Synthetism and analytism in the Celtic languages: Applying some newer typological indicators based on rank-frequency statistics

Andrew Wilson
Lancaster University
a.wilson@lancaster.ac.uk

Róisín Knight
Lancaster University
r.knight1@lancaster.ac.uk

1 Introduction

This study applies some newer quantitative typological indicators to elucidate relationships and evolution within the Celtic language family. These indicators are distinctive from earlier typological indicators (such as Greenberg's [1960] synthetism index) in that they require no morphosyntactic analysis but rely purely on rank-frequency or type-token statistics (Popescu & Altmann, 2008a, 2008b; Popescu, Mačutek & Altmann, 2009; Kelih, 2010). An important point about Greenberg's indices is that they require a fairly deep knowledge of the grammar in order to be applied reliably. Even with such knowledge, they involve substantial effort in manual analysis. Simpler indicators that can measure the same constructs are therefore to be welcomed. Descriptively, the work extends the typological analysis of Tristram (2009) on Celtic, which excluded three of the languages (Manx, Cornish, and Scottish Gaelic).

2 Theory

The power-law function for ranked word frequencies typically does not fit exactly and usually crosses the observed frequencies somewhere within the hapax legomena. Popescu and Altmann (2008a) have observed that, if the curve crosses the observed frequencies early, so that most of the hapax legomena lie above it, this indicates a tendency towards synthetism; however, if the curve crosses the observed frequencies late, so that most of the hapax legomena lie below it, then this indicates a tendency towards analytism. This is because analytic languages tend to use the same word-form multiple times whilst synthetic languages use a greater number of unique forms (because the lexeme changes form to signal grammatical information). In the former case, the function underestimates the hapax legomena and, in the latter case, it overestimates them.

Kelih (2010) has also suggested that type-token statistics alone might be an indicator of typology, without needing to fit the power function. This is because an increase in the number of hapax legomena – the main underlying feature of the rank-frequency-based indicators – necessarily leads to a change in the type-token relationship overall.

2 Data

The data for this pilot study is a small translation corpus of ten Psalms per language, giving 70 texts in total. All of the Celtic languages are included: Welsh, Cornish, and Breton (the "P-Celtic" branch); and Manx, Scottish Gaelic, and Irish (the "Q-Celtic" branch). Two periods of Irish are included as separate samples. Each text was processed individually.

3 Results

For the languages where comparative data are available (Welsh, Breton, and Irish), all of the rank-frequency-based indicators are rank-order identical with Greenberg's synthetism index, as computed by Tristram (2009). Such a direct comparison has not previously been made for any language, and this finding bodes well for future applications of these indicators.

More concretely, the indicators demonstrate not only that Irish has evolved from a greater to a lesser degree of synthetism but also that synthetic versus analytic tendencies within Celtic seem not to be linked in any way to the ancestral Q- versus P-Celtic classification. This picture was not entirely clear in Tristram's (2009) study, since she did not compute Greenberg's index for Manx, Cornish, and Scottish Gaelic. In our study, Manx (a Q-Celtic language) is the most analytic of all; in contrast, Cornish (a P-Celtic language) is the second most synthetic language, more so than Modern Irish (Q-Celtic). Since the diachronic tendency in most European languages has been a move away from synthetism, it seems unlikely that disparities in text dates lie behind these results: the Cornish texts are the most recent, whilst the Manx texts only post-date the Early Modern Irish texts by around a century.

The type-token statistics tell a slightly different story, so it has to be assumed that the two approaches are actually not directly comparable. In this case, the pattern is more directly suggestive of Q- versus P-Celtic relations, with the two historical stages of Irish particularly close to one another.

4 Conclusion

This research, despite drawing only on a small pilot sample of Psalm texts, and with limitations on text dates, suggests that the newer typological indicators may be of considerable value in investigating

morphosyntactic typological variation. As far as Celtic is concerned, our continuing work drawing on other discourse types and other dates will surely tell an interesting story.

References

Greenberg, J.H. 1960. "A quantitative approach to the morphological typology of languages". *International Journal of American Linguistics,* 26: 178-194.

Kelih, E. 2010. "The type-token relationship in Slavic parallel texts". *Glottometrics,* 20: 1-11.

Popescu, I.-I. and Altmann, G. 2008a. "Hapax legomena and language typology". *Journal of Quantitative Linguistics,* 15(4): 370-378.

Popescu, I.-I. and Altmann, G. 2008b. "Zipf's mean and language typology". *Glottometrics,* 16: 31-37.

Popescu, I.-I., Mačutek, J. and Altmann, G. 2009. *Aspects of word frequencies.* Lüdenscheid: RAM-Verlag.

Tristram, H.L.C. 2009. "Wie weit sind die inselkeltischen Sprachen (und das Englische) analytisiert?" In U. Hinrichs, N. Reiter and S. Tornow (Eds.), *Eurolinguistik: Entwicklung und Perspektiven* (pp. 255-280). Wiesbaden: Harrassowitz.