Orthography in Language Modeling of Mutual Intelligibility

Andrea Fischer, Klára Jágrrová, Irina Stenger
Tania Avgustinova, Dietrich Klakow, Roland Marti

Meaningful Units of Language

Certain constructions encode specific information
possibly stark differences between languages
→ fundamental advance in computational linguistics research

Decoding as Domain Adaptation

Explicit “latent” space describing each language
→ discover informative elements of natural language

Modeling: Language as Domain

Basic idea: surprisal of statistical n-gram language models correlates with cognitive effort, but n-grams need to be adapted to process a different language

Preliminary Results: Orthography

Diachronically-based assumptions tested on parallel list of Pan-Slavic vocabulary for each language pair (high cognate rate)

Applicability of Diachronically-Based Rules

1) Orthographically identical words (8.79% in CZ-PL vs. 21.25% in BG-RU),
2) Application of transformation rules on remaining word pairs (91.21% vs. 79.75%), but
3) not all word pairs could be covered by rules: morphological differences
→ will be explored in next project phase
→ rules also tested on other word sets (internationalism)
Swadesh lists with wider vocabulary range/ higher (non-)cognate rates

Models are sensitive to individual texts and scores comprehensibility as conformity to alphabet usage of L1

Next Steps

Linguistically:
Lexis: "false friends" and closed word classes
Morphology: correspondences in grammar
Syntax: word order, complexity of constructions

Information-Theoretically:
Suitable model classes
Most informative features
Inter/Intra-language patterns

Orthography in Language Modeling of Mutual Intelligibility

Andrea Fischer, Klára Jágrrová, Irina Stenger
Tania Avgustinova, Dietrich Klakow, Roland Marti

Meaningful Units of Language

Certain constructions encode specific information
possibly stark differences between languages
→ fundamental advance in computational linguistics research

Decoding as Domain Adaptation

Explicit “latent” space describing each language
→ discover informative elements of natural language

Modeling: Language as Domain

Basic idea: surprisal of statistical n-gram language models correlates with cognitive effort, but n-grams need to be adapted to process a different language

Preliminary Results: Orthography

Diachronically-based assumptions tested on parallel list of Pan-Slavic vocabulary for each language pair (high cognate rate)

Applicability of Diachronically-Based Rules

1) Orthographically identical words (8.79% in CZ-PL vs. 21.25% in BG-RU),
2) Application of transformation rules on remaining word pairs (91.21% vs. 79.75%), but
3) not all word pairs could be covered by rules: morphological differences
→ will be explored in next project phase
→ rules also tested on other word sets (internationalism)
Swadesh lists with wider vocabulary range/ higher (non-)cognate rates

Models are sensitive to individual texts and scores comprehensibility as conformity to alphabet usage of L1

Next Steps

Linguistically:
Lexis: "false friends" and closed word classes
Morphology: correspondences in grammar
Syntax: word order, complexity of constructions

Information-Theoretically:
Suitable model classes
Most informative features
Inter/Intra-language patterns

Orthography in Language Modeling of Mutual Intelligibility

Andrea Fischer, Klára Jágrrová, Irina Stenger
Tania Avgustinova, Dietrich Klakow, Roland Marti

Meaningful Units of Language

Certain constructions encode specific information
possibly stark differences between languages
→ fundamental advance in computational linguistics research

Decoding as Domain Adaptation

Explicit “latent” space describing each language
→ discover informative elements of natural language

Modeling: Language as Domain

Basic idea: surprisal of statistical n-gram language models correlates with cognitive effort, but n-grams need to be adapted to process a different language

Preliminary Results: Orthography

Diachronically-based assumptions tested on parallel list of Pan-Slavic vocabulary for each language pair (high cognate rate)

Applicability of Diachronically-Based Rules

1) Orthographically identical words (8.79% in CZ-PL vs. 21.25% in BG-RU),
2) Application of transformation rules on remaining word pairs (91.21% vs. 79.75%), but
3) not all word pairs could be covered by rules: morphological differences
→ will be explored in next project phase
→ rules also tested on other word sets (internationalism)
Swadesh lists with wider vocabulary range/ higher (non-)cognate rates

Models are sensitive to individual texts and scores comprehensibility as conformity to alphabet usage of L1

Next Steps

Linguistically:
Lexis: "false friends" and closed word classes
Morphology: correspondences in grammar
Syntax: word order, complexity of constructions

Information-Theoretically:
Suitable model classes
Most informative features
Inter/Intra-language patterns

