Data linguistica 28

Steg för steg

Naturvetenskapligt ämnesspråk som räknas

av Judy Ribeck

Akademisk avhandling för filosofie doktorsexamen
i språkvetenskaplig databehandling,
som enligt beslut av humanistiska fakultetsnämnden
vid Göteborgs universitet kommer att försvaras offentligt
fredagen den 4 december 2015 kl. 13.15 i Lilla hörsalen, Humanisten.

GÖTEBORGS UNIVERSITET
HUMANISTISKA FAKULTETEN

Göteborg 2015

TITLE: Steg för steg
Naturvetenskapligt ämnesspråk som räknas
ENGLISH TITLE: Step by step
A computational analysis of Swedish textbook language
LANGUAGE: Swedish
AUTHOR: Judy Ribeck

Abstract

In this work, I present a linguistic investigation of the language of Swedish textbooks in the natural sciences, i.e., biology, physics and chemistry. The textbooks, which are used in secondary and upper secondary school, are examined with respect to traditional readability measures, e.g., LIX, OVIX and nominal ratio. I also extract typical linguistic features of the texts, typicality being determined using a proposed quantitative method, labelled *the index principle*. This empirical, corpus-based method relies on automatic linguistic annotations produced by language technology tools to calculate what I call *index lists*, rank-ordered lists of characteristic linguistic features of specific text corpora as compared to reference texts.

I produce index lists for typical vocabulary, noun phrase structures and syntactic structures, extracted from a 5.2 million word textbook corpus, compiled as a part of the work presented. As well as being frequent and well dispersed, the linguistic variables selected for the index lists are also characteristic of the text type in question, as is evident when they are compared to a reference corpus, comprising textbooks in the social sciences and mathematics, as well as narrative and academic (university-level) texts.

The results show that textbooks in natural science contain a lot of content-specific, technical vocabulary. This characteristic not only distinguishes natural scientific language from everyday language, but also from social scientific language, which on the lexical level has more in common with narrative texts. On the other hand, the textbook language as a whole is structurally distinguishable from narrative texts, as clearly seen, e.g., in its noun phrase complexity.

In the transition between secondary and upper secondary school, the scores of almost every readability measure go up, indicating an increase in linguistic demands on the readers. In the upper secondary textbooks the words are longer, the vocabulary more varied, the noun phrases longer and more elaborate, and the most typical syntactic structures more complex. Notably, the linguistic development between the form levels is more marked in the natural-science textbooks, compared to social sciences and mathematics. Nevertheless, the textbook language overall shows a relatively low complexity in comparison to academic language.

KEYWORDS: academic language, computational linguistics, corpus linguistics, language technology, natural language processing, scientific language, subject-specific language, Swedish textbooks, quantitative stylistics

DISTRIBUTION:
Department of Swedish
University of Gothenburg
Box 200
SE-405 30 Gothenburg, Sweden

Data linguistica 28
ISSN 0347-948X
ISBN 978-91-87850-59-2
E-publication http://hdl.handle.net/2077/40506