Researching legal translation with corpora: mixed-methods opportunities and challenges

Dr hab. Łucja Biel
Institute of Applied Linguistics
University of Warsaw
l.biel@uw.edu.pl
**Corpus-Based Translation Studies (CBTS)**

Corpus Translation Studies (CTS)

- CBTS – applies corpus linguistics as a method to study translations
Translation-induced variation (translationese, transgenre)

- **Translations as a distinct genre** falling under the category of internal variation
  - Literary translation - a distinct literary genre governed by its own norms and purposes (Ortega y Gasset 2004[1937]: 61; James 1989: 35); literary translation implies loss and it “is not the work, but a path toward the work” (Ortega y Gasset 2004[1937]: 61)
  - **Translationese** – M. Baker (1990s) as a product of translation process & **translation universals/features**, first mainly literary translation
  - **Transgenre** a genre which is ‘exclusive to translation’ and differs from the source genre and target genre (Borja et al. 2009: 62, 68)
Development of LTS thanks to corpora

- Legal translation – highly interdisciplinary field; systemic nature of law, language-pair-specific
- Expansion of LTS in the late 2000s into new methodologies, new angles and areas of interest
- Methodological development of the field through an increased methodological awareness, **rigour** and methodological **eclecticism**
  - Testing and triangulating diverse methods
  - **Shift from qualitative only to empirical, quantitative and mixed approaches**
  - Shift from prescriptive to **descriptive**
- **Corpora** as one of the main methods, new types of data: n-grams, keywords, collocations
Availability of legal corpora

- underdeveloped corpus resources, in particular for lesser used languages
- limited accessibility of corpus resources
- confidentiality of legal texts and related corpus design limitations → legicentrism, representativeness and balance as regards genres
- but the pool of resources is growing
Legal corpora

Monolingual
- BoLC Bononia Legal Corpus
- JuReko German Legal Reference Corpus
- Polish Law Corpus
- Cambridge Corpus of Legal English
- Old Bailey Corpus

Multilingual
- Bilingwis Swiss Law Text Collection (Höfler and Sugisaki 2014)

Multilingual - institutional resources; for derived products, e.g. translation memories, machine translation systems, term ontologies (Heylen et al. 2014: 9)
- The United Nations Parallel Corpus v1.0 - a multilingual parallel corpus, 6 lgs, released in 2016
- The EU: the JRC-Acquis parallel corpus; DGT-Acquis; translation memories of the DGT, EAC-TM and ECDC-TM; the Digital Corpus of the European Parliament DCEP; Europarl and the EUR-Lex corpus on Sketchengine (24 languages, 840m tokens of English, 3.9m documents (Baisa et al. 2016))
Theoretically-oriented research

- **K. McAuliffe’s** ERC Starting Grant **LLECJ** (*Law and Language at the European Court of Justice*)
- **F. Prieto Ramos’s** **LETRINT** project (*Legal Translation in International Institutional Settings: Scope, Strategies and Quality Markers*)
- **L. Mori’s** **Eurolect Observatory** project to analyse 11 Eurolects through a corpus of EU directives
- **Ł. Biel’s** **Polish Eurolect** project involving a genre-based analysis of the Polish Eurolect across four genres (legislation, judgments, reports, websites for citizens) and its impact on administrative Polish
- **G. Pontrandolfo’s** Ph.D. dissertation on the contrastive analysis of Spanish, Italian and English phraseology in translated EU judgments and nontranslated national judgments (2016)
- **M. Orozco-Jutoran, C. Bestue** (Autonomous University of Barcelona) corpus of interpreting at criminal proceedings
- **Bergen Translation Corpus (BTC)** of Translations from the Norwegian National Translator Accreditation Exam (Simmonaes, 2012)
Practically-oriented applied research

Developing resources for legal practitioners

- **Qualetra project** (Kockaert et al.) on training, assessment, certification and accreditation of legal translators in criminal proceedings
- the **GENTT** (Textual Genres for Translation) Research Group (Borja Albi 2013, Borja Albi et al. 2014) - the JudGENTT web platform
- Torres-Hostench and Bestué Salinas’ **LAW10n** project on terminology in software licensing agreements (2015)
- **TermWise** – a CAT-tool cloud-based database derived from the Dutch-French parallel corpus (Heylen et al. 2014)
Variation and genres

- Corpus studies into translation show that differences between translations and nontranslations (in particular features of translations) are dependent on genres (Teich 2003: 147; Delaere et al. 2012, de Sutter et al. 2012).

