
La variation dé nominative dans les textes scientifiques en sciences de la Terre : proposition d'une méthodologie pour l'annotation de la variation

José Cornejo Cárcamo^{*1}

¹Approches Linguistiques Théoriques, Appliquées et Expérimentales : langues et cultures connectées – Université Paris Cité – France

Résumé

Les linguistes s'intéressent à la variation terminologique depuis plus de vingt ans (Cabré, 1999 ; Collet, 2004 ; Desmet, 2006), démontrant la présence et l'usage des variantes dé nominatives dans le discours spécialisé. Cette communication, réalisée dans le cadre du projet MaTOS (Machine Translation for Open Science), examine la variation dé nominative intratextuelle dans les textes scientifiques en anglais en sciences de la Terre. L'objectif est de décrire cette variation, présenter une méthode d'annotation ainsi que les difficultés rencontrées lors de l'annotation des termes pour améliorer la traduction automatique.

Notre méthode s'appuie sur une approche onomasiologique et contextuelle d'identification des variantes à partir de "termes de base" (première occurrence lexicale d'un concept). S'inscrivant dans les travaux sur la variation terminologique (Pecman 2012, 2014 ; Freixa, 2022 ; Fernández-Silva, 2019, 2022), l'étude propose une typologie de la variation adaptée d'après Biel (2023), Candel-Mora (2024) et Carrasco & León-Araúz (2023).

Nous adoptons une approche manuelle et semi-automatique : (1) identification des termes de base ; (2) recherche de marqueurs discursifs introduisant une possible équivalence (parenthèses, also, or) (Suárez de la Torre, 2004) ; (3) recherche systématique par constituants (Hamon & Nazarenko, 2001) ; (4) validation externe à l'aide de Termium Plus, ARTES (Kübler & Pecman, 2012; Pecman, 2021), Google Scholar et un corpus comparable interne à Université Paris Cité.

L'annotation sur Label Studio permet d'établir visuellement les liens entre les termes de base et les variantes à l'aide d'une typologie ayant six catégories de variation : graphique (*acoustic Doppler current profiler / ADCP*), morphosyntaxique (*mantle exhumation / exhumation of the mantle*), par expansion (*barriers / seismic barriers*), par réduction (*forearc basins / basins*), lexicale (*slow slip / aseismic slip*) et changements multiples (*slow slip event / silent earthquake*).

L'identification des termes complexes révèle quelques défis (e.g. Volanschi, 2008 ; Gledhill & Pecman, 2018) puisqu'ils se trouvent imbriqués au sein des syntagmes nominaux complexes (Benetti, 1995 ; Kübler et al., 2018, 2022) et partagent cet environnement avec des collocations spécialisées. Les frontières entre ces phénomènes semblent donc difficiles à délimiter. Des cas tels que *earthquake rupture* au sein du SNC *preventing fast earthquake rupture and*

^{*}Intervenant

promoting fault creep ou slow slip event (SSE) ayant des occurrences comme the episodes of slow slip ou long episodes of aseismic slip illustrent ce problème.

Pour la traduction automatique, ces données annotées permettraient de créer des listes de formes attestées de termes et de variantes pour entraîner et tester des modèles de langues. L'objectif serait d'évaluer leur gestion de la variation terminologique en contexte discursif afin d'améliorer potentiellement la cohérence et la qualité des traductions scientifiques.