Computational Study of Lexical Productivity in Ancient Greek

Nadège M. Y. Rollet
nrollet@ucm.es

Lunes 30 de noviembre de 2020, 17:00 – CCHS, CSIC

Spanish/ English abstract

Estudio computacional de la productividad léxica en antiguo griego

El calculo de la productividad léxica de una clase de palabras, realizado mediante un modelado computacional, resulta revelador para entender mejor la dínamica de la gramática léxica de los hablantes de lenguas modernas. Ahora bien, cuando uno pretende estudiar dicha productividad en una lengua antigua cuyo corpus es histórico y cerrado, ¿se pueden aplicar metodologías similares a las desarolladas a partir de un corpus sincrónico?

El propio Baayen (2009), en su definición de la productividad realizada de una categoría de formación de palabras a partir de un corpus sincrónico, compara el estudio de aquel corpus con el de un modelo diacrónico, igual que en la vida se acumulan nuevos extractos de texto a la experiencia cumulativa del locutor. Nuestro objetivo es precisamente proponer una adaptación de la medición computacional de los tres tipos de productividad léxica – la productividad realizada, potencial y de expansión – tal como la define Baayen, pero a partir de un corpus realmente histórico como es el del griego antiguo.

Tal metodología no se podía desarollar sin el aporte de los datos y recursos de programación disponibles hoy en día. Consecuentemente, trataré también de ejemplificar técnicas de procesamiento de lenguajes naturales aplicadas a lenguas antiguas compiladas por la biblioteca python CLTK (Johnson et al., 2014-2020) y la optimización de sus lematizadores con el uso de la reciente lista de palabras MAGWL (Riaño Rufilanchas, 2019).

El objetivo de mi presentación es por tanto ofrecer un caso de estudio que combina la teoría lingüística actual y aplicada a las lenguas habladas, con el empleo de nuevas tecnologías de procesamiento de lenguaje que he aplicado al estudio de las lenguas antiguas, con el fin no solo de obtener mejores fundamentos para el estudio de las mismas, sino también de mantener vivo el diálogo entre la lingüística general y la histórica.

REFERENCIAS

Baayen, R. Harald 2009. Corpus linguistics in morphology: morphological productivity. *Corpus linguistics. An international handbook.* 900-919

Johnson, Kyle P. et al.. 2014-2020. CLTK: The Classical Language Toolkit. DOI 10.5281/zenodo.60021 [recurso online]

Riaño Rufilanchas, Daniel 2019. MAGWL : Madrid Ancient Greek Wordlist, Grupo de Lingüística Griega del ILC. https://glg.csic.es/MadridWordList/MadridAncientGreekWord List.html [recurso online]

Computational Study of Lexical Productivity in Ancient Greek

The calculation of the lexical productivity of a given word-formation pattern that we obtain through computational modelling may be revealing in order to achieve a better understanding of the dynamics in play in the lexical grammar of modern languages speakers. The question still remains, however, when one tries to study the same kind of lexical productivity within the closed data corpus of an ancient language: can we apply the same methodology for the study of a synchronic corpus?

Baayen (2009) compares measuring the realized productivity of given class of words in a synchronic data corpus to a model of diachrony: "A corpus can also be viewed as a (simplified) model of diachrony, as through life, new samples of text are continuously added to one's cumulative experience." Our goal is precisely to adapt his methodology to the study of ancient languages in order to measure the three types of productivity he defines for a given pattern –realized, potential and expanding productivity– from the quantitative study of a diachronic and historical data corpus such as the Ancient Greek one.

It would not have been possible to engineer such a methodology without the data and computer resources at hand nowadays. Consequently, we also aim to illustrate the use of Natural Language Processing techniques specialized in the study of ancient languages and compiled by the CLTK library (Johnson et al., 2014-2020) as well as the use of the recent MAGWL (Riaño Rufilanchas, 2019) to obtain a more precise output than was possible using previous lemmatizers.

Thereby, our goal is to present a case study where present linguistic theory usually applied to the study of spoken languages crosses paths with new technologies especially developed to study ancient languages in a way that keeps alive the dialog between General Linguistics on one hand and historical linguistics on the other.

REFERENCES

Baayen, R. Harald 2009. Corpus linguistics in morphology: morphological productivity. *Corpus linguistics. An international handbook.* 900-919

Johnson, Kyle P. et al.. 2014-2020. CLTK: The Classical Language Toolkit. DOI 10.5281/zenodo.60021 [online resource]

Riaño Rufilanchas, Daniel 2019. MAGWL : Madrid Ancient Greek Wordlist, Grupo de Lingüística Griega del ILC. https://glg.csic.es/MadridWordList/MadridAncientGreekWord List.html [online resource]