Orthography in Language Modeling of Mutual Intelligibility

Andrea Fischer, Klára Jágrrová, Irina Stenger
Tania Avgustinova, Dietrich Klakow, Roland Marti

Meaningful Units of Language

Certain constructions encode specific information
possibly stark differences between languages
→ fundamental advance in computational linguistics research

Decoding as Domain Adaptation

Explicit “latent” space describing each language
→ discover informative elements of natural language

Modeling: Language as Domain

Basic idea: surprisal of statistical n-gram language models correlates with cognitive effort, but n-grams need to be adapted to process a different language

Preliminary Results: Orthography

Diachronically-based assumptions tested on parallel list of Pan-Slavic vocabulary for each language pair (high cognate rate)

Applicability of Diachronically-Based Rules

1) Orthographically identical words (8.79% in CZ-PL vs. 21.25% in BG-RU),
2) Application of transformation rules on remaining word pairs (91.21% vs. 79.75%), but
3) not all word pairs could be covered by rules: morphological differences
→ will be explored in next project phase
→ rules also tested on other word sets (internationalism)
Swadesh lists with wider vocabulary range/ higher (non-)cognate rates

Models are sensitive to individual texts and scores comprehensibility as conformity to alphabet usage of L1

Next Steps

Linguistically:
Lexis: "false friends" and closed word classes
Morphology: correspondences in grammar
Syntax: word order, complexity of constructions

Information-Theoretically:
Suitable model classes
Most informative features
Inter/Intra-language patterns

Orthography in Language Modeling of Mutual Intelligibility

Andrea Fischer, Klára Jágrrová, Irina Stenger
Tania Avgustinova, Dietrich Klakow, Roland Marti

Meaningful Units of Language

Certain constructions encode specific information
possibly stark differences between languages
→ fundamental advance in computational linguistics research

Decoding as Domain Adaptation

Explicit “latent” space describing each language
→ discover informative elements of natural language

Modeling: Language as Domain

Basic idea: surprisal of statistical n-gram language models correlates with cognitive effort, but n-grams need to be adapted to process a different language

Preliminary Results: Orthography

Diachronically-based assumptions tested on parallel list of Pan-Slavic vocabulary for each language pair (high cognate rate)

Applicability of Diachronically-Based Rules

1) Orthographically identical words (8.79% in CZ-PL vs. 21.25% in BG-RU),
2) Application of transformation rules on remaining word pairs (91.21% vs. 79.75%), but
3) not all word pairs could be covered by rules: morphological differences
→ will be explored in next project phase
→ rules also tested on other word sets (internationalism)
Swadesh lists with wider vocabulary range/ higher (non-)cognate rates

Models are sensitive to individual texts and scores comprehensibility as conformity to alphabet usage of L1

Next Steps

Linguistically:
Lexis: "false friends" and closed word classes
Morphology: correspondences in grammar
Syntax: word order, complexity of constructions

Information-Theoretically:
Suitable model classes
Most informative features
Inter/Intra-language patterns

Orthography in Language Modeling of Mutual Intelligibility

Andrea Fischer, Klára Jágrrová, Irina Stenger
Tania Avgustinova, Dietrich Klakow, Roland Marti

Meaningful Units of Language

Certain constructions encode specific information
possibly stark differences between languages
→ fundamental advance in computational linguistics research

Decoding as Domain Adaptation

Explicit “latent” space describing each language
→ discover informative elements of natural language

Modeling: Language as Domain

Basic idea: surprisal of statistical n-gram language models correlates with cognitive effort, but n-grams need to be adapted to process a different language

Preliminary Results: Orthography

Diachronically-based assumptions tested on parallel list of Pan-Slavic vocabulary for each language pair (high cognate rate)

Applicability of Diachronically-Based Rules

1) Orthographically identical words (8.79% in CZ-PL vs. 21.25% in BG-RU),
2) Application of transformation rules on remaining word pairs (91.21% vs. 79.75%), but
3) not all word pairs could be covered by rules: morphological differences
→ will be explored in next project phase
→ rules also tested on other word sets (internationalism)
Swadesh lists with wider vocabulary range/ higher (non-)cognate rates

Models are sensitive to individual texts and scores comprehensibility as conformity to alphabet usage of L1

Next Steps

Linguistically:
Lexis: "false friends" and closed word classes
Morphology: correspondences in grammar
Syntax: word order, complexity of constructions

Information-Theoretically:
Suitable model classes
Most informative features
Inter/Intra-language patterns