Data linguistica 24

I see what you mean

Assessing readability for specific target groups

av Katarina Heimann Mühlenbock

Akademisk avhandling för filosofie doktorsexamen i språkvetenskaplig databehandling vid Göteborgs universitet, som enligt beslut av humanistiska fakultetens dekanus kommer att försvaras offentligt fredagen den 26 april 2013 kl. 10.15 i Stora hörsalen, Humanisten.

GÖTEBORGS UNIVERSITET
HUMANISTISKA FAKULTETEN

Göteborg 2013

TITLE: I see what you mean
Assessing readability for specific target groups
LANGUAGE: English
AUTHOR: Katarina Heimann Mühlenbock

Abstract

This thesis aims to identify linguistic factors that affect readability and text comprehension, viewed as a function of text complexity. Features at various linguistic levels suggested in existing literature are evaluated, including the Swedish readability formula LIX. Natural language processing methods and resources are employed to investigate characteristics that go beyond traditional superficial measures.

A comparable corpus of easy-to-read and ordinary texts from three genres is investigated, and it is shown how features present at various levels of representation differ quantitatively across text types and genres. The findings are confirmed in significance tests as well as principal component analysis. Three machine learning algorithms are employed and evaluated in order to build a statistical model for text classification. The results demonstrate that a proposed language model for Swedish (SVIT), utilizing a combination of linguistic features, actually predicts text complexity and genre with a higher accuracy than LIX.

It is suggested that the SVIT language model should be adopted to assess surface language properties, vocabulary load, sentence structure, idea density levels as well as the personal interest of different texts. Specific target groups of readers may then be provided with materials tailored to their level of proficiency.

KEYWORDS: Readability, text complexity, computational linguistics, language resources, language technology, linguistic features, LIX, SVIT, corpus linguistics, text classification, quantitative methods, natural language processing, multilevel text analysis.

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

Data linguistica 24
ISSN 0347-948X
ISBN 978-91-87850-50-9
GUPEA <http://hdl.handle.net/2077/32472>

PRINTED in Sweden by Ineko AB Göteborg 2013