine de spécialité mais de la linguistique. Notre projet ne s'arrête pas à l'identification des termes mais s'enchaînera à l'étude de leurs formes et de leurs formations pour pouvoir constituer leurs relations sémantique et/ou conceptuelle dans le domaine. Nous avons, donc, intérêt à faire appel à des critères majoritairement linguistiques pour l'étude des formes et des formations terminologiques (les critères morphologiques, les critères fonctionnels, les critères sémantique, etc.)(Cabré : 1998). Les autres critères qui ont rapport aux connaissances du domaine n'interviendront qu'à l'étude du système conceptuel.

En ce qui concerne l'outils pour le traitement automatique à utiliser pour chaque étape respective, nous devrions recourir à un concordancier qui sait faire apparaître la fréquence, la répartition et la variation des termes aussi bien dans un texte qu'au niveau des textes linguistiquement différents.

En résumé, les textes issus de différents types de documents devraient être classés de manière à pouvoir disposer, après la fouille semi-automatique, des résultats fiables aussi bien du point de vue linguistique que du point de vue terminologique mais pour une analyse linguistique.

3. Les méthodologies de classification des textes

Parlant de la linguistique de corpus, Il existe généralement deux démarches de classification différentes mais pouvant être complémentaires à savoir la démarche inductive et la démarche déductive. La première est apriorique et se manifeste par la détection intuitive de certains éléments de classements de textes, tandis que la deuxième est empirique voire automatique et s'opère après que le corpus soit préalablement constitué. Malrieu et Rastier ont essayé de coupler la démarche inductive avec la démarche déductive pour étudier les variations relatives

à chaque groupe de textes (Marlieu et Rastier, 2001). Ils se sont basés sur l'analyse semi-automatique des variables existant dans chaque classe de textes. Donc, notre démarche est inductive au sens qu'elle s'effectue avant la fouille semi-automatique.

L'étude du concept de classement de textes remonte à l'Antiquité grecque. Aristote en était le pionnier qui s'y est initié. Il est suivi par bon nombre de chercheurs occidentaux où la civilisation de l'écriture a beaucoup évolué. Actuellement, aucune norme ni terminologie en matière de classement de textes n'arrive à s'imposer ni pour le domaine général, ni pour le domaine de spécialité. La fluctuation terminologique pour la désignation du concept ne fait que confirmer ce propos (registre, genre, discours…). Tout de même, des travaux ont été faits aussi bien par des terminologues que par des linguistes. Nous allons voir ci-après les méthodologies existantes avancées par les prédécesseurs.

3.1. La classification des textes dans le domaine de spécialité.

Nous allons mettre de côté les classifications terminologiques déductives qui ont généralement pour but de classer, a posteriori, les textes ou les termes par l'utilisation d'outils automatique (construction d'ontologies, construction de ressources terminologiques, acquisition des connaissances, modélisation des connaissances, etc.). Ce qui nous emmène aux travaux de Pearson (Pearson, 1998) avec qui L'Homme (L'Homme, 2004) est du même avis. La classification de Pearson se fait en deux étapes dont la classification pour la sélection des textes et la classification pour la requête d'informations et/ou la génération de petit corpus pour une étude particulière. Pour l'auteur, les textes collectés selon les critères basés sur la fonction communicative (expéditeur – message – récepteur) devrait être reclassés selon d'autres critères externes (genre, mode de production, les sources, objectifs) et internes (sujet ou matière, modèle ou style) pour une étude donnée.

3.2. La classification des textes dans le domaine général.

Comme nous avons mentionné ci-haut, l'idée de classer les textes n'est pas nouvelle mais c'est son application dans un domaine déterminé ou pour un objectif spécifique qui varie. La plupart des recherches entamées concernent le domaine général et se sont basées soit sur la théorie de la communication de Jackobson, soit sur la classification littéraire. Bronckart (1996) voulait partir des genres historiques (source d'inspiration des locuteurs) et les typologies linguistiques pour une proposition de méthodologie de classification de textes. S'inquiétant surtout sur les principes et théories de textes, l'auteur n'est pas arrivé à préconiser une méthodologie concrète. Après Habert (Habert, 2001), Rastier (Marlieu et Rastier, 2001 ; Rastier, 2001) est allé plus loin dans la théorisation des textes. Il est parvenu à trouver une hiérarchie de classification des concepts. En se basant sur le fondement de la linguistique de corpus en tant que science du langage naturel humain, il met la langue tout au dessus de la pyramide. Le discours et ses champs génériques viennent après et s'en suivent les genres, les textes, les sections, les morphologies et les usages génériques. Alors que la typologie de genre est subordonnée à la typologie de discours, Rastier opte pour la primauté du « genre » pour la constitution de corpus et la classification de ses textes.

Quelle approche choisir, alors, pour la classification de nos textes ?

4. La méthodologie de classification adoptée

Le choix méthodologique que nous avons adopté est mixte au sens qu'il reflète à la fois l'approche terminologique de L'Homme et Pearson (L'Homme, 2004 ; Pearson, 1998) et la classification généraliste synthétisée par Rastier (Marlieu et Rastier, 2001 ; Rastier 2001).

