Phonetic Similarity in Writing
A Computational Model of Orthographic Productivity

Philip Reuter
preuter@uos.de

Orthographic Knowledge
- Like morphology or syntax, orthography is part of a language’s grammar
- Orthographies (writing systems) are usually strictly normed, so a writer has almost no flexibility in representing a word form graphically
- Thus, orthographic knowledge is highly lexicalized: When writing some word form, its (ortho-)graphic representation is accessed in a mental lexicon

Productive Orthographic Processes
- Writing competence includes some productive mechanism: We know how to write novel or unknown words
- Example: While planning a workshop, sb. is doing phone calls to book a room. Afterwards, he writes:

\[
\text{Daher habe ich mir den Cemilienaal (schreibt man das so?) reservieren lassen.}
\]

[So I booked Cemilienaal (do you write it like this?)]

Research Question
- What kind of knowledge is used in productive orthographic processes?
- One answer is: Phonetic similarity between novel and known word forms
- How can this process be modelled computationally?

Approach
basis
- phonetic word form (what is to be written, e.g. [tsa.'me:ljan])
- orthographic word form (what is written, e.g. <Cemilien>)
- mental lexicon as list of (phonetic, orthographic) word form pairs (e.g. ([fa.'mi:ljan], <Familien>, ...)

model phonetic similarity
- phonetic word form as sequence of phonemes
- apply sequence alignment algorithm: minimum edit distance
- phoneme as n-dimensional vector; dimensions are phonetic features; e.g.
  - [p]: [place: bilabial, manner: plosive, voice: no]
  - [z]: [place: alveolar, manner: fricative, voice: yes]
  - [p]: [0, 0, 0]
  - [z]: [0.3, 0.57, 1]
- edit distance is some metric on phoneme vector (e.g. euclid)
- low edit distance between phon. forms si considered as high similarity

search strategy
- organize (mental) lexicon in a tree structure
- perform a k-nearest-neighbour search

Next Steps
phonetic similarity measure
- try different feature representations
- apply different metrics
- test on empirical data to find optimal measure

mental lexicon
- try different size: basic words vs. large full-form lexicon
- vary search strategy: time efficiency vs. max. quality
- apply to empirical data
- output of modeling is numerical – how to interpret?
- competition with different approaches (e.g. orthographic productivity as phoneme-to-grapheme-conversion)