Case and Content: A Cross-Linguistic Corpus Study
Drew Reisinger (reisinger@cogsci.jhu.edu)
Johns Hopkins University

Overview
Grimm (2005, 2011)’s hypothesis: Proto-role entailments of verbal arguments determine morphological case marking
Proto-role projection hypothesis (NEW): aligned arguments in a translated sentence pair have the same proto-role entailments
Evaluate these hypotheses in English and Czech using automatic alignment and computational modeling

Motivation
• Manually annotating an SPR corpus = expensive, time-consuming, infeasible for low-resource and endangered languages
• Cross-lingual projection techniques leverage resources in well-supported languages (English) to generate datasets/models in other languages (e.g., Pado & Lapata 2005, 2009)
• For this project, Czech is the target language
• Slavic case system complex enough to test Grimm’s hypothesis
• Prague Czech-English Dependency Treebank (PCEDT), a large English/Czech parallel corpus (Hajič et al. 2012)

Background
• Dowty (1991)’s thematic proto-roles
• Thematic roles replaced with sets of proto-role entailments, e.g., VOLITION and CHANGE OF STATE
• The argument with the most proto-Agent entailments is realized as the subject, and the one with the most proto-Patient entailments the object.

Projection by Alignment

<table>
<thead>
<tr>
<th>Výroba byla včera obnovena</th>
<th>Production resumed yesterday</th>
</tr>
</thead>
</table>

Error Analysis
• Due to case annotation errors in PCEDT caused by synonymy
1. Finanční ředitel William J. Montgoris závěratně znížil výchozí příjmy a náklady, že [...] (model guesses ACC)
2. “William Montgoris, chief financial officer, defended the lofty salaries at Bear Stearns.”
• Due to alignment errors
3. Výroba byla včera obnovena (model guesses ACC)
4. Production resumed yesterday.
• Due to failures of modeling assumptions (passive, pro-drop)
5. “Production resumed yesterday.”
• Due to learning limitations? Skewed distribution of cases in data?
6. Tento dopis přišel o podnictví ministerstv uchazečů za druhou příchu, že [...] (model guesses NOM)
7. “The letter, along with a detailed prosecution memo, sat on the desks of justice officials [...]”

Experiments
• Predict case on projected Czech SPRL corpus
• Predict grammatical function on English Universal Dependencies treebank (Nivre et al. 2015) annotated with SPRL (White et al. 2016)

Results

<table>
<thead>
<tr>
<th>Model accuracy on English UD</th>
<th>Model accuracy on Czech MSJ</th>
</tr>
</thead>
</table>

Conclusions
• MT alignment models fail to capture shared predicate-argument structure in translation pairs, reflect distinct objectives
• Corpus verification of Grimm must confront skewed distribution of cases in real data
• Other morphosyntactic phenomena (pro-drop, valency change, Japanese topic/focus markers, Tagalog case system) challenge Grimm’s approach