Distant Diplomatics: new chances for "distant linguistics" on medieval charters?

Medieval charters have been and are being used as a source for linguistic studies of various descriptions. In my presentation, I shall first sketch an overview of the types of such linguistic research. After that, I shall focus on three types in which I have most experience and which I reckon will most benefit from distant-diplomatic advances. At the same time, I shall also examine the specific requirements for and potential pitfalls of such "distant" approaches. The discussion will be mainly about Latin data.

The types of linguistic research to be discussed are: 1) Late Latin case marking (system of the language in focus), 2) Latin as a second-language acquisition in the Middle Ages (language as an indicator of socio-historical change), 3) variation of documentary formulae (language as an indicator of cultural and personal connections, which, in turn, feed back to linguistic research). The distinction between 1 and 2–3 reflects self-contained and instrumental uses of linguistics, respectively, and is fundamental to the understanding of the radically differing premises, aims, and data requirements of the different paradigms of linguistic research. Namely, if a study accumulates new information on the very system of a given language, it can be termed purely linguistic (Korkiakangas 2016); if it exploits language instrumentally to attain a language-external goal (e.g., mechanisms of second-language acquisition), it can still be called linguistic, but it departs from the core linguistics in that it compares the data to what is aprioristically known of the system of the language (e.g., Korkiakangas [forthcoming]).

The methodologies of distant reading mostly apply to quantitative research settings, typically conceptualized as corpus linguistics. The examples 1–3 are all quantitative. However, even qualitative linguistic approaches can draw on the developments of distant diplomatics. Indeed, perhaps the greatest part of language-related studies on charters are about vocabulary, i.e., historical toponymy and onomastics (e.g., Lamberg 2021), which mainly serve historiography, and, on the other hand, etymology and historical dialectology (e.g., Francovich Onesti 2013), which are purely linguistic fields in their own right. It is probably here that the large-scale digitization of charters and the application of advanced HTR models to enable key word spotting have their lowest-hanging fruit to pick in terms of providing new digital textual data.

To investigate the system of a language using charters obviously requires that their language is produced by native speakers. Yet, with Latin, such charters are only available up to the $10^{th}/11^{th}$ centuries, after which Latin is more clearly a second language even in the Romance world. Another problem is that coherent reliable bodies of charters in electronic text with enough exemplars from a certain document type, type of issuer, period, and place are seldom available to enable contrastive statistical study. Therefore, apart from promoting massive digitization and HTR, distant diplomatics would best foster linguistics by developing automated methods of metadata extraction/creation (e.g., document typologization and dating; cf. Tilahun & al. 2012). Equally important would be the automated detection of diplomatic parts (e.g., Galuščáková & Neužilová 2018), indispensable to all linguistic research, given that the formulaic and non-formulaic diplomatic parts of documents tend to reflect different linguistic realities. For these purposes, computer vision might be used more efficiently along with text-based methods: similar layouts may indicate similar contents (cf. Christlein 2018).

References

Christlein, Vincent. 2018. 'Automatic Detection of Illuminated Charters', *Illuminierte Urkunden: Beiträge aus Diplomatik, Kunstgeschichte und Digital Humanities*. AfD Beiheft 16, 45–52.

Francovich Onesti, Nicoletta. 2013. 'Il nome Lapo e i suoi antefatti nella documentazione altomedievale', *Le regine dei longobardi e altri saggi*, 31–40.

Galuščáková, Petra & Neužilová, Lucie. 2018. 'Low Resource Methods for Medieval Document Sections Analysis', *LREC 2018 Proceedings*, 2344–8.

Korkiakangas, Timo. 2016. *Subject Case in the Latin of Tuscan Charters of the 8th and 9th Centuries*.

Korkiakangas, Timo. [forthcoming] 'Spelling correctness as a witness of changing documentary culture in Tuscia (VIII–IX centuries)', *Early Medieval Europe*.

Lamberg, Marko. 2021. 'Människa och miljö i medeltida Åbo', *Folkmålsstudier* 59, 71–102.

Tilahun, Gelila, Feuerverger, Andrey & Gervers, Michael. 2012. 'Dating medieval English charters', *Annals of Applied Statistics* 6(4), 1615–40.

Verdo, Rémy. 2010. *La reconfiguration de latin mérovingien sous les Carolingiens: étude sociolinguistique des diplômes royaux et des réécritures hagiographiques*. Thèse non publié.

Bio

Timo Korkiakangas (Academy of Finland Research Fellow; Docent, PhD, Uni Helsinki) is a Latin linguist specialized in early medieval Latin charters. Since 2010, he has carried out four individual projects that combine treebank-based corpus-linguistics with other digital methods, such as handwriting quantification and network analysis.