

El gran proyecto lexicográfico neerlandés: *De historische woordenboeken van het Nederlands op internet**

Marta Gómez Martínez

Cilengua - Instituto Historia de la Lengua

Tanneke Schoonheim (1965) estudió Lengua y Literatura Neerlandesa en la Universidad de Leiden. Se especializó en Lingüística histórica y se graduó en 1986. Obtuvo el Grado de Doctora en 2004 con una tesis dedicada al estudio de los nombres femeninos de persona en neerlandés antiguo y medieval. Trabaja en el Instituto de Lexicología Neerlandesa desde 1986, primero como editora del Vroegmiddelnederlands Woordenboek (VMNW, Diccionario de Neerlandés Medieval Temprano), después como editora y editora-jefe del Oudnederlands Woordenboek (ONW, Diccionario de Neerlandés Antiguo), editora-jefe del Etymologisch Woordenboek van het Nederlands (EWN, Diccionario Etimológico del Neerlandés) y editora-jefe de la versión en línea del Woordenboek der Nederlandsche Taal (iWNT, Diccionario de Lengua Neerlandesa en línea).

Resumen: en enero de 2008 hizo un año desde que los diccionarios históricos neerlandeses, *De historische woordenboeken van het Nederlands op internet*, estuviesen disponibles en la web. Se trata de un proyecto llevado a cabo por el Instituto de Lexicología Neerlandesa de Leiden con financiación del Ministerio de Educación, Cultura y Ciencia y del Ministerio de la Comunidad Flamenca. Hasta la fecha se han informatizado dos de los diccionarios: *Woordenboek der Nederlandsche Taal op Internet* (iWNT) que recoge las voces del período que va desde 1500 a 1976 y *Vroegmiddelnederlands Woordenboek* (VMNW) de la época entre 1200 y 1300. En esta entrevista pretendemos hacer balance de lo que ha supuesto este proyecto para la lexicografía del neerlandés en el último año y mostrar los planes de trabajo futuro de este instituto. Además, de la mano de Tanneke Schoonheim veremos cómo se han enfrentado a la tarea de informatizar, y así unificar, compilaciones

* Gracias a la Prof.^a Tanneke Schoonheim por esta entrevista. Instituut voor Nederlandse Lexicologie afdeling Descriptieve Historische Projecten. Postbus 9515. NL - 2300 RA Leiden. schoonheim@inl.nl.

léxicas tan dispares realizadas originalmente en papel y por diversos investigadores.

MARTA GÓMEZ: *A bit of history on the Instituut voor Nederlandse Lexicologie (INL).*

TANNEKE SCHOONHEIM: The Instituut voor Nederlandse Lexicologie was founded in 1967. It is a Flemish-Dutch institute and board, employees and subsidies all come from Belgium and the Netherlands.

There are approximately 50 people working at INL, mainly lexicographers, (historical) linguists, computational linguists and lexicographical assistants. Their main activity is to collect and study Dutch words, to store them in databases –along with various additional data– and to use them to make scholarly dictionaries.

At the moment the main lexicographical projects at the INL are: the *Oudnederlands Woordenboek* (ONW, *Old Dutch Dictionary*), the *Etymologisch Woordenboek van het Nederlands* (EWN, *Etymological Dictionary of Dutch*), the *Algemeen Nederlands Woordenboek* (ANW, *General Dutch Dictionary*).

Two major historical dictionaries have already been completed: the *Vroegmiddelnederlands Woordenboek* (VMNW, *Dictionary of Early Middle Dutch*) and the *Woordenboek der Nederlandsche Taal* (WNT, *Dictionary of the Dutch Language*). Both dictionaries are available online at gtb.inl.nl.

The INL does not only make digital dictionaries; it also has several large digital text files (corpora) at its disposal. Linking these files to the dictionaries offers new research possibilities. That's why the INL is developing a large, integrated language database, a coherent whole in which both modern and historic, new and existing, dictionaries and corpora can be linked to each other.

We also manage and preserve digital language resources and facilitate their availability (tst.inl.nl).

M.G.: *One of the projects of this institute is the De historische woordenboeken van het Nederlands op internet which is the result of the process of digitalization of different dictionaries. Which ones have been put online? Could you please describe them shortly (period of time covered, number of lemmata, authors, date of composition)?*

T.S.: The project *De historische woordenboeken van het Nederlands op internet* will put the major historical dictionaries of the Dutch language online.

