HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI
HUMANISTINEN TIEDEKUNTA
HUMANISTISKA FAKULTETEN
FACULTY OF ARTS

Normalizing Early English Letters for Neologism Retrieval

Mika Hämäläinen,
Tanja Säily
Eetu Mäkelä

Department of Digital Humanities

INTRODUCTION

In this abstract we describe the current state of our method for detecting neologisms. The problem we are facing at the moment is the fact that our corpus consists of **non-normalized text**. Therefore, normalization is the first step we need to solve before we can apply automatic methods to the whole corpus.

CORPUS

We use **CEEC (Corpora of Early English Correspondence)** as the corpus for our research.

- letters ranging from the 15th century to the 19th century
- a wide social spectrum, richly documented in the metadata information
- e.g. socioeconomic status, gender, age, domicile
- the relationship between the writer and recipient.

DIFFERENT APPROACHES

For the non-normalized words, we have tried a number of different approaches.

- Rules
- SMT
- NMT
- Edit distance, semantics and pronunciation

Hand-written **VARD2 normalization rules**

We have also trained a **statistical machine translation model (with Moses)** and a **neural machine translation model (with Open NMT).**

- models are character based
- the known non-normalized to normalized word pairs
- The language model is the British National Corpus (BNC).

One more approach

- Levenshtein edit distance.
- Filtering by semantic similarity
- Soundex pronunciation by edit distance.