TEXT VARIATION EXPLORER

INTERACTIVE VISUALIZATION FOR CORPUS LINGUISTICS

DAMMOC

PROJECT FUNDED BY THE ACADEMY OF FINLAND FOR 2009–2011, LED BY:

- HEIKKI MANNILA, AALTO UNIVERSITY
  - ALGODAN – ALGORITHMIC DATA ANALYSIS (NATIONAL CENTRE OF EXCELLENCE)
- TERTTU NEVALAINEN, UNIVERSITY OF HELSINKI
  - VARIENG – RESEARCH UNIT FOR VARIATION, CONTACTS AND CHANGE IN ENGLISH (NATIONAL CENTRE OF EXCELLENCE)
- KARI-JOUKO RÄIHÄ, UNIVERSITY OF TAMPERE
  - TAUCHI – TAMPERE UNIT FOR COMPUTER-HUMAN INTERACTION

PEOPLE DEALING WITH LANGUAGE DATA NEED TOOLS FOR LOOKING INSIDE TEXTS

- FOR QUICKLY FINDING OUT MORE ABOUT TEXTS AND TEXT CORPORA
- FOR EASILY COMPARING HOW TEXTS/CORPORA DIFFER FROM EACH OTHER

TVE, A SIMPLE BUT FLEXIBLE TOOL FOR EXPLORATORY DATA ANALYSIS,

- HELPS YOU CHOOSE DATA FOR ANALYSIS
- HELPS YOU INTERPRET YOUR RESULTS

TVE PROVIDES LINE GRAPHS OF 3 COMMON TEXT MEASURES THAT DESCRIBE THE VOCABULARY RICHNESS AND STYLE OF ANY TEXT

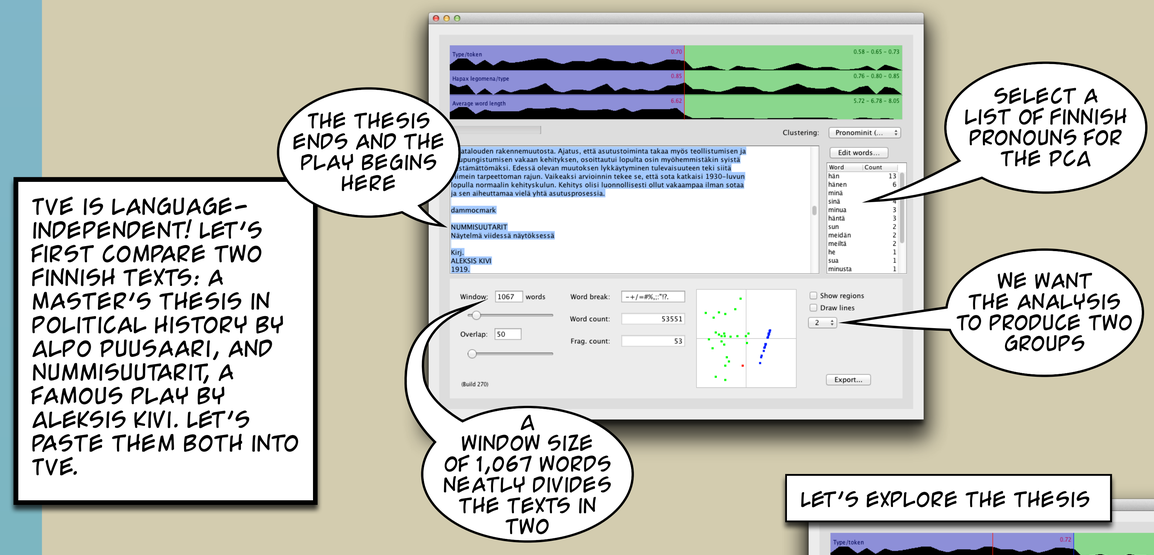
- TYPE-TOKEN RATIO (TTR)
- PROPORTION OF WORDS THAT ONLY OCCUR ONCE (HAPAX LEGOMENA)
- AVERAGE WORD LENGTH

TVE ALSO CLUSTERS TEXT FRAGMENTS ACCORDING TO A GIVEN SET OF WORDS (PRINCIPAL COMPONENT ANALYSIS)

- THE PCA VIEW DISPLAYS EACH TEXT FRAGMENT AS A POINT AND SHOWS THE VALUES OF THE FIRST TWO PCS FOR THE FRAGMENT

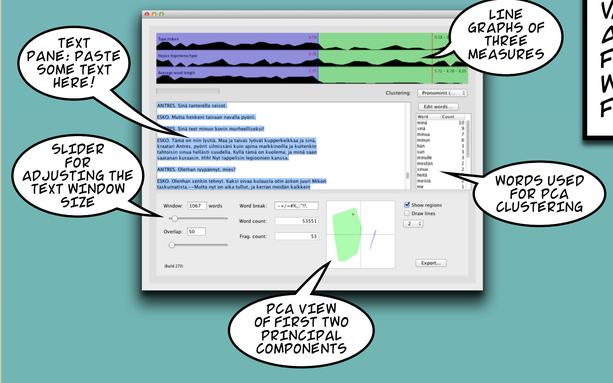
TVE ALLOWS THREE-WAY BRUSHING: CLICK A POINT ON ANY OF THE VIEWS (TEXT, LINE GRAPH, PCA) AND THE OTHER TWO WILL UPDATE TO SHOW THE RELEVANT PART