- Studies into legal language have demonstrated a high variation of lexical bundles across legal genres (cf. Goźdź-Roszkowski 2011)

- Intensity of translation effects across genres
# Variation in legal language

<table>
<thead>
<tr>
<th>Trajectory 1</th>
<th><strong>External variation</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>How does legal language differ from general language and other languages for special purposes?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trajectory 2</th>
<th><strong>Internal variation</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>How do legal genres/sub-genres differ from each other?</td>
</tr>
<tr>
<td></td>
<td>Description of legal genres (macroanalysis and microanalysis)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trajectory 3</th>
<th><strong>Diachronic variation/change</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>How does the current legal language differ from a historic one?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trajectory 4</th>
<th><strong>Cross-linguistic variation</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>How do legal genres differ across languages?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trajectory 5</th>
<th><strong>Idiosyncratic/individual variation</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Studies within forensic linguistics to provide linguistic evidence in witness statements, blackmail or suicide notes, as well as to attribute authorship and detect plagiarism by identifying the ‘linguistic fingerprint’ of individuals (cf. McEnery et al. 2006: 116; Stubbs 2004: 124).</td>
</tr>
</tbody>
</table>
The Eurolect project

• *The Eurolect: an EU variant of Polish and its impact on administrative Polish*: Sonata BIS grant, National Science Centre, 2015-2019)

• A follow-up of the Eurofog project, which has demonstrated divergent textual fit; improved comparability

• The team: Łucja Biel, Dariusz Koźbiał, Katarzyna Wasilewska
The Polish Eurolect

• The Polish Eurolect is a new linguistic phenomenon, i.e. a hybrid variant of the Polish language used in the highly multilingual EU context as one of the official EU languages and emerging via translation
• A translator-mediated variant of Polish
Objectives of the project

1. to extensively investigate the Eurolect to understand the processes and factors behind its formation
2. to track the impact of the Eurolect on post-accession Polish
Research questions - variation

1. **Internal variation (textual fit/transgenre):** How does the eurolect differ from naturally occurring administrative Polish?

2. **Internal variation - cross-generic:** How does the eurolect differ internally across four genres (legislation, judgments, reports, official websites for citizens)?

3. **Variables:** How is the eurolect affected by the genre, the source language, institutionalisation of the translation process, the translator profile and translation universals?

4. **Diachronic variation → Europeanisation of administrative Polish:** How has post-accession Polish been affected by the huge inflow of EU translations (comparison of pre-accession Polish (1999/2000) and post-accession Polish (2015))?
Corpus design: genre-based comparable-parallel corpora
Sampling frame 2011-2015

- Corresponding English texts
- Polish Eurolect corpus
- Administrative Polish
- National Corpus of Polish

- Genre-based parallel corpus
  \((EU\ EN)\)
- Genre-based Eurolect corpus
  \((EU\ PL)\)
- Genre-based reference corpus
  \((PL)\)
- Large general reference corpus
  \((PL)\)
Generic composition of corpora

- Legislation – enacting terms
- Judgments
- Institutional reports
- Websites for citizens

Time frame:
- Europeanisation of the administrative Polish: 1999 v 2015
## Characteristics of genres

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Legislation</th>
<th>Judgments</th>
<th>Reports</th>
<th>Websites</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Subgenres</strong></td>
<td>regulations, directives, decisions, recommendations Category A</td>
<td>judgments, orders, decisions Category A</td>
<td>administrative reports Category B</td>
<td>hypergenre, hybrid of genres Category C</td>
</tr>
<tr>
<td><strong>Function</strong></td>
<td>prescriptive, normative, constitutive</td>
<td>interpretative, regulative</td>
<td>informative, promotional</td>
<td>informative, promotional</td>
</tr>
<tr>
<td></td>
<td>Prescribe law, impose obligations and create rights</td>
<td>Interpret law, ensure uniform application of EU law</td>
<td>Inform experts, create a positive image of an institution</td>
<td>Explain; inform about rights, create a positive image of the EU</td>
</tr>
<tr>
<td><strong>Discourse community</strong></td>
<td>Expert-expert</td>
<td>Expert-expert</td>
<td>Expert-expert</td>
<td>Expert-lay; institution-citizen</td>
</tr>
<tr>
<td><strong>Mode of translation</strong></td>
<td>inhouse translation service</td>
<td>Lawyer-linguists</td>
<td>In-house translators or external contractors</td>
<td>Dedicated in-house unit</td>
</tr>
<tr>
<td>Genre</td>
<td>Number of texts</td>
<td>Tokens</td>
<td>Types</td>
<td></td>
</tr>
<tr>
<td>------------------------</td>
<td>-----------------</td>
<td>----------------</td>
<td>----------</td>
<td></td>
</tr>
<tr>
<td><strong>EU PL Legislation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regulations</td>
<td>925</td>
<td>1,899,403</td>
<td>39,409</td>
<td></td>
</tr>
<tr>
<td>Directives</td>
<td>92</td>
<td>768,187</td>
<td>24,288</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1,017</strong></td>
<td><strong>2,667,590</strong></td>
<td><strong>45,249</strong></td>
<td></td>
</tr>
<tr>
<td><strong>PL PL Legislation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statutes</td>
<td>144</td>
<td>1,859,880</td>
<td>35,142</td>
<td></td>
</tr>
<tr>
<td><strong>EU PL Judgments</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CJEU judgments</td>
<td>897</td>
<td>4,693,968</td>
<td>72,310</td>
<td></td>
</tr>
<tr>
<td>General Court judgments</td>
<td>384</td>
<td>4,515,866</td>
<td>65,318</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1,281</strong></td>
<td><strong>9,209,834</strong></td>
<td><strong>97,335</strong></td>
<td></td>
</tr>
<tr>
<td><strong>PL PL Judgments</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supreme Court judgments</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>EU PL Reports</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reports of EU institutions</td>
<td>498</td>
<td>2,329,498</td>
<td>60,466</td>
<td></td>
</tr>
<tr>
<td><strong>PL PL Reports</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reports of PL institutions</td>
<td>110</td>
<td>2,556,114</td>
<td>77,417</td>
<td></td>
</tr>
<tr>
<td><strong>EU PL websites</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Websites of EU institutions</td>
<td>873</td>
<td>679,631</td>
<td>873</td>
<td></td>
</tr>
</tbody>
</table>
Operationalisation for comparable corpora

