Diversité lexicale et longueur du texte en évaluation du langage

Yves Bestgen

Université catholique de Louvain – yves.bestgen@uclouvain.be

Abstract

Estimating the lexical diversity of a text has been one of the main subjects of study in lexicometrics since its beginnings. It is also a question that has seen a large number of applications, particularly in the field of language acquisition and deterioration. In this field, indices based on probability laws, such as Muller's index, are considered too affected by length differences to be recommended, even though these indices are insensitive by construction. The aim of this paper is to understand why this research has led to an erroneous result, and to present a study carried out on 600 texts, written by learners of German, Czech and Italian, in order to rigorously evaluate the lexical diversity indices used in this field.

Keywords: Lexical diversity, random sampling, intra-class correlation coefficient, individual profiles.

Résumé

Estimer la diversité lexicale d'un texte est un des sujets principaux d'étude de la lexicométrie depuis sa naissance. C'est aussi une question qui a connu un grand nombre d'applications, tout particulièrement dans le domaine de l'acquisition et de la détérioration du langage. Dans ce domaine, des indices basés sur des lois de probabilité, comme l'indice de Muller, sont considérés comme trop affectés par les différences de longueur pour être recommandés. Il est pourtant bien établi que ces indices sont insensibles par construction. Cette communication a pour objectif de comprendre pourquoi ces recherches ont abouti à un résultat erroné et de présenter une étude menées sur 600 textes, rédigés par des apprenants de l'allemand, de l'italien et du tchèque, afin d'évaluer d'une manière rigoureuse les indices de diversité lexicale employés dans ce domaine.

Mots clés : Diversité lexicale, échantillonnage aléatoire, coefficient de corrélation intra-classe, profils individuels.