Primo, notre corpus a été constitué d'après les propositions de L'Homme (L'Homme, 2004), celles avancées par Pearson (Pearson, 1998) dans la première étape de sa classification. La raison en est que la catégorisation des textes par niveau de spécialisation garantit l'existence de termes et l'adéquation de leurs utilisations. Nous y avons fait une légère adaptation pour les textes de notre corpus. De ce fait, nous les avons catégorisés en trois niveaux de spécialisation selon que les locuteurs soient confirmés, apprentis ou intermédiaires. Les articles scientifiques sont classés dans le groupe des textes produits par les locuteurs confirmés parce que ce sont des textes dans lesquels se trouve une forte concentration de termes couplée à une forte adéquation de leurs utilisations. Les mémoires de maîtrise ont été classés dans le deuxième groupe qui contient les textes produits par les locuteurs apprentis qui sont des spécialistes en devenir. Nous précisons que ces étudiants ont été formés dans une Faculté des Lettres et sciences Humaines mais spécialisés en Valorisation du Patrimoine naturel. Le troisième groupe comprend des textes écrits par des locuteurs issus du domaine général ou de domaines connexes (droit et développement durable) et qui n'ont que partiellement des connaissances en environnement.

Secondo, nous avons essayé de trouver le modèle approprié à notre situation pour la deuxième étape de Pearson (la classification par sujet/matière et modèle/style). Cette fois, nous avons essayé de trouver des places pour nos textes dans la hiérarchisation de Rastier citée ci-haut (sous-paragraphes 3.2). À la différence des textes littéraires, les articles scientifiques et les mémoires scientifiques se trouvent dans le modèle scientifique. Les textes de droit, les chartes appartiennent au modèle juridique. Et enfin, les articles des journaux ainsi que des quelques articles parues dans des revues à caractère journalistique sont regroupés dans le modèle journalistique.

Quant au choix terminologique, nous avons choisi, à partir des classifications de Rastier, trois termes correspondants aux trois niveaux de typologie adapté à notre corpus. La *typologie de documents* (articles, mémoires, charte, textes de droit) désigne la première et se trouve au plus bas de l'échelle. Le niveau supérieur est indiquée par la *typologie des langues.* La *typologie de discours* ou classement en *champs génériques* se trouve au niveau intermédiaire. Ce niveau est susceptible d'avoir un autre sous-niveau selon la diversité des documents.

Tableau1 : Classification des textes

Niveau	Typologies	Types		
A	Typologie de langues	Langue spécialisée vs langue générale		
B	Typologie de discours	Discours scientifiques	Discours juridiques	Discours journalistiques
C	Typologie de documents ou champs génériques	Articles scientifiques et mémoires	Textes de droit, chartes, contrats de transfert de gestion.	Articles des journaux et autres articles

5. Conclusion

Pour l'identification des termes après la fouille dans un corpus électronique spécialisé en vue de l'analyse linguistique, il est essentiel de procéder à la classification des textes selon des critères et terminologiques et linguistiques. La première classification de Jennifer, représentant les terminologues contextuels ou textuels, qui est plutôt relative à la fonction communicative des textes permet de classer les textes par niveaux de spécialisation et pourrait favoriser l'obtention des termes et leurs utilisations dans le domaine. Or un terme, outre son appartenance au domaine possède des caractères linguistiques qui sont essentiels surtout pour l'étude de ses relations sémantiques et conceptuelles avec d'autres termes. Ces caractères favoriseraient également l'étude des variations terminologiques selon les modèles discursifs donnés. Sans donner une piste clarifiée, l'auteur elle-même en est conscient en proposant l'utilité de recourir à des modèles pour une classification supplémentaire. Du coup, nous y apportons les niveaux de typologies, de Marlieu et Rastier, légèrement adaptés à nos textes. Ce qui a donné lieu à trois niveaux de classification à savoir la typologie de langues, la typologie de discours, la typologie de documents ou champs générique. Les chercheurs pourront y introduire d'autres sous-niveaux de classification en fonction des types de documents en sa disposition et par rapport à ses objectifs. Ce qui va risquer de poser un autre problème terminologique. En effet, Rastier utilisait le terme *typolgie de genre* pour désigner le niveau inférieur à celui de la *typolgie de discours*. Est-ce que ce terme pourrait être utilisé pour classifier les textes spécialisés ? Autrement dit, quels sont les rapports qui existent entre *genre* et *discours* dans le domaine général et dans le domaine de spécialité ? Bref, cette méthodologie tend à rapprocher de plus en plus la terminologie à la linguistique

Références

L'Homme M-C. (2004). *La terminologie : principes et techniques*. Les Presses de l'Université de Montréal.

Pearson J. (1998). *Terms in Context.* John Benjamins publishing company.

Cabré M.T. (1998). *La terminologie : Théorie, méthode et applications.* Les Presses Universitaires d'Ottawa et Armand COLIN.

Bronckart J. P. (1996). Genres de textes, types de discours et opérations discursives. *Enjeux*, (37-38), 31-47 [en ligne]. Disponible sur : <http://rtpdoc.enssib.fr/fichiers/definitionDocument/Bibliothèque_document/genre%20de%20texte%type%20discours.pdf>.

Marlieu D. et Rastier F. (2001). Genres et variations morphosyntaxiques. *Traitement automatique des langues*, vol.(42) n°2/2001 : 547- 577.

Rastier F. (2001). Eléments de théorie des genres. *Texto !* juin 2001 [en ligne]. Disponible sur : <http://www.revue-texto.net/Inedits/Rastier/Rastier_Elements.html>.

Habert B. (2001). Des corpus représentatifs : de quoi, pour quoi, comment ? *Linguistique sur corpus. Études et réflexions*. éd. par Bilger M. Presses Universitaires de Perpignan.