(Mis)Adventures in Corpus Creation: Lemmatization, PoS tagging, and the Creation of a Dictionary of Forms for Old Spanish.

Francisco Gago-Jover
College of the Holy Cross

For the last four years a small team of hispano medievalist have been working on the development of the *Old Spanish Textual Archive*, a morphologically tagged and lemmatized corpus of more than 25 million words, based on the more than 400 paleographic transcriptions of medieval texts written in Castilian, Asturian, Leonese, Navarro-Aragonese and Aragonese prepared by the collaborators of the Hispanic Seminary of Medieval Studies (HSMS).
Once we established the basic architecture of the corpus, we proceeded to evaluate a variety of taggers/lemmatizers—Freeling, LaPOS, Marmot, neoTAG, and TreeTagger—trying to find one that allowed us to preserve the orthographical variation present in medieval texts written in Castilian, Aragonese, Navarrese, and Leonese, without the need to normalize the language prior to processing, as this would have eliminated most variation and with it, the linguistic value of the corpus.
We ended up selecting Freeling as it had already been trained and tested with a small part of the HSMS corpus (Sánchez Marco 2011, 2012), thus avoiding the need to train the other tools. FreeLing's processing pipeline is relatively simple: a set of texts is sent to the analyzer, which processes and enriches the texts with linguistic information using different modules: tokenizer, dictionary, affixer, probabilistic analyzer and unknown-word guesser, and PoS tagger. However, once we processed the whole HSMS corpus with Freeling to evaluate its real-world performance, we detected a fairly large number of problems caused by the dictionary that Freeling employs. Correcting and improving that dictionary became our main task for the last two years. In this presentation I will discuss 1) the origins and structure of the original Freeling dictionary and why unsupervised methods of enrichment can lead to problems, 2) the typology of problems detected and the implications they have on an accurate tagging/lemmatization, 3) the tool developed to edit and improve the dictionary, with its benefits and shortcomings, and 4) other methods used to improve the dictionary, including lemma and PoS extraction from other dictionaries, and the process of manual revision.

FRANCISCO GAGO-JOVER is Professor of Spanish at the College of the Holy Cross, in Worcester, Massachusetts. He received his licenciatura in Geography and History at the Universidad de Valladolid, and his Ph.D. in Hispano Romance Linguistics and Philology at the University of Wisconsin-Madison with a

dissertation on Medieval Spanish military lexicography. He is the author of two dictionaries, an edition of the Spanish version of the Art of Dying Well, numerous articles on lexicography, and several paleographical transcriptions of medieval and early colonial Spanish texts. He has taught doctorate courses in different universities in the United States (University of Massachusetts-Amherst and Boston University) and Spain (Universidad de León, Universidad de Valladolid, and Universitat de les Illes Balears). For the last 10 years he has been involved in several Digital Humanities projects. He is also the Director of Digital Projects at the Hispanic Seminary of Medieval Studies and is in charge of the Digital Library of Old Spanish Texts and the Old Spanish Textual Archive.

Gathering, cultivating and harvesting the data — the evolution of a digital resource for Old Norse prose.

Ellert Jóhannsson
University of Copenhaguen

The main theme of this paper is the data used in historical dictionary work and how technological advances have opened up new avenues for managing, exploring and exploiting lexicographic data.
The focus will be on the development of A Dictionary of Old Norse Prose (ONP), which has gone through several incarnations in its 80 years of existence. ONP was originally conceived of as a supplementary citation collection to its main 19th century predecessor. It then became an independent dictionary project, with its own set of principles and standards, centered on collecting example citations to cover the entirety of the Old Norse prose vocabulary and to register data about the medieval source material. This work then finally bore fruit in the eighties when publication of printed dictionary volumes commenced. After 15 years and several volumes covering about one fourth of the alphabet, the project became all-digital, to be published electronically online from 2010. Since then ONP Online has been improved and enhanced to its current form as a dynamic digital resource with several innovative features. The key to the success of the ONP project is meticulous registration of data and very detailed data management system. ONP sought inspiration from the Middle English Dictionary in its scope and attention to philological details, with great emphasis on the integrity of the source material. For the first forty years of the project a vast amount of data was collected, but management was limited to analogue methods, using paper and pen and several types of slips organized in filing cabinets. It was not until the computer age that it was possible to take full advantage of ONP‘s archives by digitizing the data and then later linking it together in different ways.
In this paper, I will give an overview of this development, identifying specific milestones in the evolution of the project and their significance. I will show how consistent data management and organization makes it possible through recent technological advancements to bring to light aspects of the data, which have not been discernible until now. This is evident by the latest online platform of ONP, launched in December 2019, which brings together the work of many generations of lexicographers and offers unprecedented look at all the data ONP has collected throughout its existence with extensive linking as well as several enhancements from other digital sources. This allows the user to interact with ONP and its source material in previously impossible ways, using

digital text editions, scanned manuscript pages, other lexicographic resources and even Google translate.

ELLERT THOR JOHANNSSON was born and raised in Reykjavik, Iceland, where he attended the University of Iceland, graduating in 1996 with a degree in Linguistics and Icelandic. After doing some post-graduate work in Iceland, he relocated to Ithaca, NY in 1998 to attend Cornell University where he taught Old Norse while pursuing graduate studies in Linguistics, receiving an MA and PhD degree in Historical linguistics, with focus on Old Germanic languages. In 2006 he was hired as a junior staff member at the Dictionary of Old Norse in Copenhagen and as an editor in 2012, where he has been involved in the digitization effort of the dictionary. He has in collaboration with the other editors of ONP written extensively about various aspects of the ONP-project and worked on promoting ONP as an important historical lexicographic resource.

‘Impersonal’ and ‘Reflexive’ Constructions: Verb Features peculiar to Old and Middle English.

Michiko Ogura
Chiba University

When Old English appeared in a written form for the first time, it had already lost inflections like optative, hortative, perfective, passive, etc. Making up for these morphological forms, it started, again before it was written and preserved, using periphrastic expressions with modal auxiliaries, *habban*, *beon/wesan*, *utan*, *ongan*, etc. Without having middle voice, it used ‘impersonal’ and ‘reflexive’ constructions (the single quotes mean that they included quasi-impersonals and quasi-reflexives in the real sense of the words). In this paper I focus on some such verbs as *lician*, *lystan*, *sceamian*, *þyncan* and *wer(g)ian* with their native and/or loan synonyms like *(dis)plesen*, *joien*, *remembren*, *repenten*, *semen*, etc. and their constructions used in Old and Middle English so as to maintain that their peculiar features reflect compensatory devices of the lost function before the appearance of Old English. The phonetic-morphological-syntactic merger, which occurred from late Old to early Middle English, is also discussed, in order to see the results of the rivalry of some synonyms into Modern English.

MICHIKO OGURA is Professor Emeritus, Chiba University, Japan. Her special field of study is Old and Middle English syntax and word studies. Her publications include *Old English ‘Impersonal’ Verbs and Expressions* (1986), *Verbs with the Reflexive Pronoun and Constructions with SELF in Old and Early Middle English* (1989), *Verbs in Medieval English: Differences in Verb Choice in Verse and Prose* (1996)*, Verbs of Motion in Medieval English* (2002), *Words and Expressions of Emotion in Medieval English* (2013), and *Periphrases in Medieval English* (2018). Her immediate concern is Christian terms in Old English and manuscript comparison between West Saxon Gospels.