Language change can be traced using gigantic text archives

June 26 2009

(PhysOrg.com) -- Historical collections that include everything ever written in a dozen American and British newspapers since they started are now available electronically. Donald MacQueen from Uppsala University, Sweden, has carried out the first comprehensive study that makes use of this resource in order to track changes in language usage, a method that makes it possible to attain an entirely new degree of precision in dating.

The gigantic newspaper archives contain news and feature articles as well as editorials and commercial and classified advertisements. Together they comprise tens of billions of words. In his dissertation in English linguistics, Donald MacQueen has examined the word million in English, especially how language usage shifted from the previously nearly totally dominant "five millions of inhabitants" to today's "five million inhabitants." With the help of these electronic collections of texts that only recently became available, he has succeeded in pinning down when and where the modern expression began to take over.

"When you study the occurrence of uncommon words in smaller corpora (text archives) of one or a few million words, you only get a few examples to analyze. These collections are much larger, and they have enabled me to obtain extremely reliable historical data for one year at a time. In this way I have been able to trace the shift with a precision that was not previously possible in linguistic studies," he explains.

It turns out that the modern construction took over in the American

newspapers in the middle of the 1880s and in the British The Times only in the mid 1910s. What's more, it became apparent that the transitional period was shorter in The Times. These circumstances indicate that usage in American newspapers influenced and accelerated the shift in the British newspaper.

This took place at the height of the British empire, and roughly when the US economy overtook the British for the first time. Donald MacQueen tentatively sees as an impetus for the change in usage, apart from the fact that both expressions suddenly began to be used more frequently, the greater propensity for people to embrace innovations during periods of severe social crisis, in this case the American Civil War and World War I, respectively. It is also possible that these wars entailed major population movements that could have impacted usage.

"Another discovery I made, thanks to the huge amount of data, is that when the use of the two constructions began to be roughly equal in frequency, the newspapers chose quite simply to avoid using such constructions, writing numeral expressions instead. After World War II, when there was no longer any doubt which construction was the ‘right' one, the newspapers reverted to writing number-word expressions again," he says.

The dissertation also includes a comparison with languages like French and German, where the corresponding grammatical shift regarding the word million from being a noun to an ordinary number word has not yet taken place.

"But in the long perspective we can expect this change to occur in those languages as well. The shift is a universal phenomenon when it comes to number words," says Donald MacQueen.

He defended his dissertation at Uppsala University on June 8.