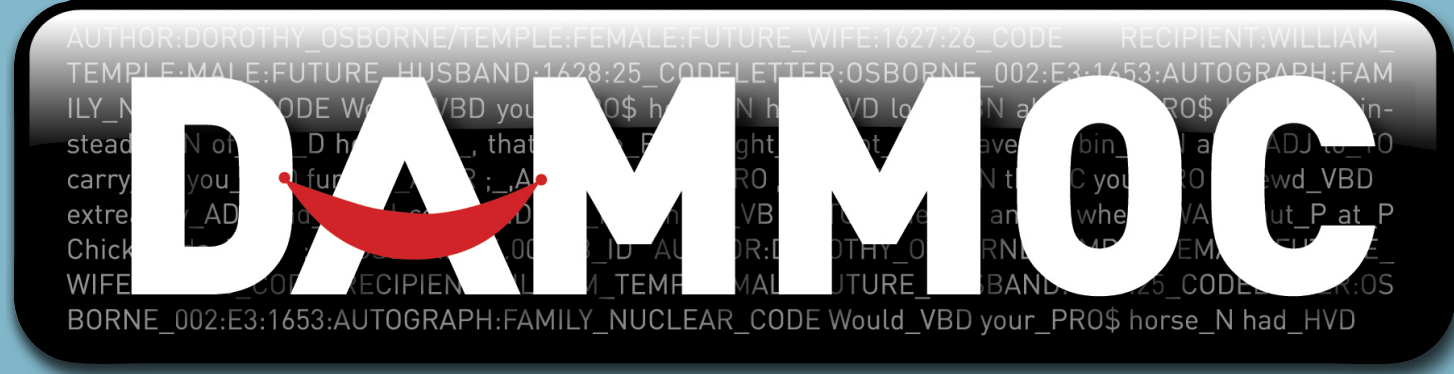


TEXT VARIATION EXPLORER

INTERACTIVE VISUALIZATION FOR CORPUS LINGUISTICS



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 RÄIHA, 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

TVE IS LANGUAGE-INDEPENDENT! LET'S FIRST COMPARE TWO FINNISH TEXTS: A MASTER'S THESIS IN POLITICAL HISTORY BY ALPO PUUSAARI, AND NUMMISUUTARIT, A FAMOUS PLAY BY ALEKSIS KIVI. LET'S PASTE THEM BOTH INTO TVE.