COMPARING A THESIS WITH NUMMISUUTARIT

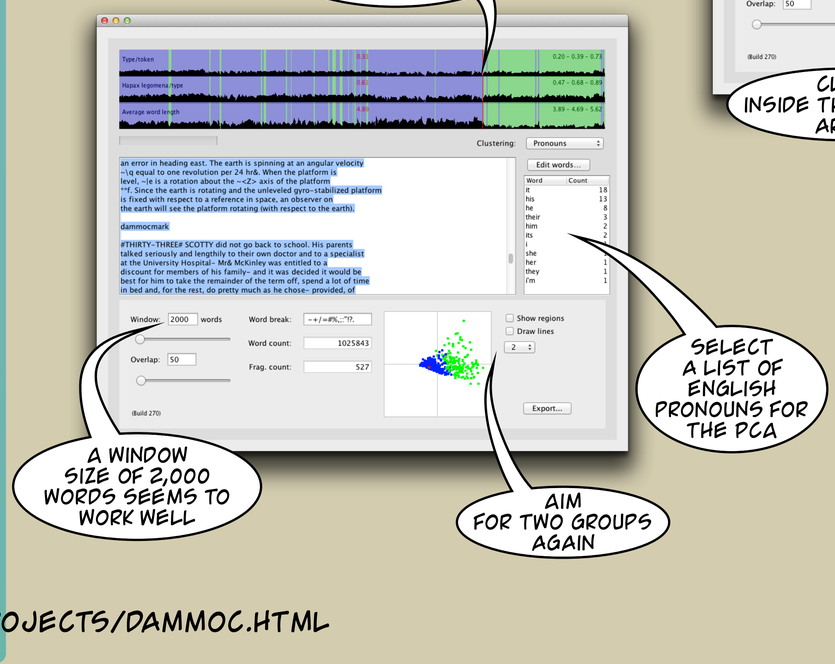


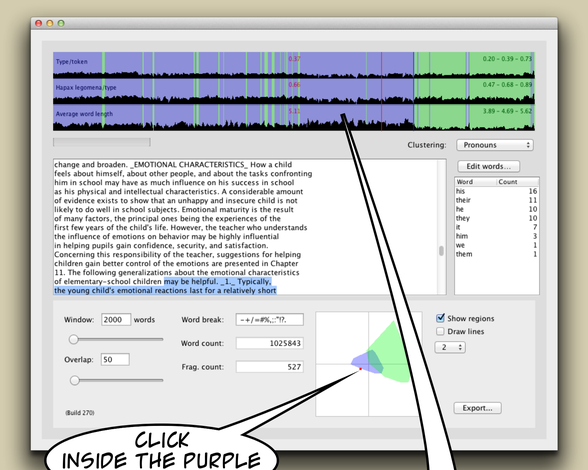
FACT AND FICTION IN THE BROWN CORPUS

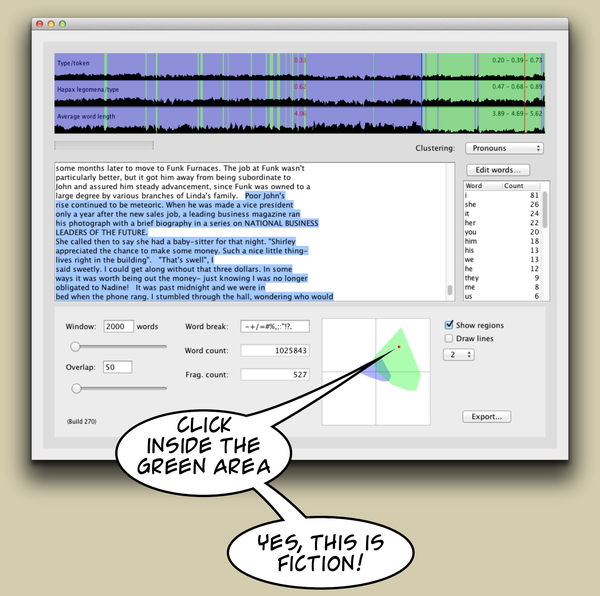
TVE USER INTERFACE



NOW LET'S EXPLORE THE BROWN CORPUS, WHICH CONTAINS VARIOUS GENRES OF AMERICAN ENGLISH FROM THE 1960S. CAN WE DISTINGUISH FACT FROM FICTION?







IN BOTH ENGLISH AND FINNISH, FICTION CAN BE SEPARATED FROM NON-FICTION WITH THE HELP OF PERSONAL PRONOUNS AND AVERAGE WORD LENGTH.

PROJECT WEB PAGE:

HTTP://WWW.UTA.FI/SIS/TAUCHI/VIRG/PROJECTS/DAMMOC.HTML