- **Global variables** (e.g. Biber and Conrad’s Multi-Dimensional (MD) register analysis: part-of-speech classes, semantic categories for word classes, grammatical features)
- **Keyness analysis**: Genre-specific functional variables

- **Differences**: overrepresented, underrepresented (→ unique items), atypical patterns
- **Similarities**: similar distribution of patterns in translations and non-translations
- **N-grams**
- **Feature aggregation**, multi-variant analysis

**Software**: Wordsmith Tools, SketchEngine, LEM
Functional categories for legislation

1. Mental models of legal reasoning (if-than, purpose clauses, cause-effect, interference)
2. Deontic modality
3. Impersonality
4. Logical relations between discourse units: parataxis and hypotaxis
5. Frames, qualifications
6. Deixis & textual mapping
7. Term variation
Modal verbs: deontic and epistemic uses

<table>
<thead>
<tr>
<th></th>
<th>EU Reg</th>
<th>EU Dir</th>
<th>PL</th>
<th>CJEU judg</th>
<th>GC judg</th>
<th>PL</th>
<th>EU report</th>
<th>PL</th>
<th>EU www</th>
<th>PL</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>musi, muszą</em> [must]</td>
<td>644</td>
<td>1026</td>
<td>69</td>
<td>506</td>
<td>512</td>
<td>322</td>
<td>580</td>
<td>109</td>
<td>1209</td>
<td>820</td>
</tr>
<tr>
<td><em>należy</em> [(it) should]</td>
<td>379</td>
<td>375</td>
<td>526</td>
<td>4192</td>
<td>4611</td>
<td>1150</td>
<td>1184</td>
<td>803</td>
<td>797</td>
<td>1052</td>
</tr>
</tbody>
</table>

*Normalised frequency per 1 mln word
# Conditionals

*Normalised frequency per 1 mln word*
Framing through complex prepositions

\[w \text{ odniesieniu do} \] [in relation to]

<table>
<thead>
<tr>
<th>EU Reg</th>
<th>EU Dir</th>
<th>PL</th>
<th>CJEU judg</th>
<th>GC judg</th>
<th>PL</th>
<th>EU report</th>
<th>PL</th>
<th>EU www</th>
<th>PL</th>
</tr>
</thead>
<tbody>
<tr>
<td>2493</td>
<td>1721</td>
<td>130</td>
<td>1156</td>
<td>1402</td>
<td>336</td>
<td>1673</td>
<td>286</td>
<td>442</td>
<td>130</td>
</tr>
</tbody>
</table>

*Normalised frequency per 1 mln word*
Analysis

• Shared features: EU-specific references, increased variation of terminology; increased passivisation and subordination, preference for framing with some complex prepositions

• Translation features/effects strongly correlated with genre and its communicative purpose, esp. in the hybrid multilingual environment
  • 3-grams: overlaps between some genres: navigation bundles in legislation and judgments (*na podstawie, w rozumieniu art.*), explanatory bundles in reports (*ze względu na, w odniesieniu do, w związku z*)

• The modular design of comparable corpora allows us to single out and research certain variables related to **textual fit**, such as the impact of source language, genre, institutionalisation of translation process, diachronic shifts in translation quality
Parallel corpus and qualitative analysis

- Strongly overrepresented and underrepresented patterns verified in corresponding parallel corpora to understand the causes and processes
- Sketchengine
- Unsophisticated techniques for searching parallel corpora, limited potential to identify equivalents in the EN-PL lg pair – manual and qualitative analysis of extracted aligned concordance lines
- Study of legal equivalence requires qualitative analysis
Constraints and limitations

- Corpus-design issues: balance, representativeness and comparability of translational corpora; sampling frames
- Limited potential to research equivalence
- Limited ability to study other dimensions of translation: the context of production and reception, the process and translators; strong focus on the product
- Triangulation of methods needed: quantitative and qualitative analysis
Thank you for your attention

l.biel@uw.edu.pl