THE THESIS ENDS AND THE PLAY BEGINS HERE

SELECT A LIST OF FINNISH PRONOUNS FOR THE PCA

WE WANT THE ANALYSIS TO PRODUCE TWO GROUPS

A WINDOW SIZE OF 1,067 WORDS NEATLY DIVIDES THE TEXTS IN TWO

LET'S EXPLORE THE THESIS

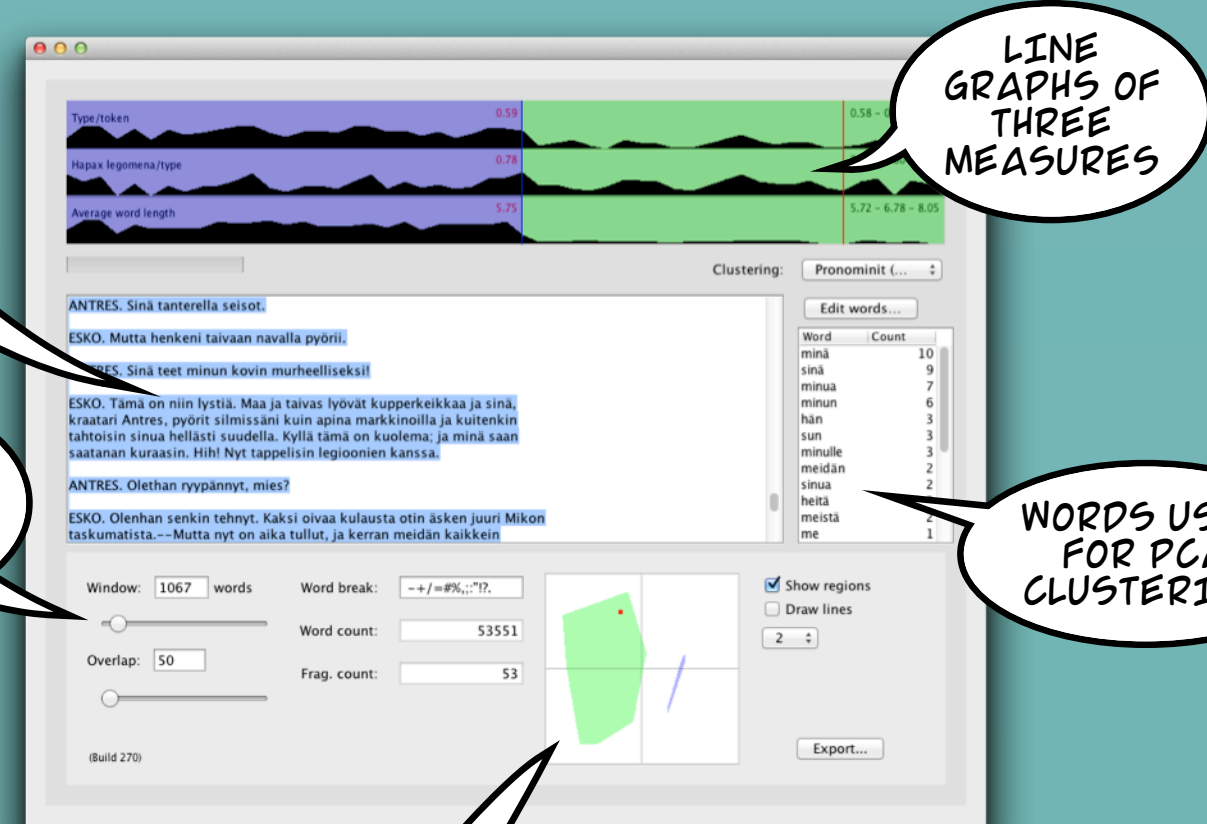
LONGER WORDS, HIGHER TYPE-TOKEN RATIO

LET'S SEE HOW THE PLAY LOOKS

SHORTER WORDS, LOWER TYPE-TOKEN RATIO

FACT AND FICTION IN THE BROWN CORPUS

TVE USER INTERFACE



TEXT PANE: PASTE SOME TEXT HERE!