At this moment, both VMNW and WNT are already available. Many people have worked on these projects, both lexicographers and lexicographical assistants. The VMNW was compiled between 1988 and 1998 and has ca.26.000 headwords and ca.195.000 citations, covering Dutch language from the 13th century. The WNT was compiled between 1864 and 2001 and contains ca.450.000 headwords and related entries and ca.1.500.000 citations, covering the period from ca.1550 to 1976.

M.G.: *When will the other ones be released? Could you please describe them shortly (period of time covered, number of lemmata, authors, date of composition)?*

T.S.: The lexicographical work on the ONW started in 1999 and will be finished by the end of this year. The release on the Internet is expected for 2009, as is the online version of the *Middelnederlandsch Woordenboek* (MNW). The ONW, with an editorial team of three lexicographers and two lexicographical assistants, will contain ca.4.500 headwords and ca.41.500 citations and covers the period from ca.500 to 1200. The MNW, covering the period from ca.1250 to 1550, was compiled between 1885 and 1929 and contains ca.75.000 headwords and ca.400.000 citations.

M.G.: *What will be the total number of lemmata once all the dictionaries are online?*

TS: We do not yet have a precise idea of the degree of overlap between the different dictionaries. I would think the total number of headwords will be between 450.000 and 475.000, but only the future will tell.

M.G.: *It has been a year since the first ones were released online, what are the results so far?*

T.S.: The response was rather overwhelming. Both WNT and VMNW are scholarly dictionaries on mainly historical language, and we expected a limited, rather specialised kind of public. But a large number of people have registered on a personal basis (some 56.500 at this moment) and many universities, libraries and other major educational institutions have asked access via IP address.

M.G.: *What is the target audience or users of De historische woordenboeken van het Nederlands op internet?*

T.S.: As they all are scholarly dictionaries, the main target was people working with and studying historical language. However, since we added

modern Dutch equivalents to the headwords of all dictionaries, access to the entries is made possible without needing to know anything about historical spelling. Thus we have also reached a more general public.

M.G.: *Type of information offered by the dictionaries: definition, grammatical category; examples from the corpus...*

T.S.: The structure of all dictionary-entries is mainly the same. After the headword there is a section with linguistic information at entry level, containing information on grammar, etymology, morphology and spelling. Then follows a section with the semantic analysis of the headword, containing definitions and citations. Some of the dictionaries (WNT, ONW) also have a section with related entries.

All dictionaries have got an equivalent headword in modern Dutch spelling for easy retrieval purposes and combined querying and linking between them. A mapping was also done for the part-of-speech information in the different dictionaries, as the used terminology was not the same for all dictionaries. Abbreviated forms of compounds and derivatives were expanded, also for retrieval purposes.

M.G.: *About the corpus, how have the texts been selected? Do you have all the bibliographical references and dates?*

T.S.: That differs, the recent dictionaries VMNW and ONW are both based on an exhaustive, digital corpus, containing resp. 1.500.000 and 41.500 words. The WNT and MNW on the other hand have worked with manual excerption on snippets of paper, very much like all traditional dictionary projects like the OED and the *Deutsches Wörterbuch*. All citations have bibliographical references accompanied by a date.

M.G.: *Are there different kinds of texts in the corpus of the dictionaries, that is, texts that reflect the dialectal, technical-scientific,... vocabulary?*

T.S.: Yes, all dictionaries contain different kinds of texts, both literary texts, religious texts, administrative texts and texts for common usage. A difference is though that for the ONW and the VMNW we used all available material, since the oldest period provides us with only a limited amount of sources. For the WNT and the MNW out of the enormous amount of available material a selection was made.

M.G.: *Are these usages labelled? Is it possible to search for all the meanings that are under the same label (let's say, for instance, dialectal or scientific)?*

T.S.: Labelling has been done in all dictionaries, but for the different dictionaries in a different way. We have not yet put any effort into mapping this, but that might be done in the future.

M.G.: *Regarding other types of labels, such as chronological, are out of usage meanings or neologisms labelled?*

T.S.: The definitions in each and every dictionary are put in chronological order so as to illustrate the development of the meaning of the word. The citations illustrating the given semantic structure are also presented in a chronological order. Obsolete words are marked in all dictionaries, neologisms are found and labelled in the ANW.

M.G.: *After De historische woordenboeken van het Nederlands op internet, is there anything left to be done in Dutch lexicography?*

T.S.: There are plenty of things to be done after releasing all four historical dictionaries of Dutch. For instance, the EWN is to be finished by the end of 2009 and ANW in 2013, but actually, you should never stop making new dictionaries and improving the ones you've got.

We are currently also preparing new projects, like a diachronic morphological database of Dutch, and we are playing with the idea of a diachronic wordnet.