SLIDER FOR ADJUSTING THE TEXT WINDOW SIZE

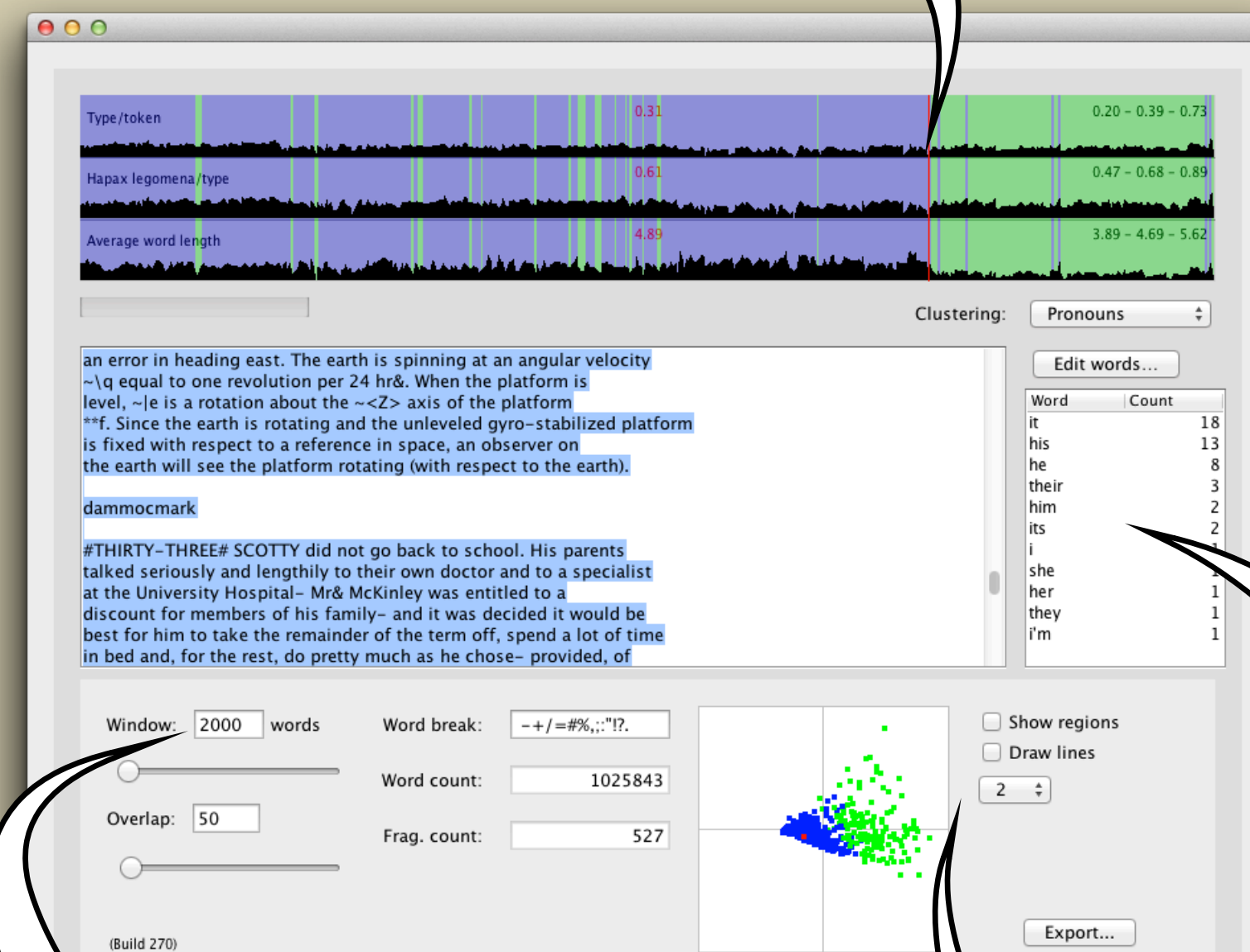
LINE GRAPHS OF THREE MEASURES

WORDS USED FOR PCA CLUSTERING

PCA VIEW OF FIRST TWO PRINCIPAL COMPONENTS

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

THIS SEEMS TO BE WHERE NON-FICTION ENDS AND FICTION BEGINS



A WINDOW SIZE OF 2,000 WORDS SEEMS TO WORK WELL

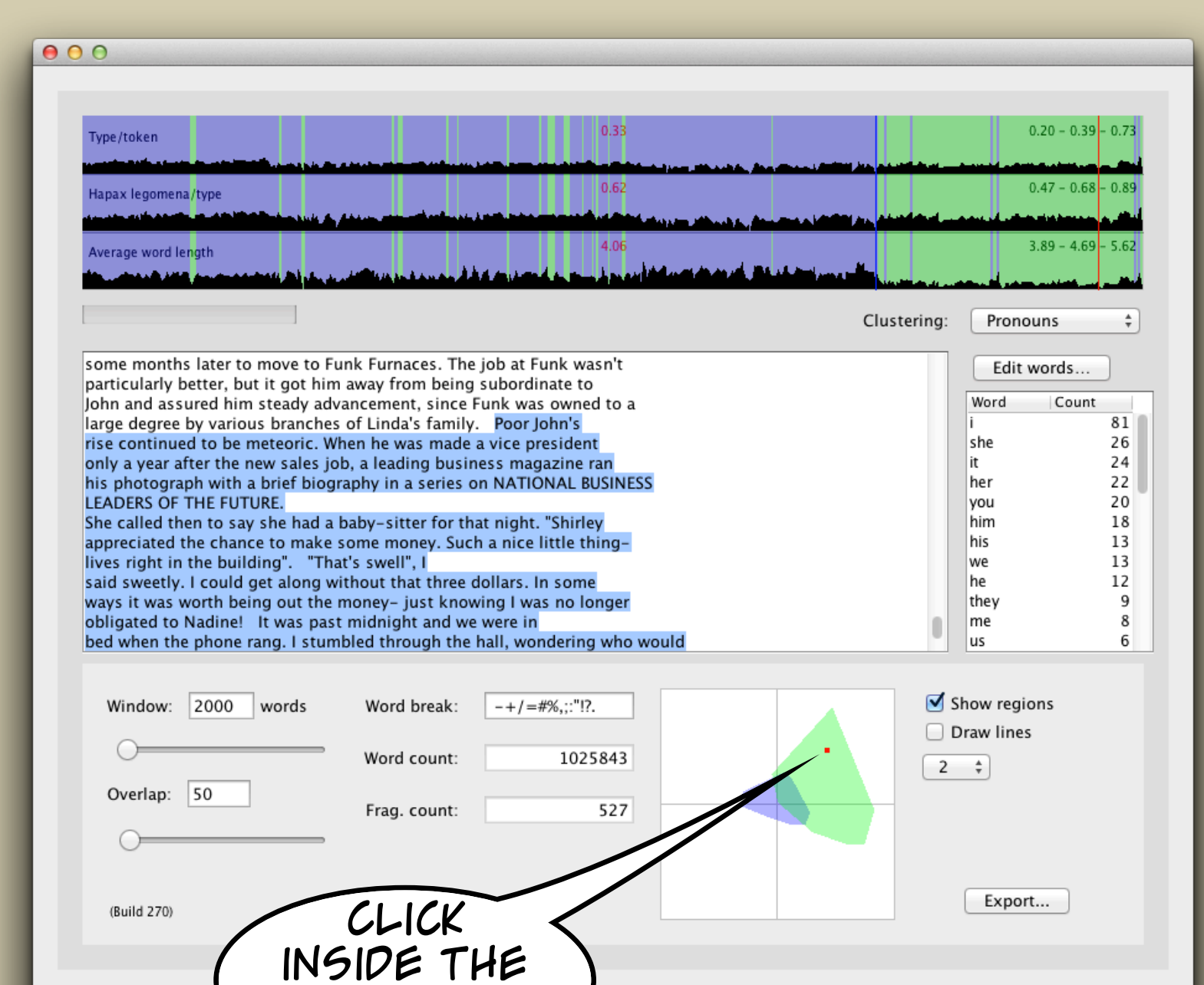
AIM FOR TWO GROUPS AGAIN

SELECT A LIST OF ENGLISH PRONOUNS FOR THE PCA

CLICK INSIDE THE PURPLE AREA

THIS IS CLEARLY NON-FICTION

THE AVERAGE WORD LENGTH IS HIGHER ON THIS SIDE



CLICK INSIDE THE GREEN AREA

YES, THIS IS 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](http://www.uta.fi/sis/tauchi/virg/projects/dammoc.html)

