A scoping review of the literature on Clinical Natural Language Processing for Languages other than English
Why address a variety of languages?

- **Access to a larger demographic**
  - Access to more patient cohorts
  - Aggregate data for rare and other diseases, e.g. autism spectrum disorder in 4 healthcare centers [Kohane et al. 2012]

- **Apply WHO protocols widely**
  - Success story in the making: IRIS, a software for automated coding of causes of death
  - Collaboration involving France, Hungary, Japan, Germany, Italy, Sweden, United States
Literature on Clinical NLP is hard to find!

- **International Medical Informatics Association (IMIA) Yearbook**
  - Clinical NLP section started in 2014
  - Survey paper, synopsis with « best papers » selection
  - For year 2017, 709 articles reviewed
    - ACL anthology: BioNLP, *ACL conferences (34% off topic)
    - Pubmed: natural language processing (35% off topic)
    - Pubmed: text mining (60% off topic)
    - Overall, 31 (4.3%) addressed a language other than English

- **Reviewing tools used**
  - Bibreview https://pypi.org/project/BibReview/
  - Integrated classifier [Norman et al. P47 Friday@9:45]
Literature on Clinical NLP is hard to find!

• American Medical Informatics Association (AMIA)
  • Panels on clinical NLP for languages other than English in 2014, 2017.

Clinical Natural Language Processing in languages other than English: opportunities and challenges
Growth of bio-clinical NLP publications in MEDLINE for the top 5 studied languages other than English

(22 languages covered in review)
Biomedical NLP in a language other than English

- **What does it consist in?**
  - Data creation: vocabularies, annotated dataset
  - Method development: NLP methods for the biomedical domain, bioNLP tasks
  - Applications

- **Is it different from bioNLP in English?**
  - Less resources
  - Language, country specificities
  - Multilingual aspects: translation, language adaptation, cross-culture comparisons
Building new systems and resources
Domain-specific NLP components

- **Morphological analyzer**
  - DeRIF, for French [Namer & Zweigenbaum 2004]

- **PoS tagger**
  - Experiments conducted for Portuguese [Oleynik et al. 2010], Polish [Marciniak and Mykowiecka 2011], Spanish [Costumero et al. 2014]

- **Parser**
  - Some work for French [Baud et al. 1999] and Finish [Haverinen et al.], but no public tool

- **Entity and concept recognition**
  - No equivalent of Metamap or cTAKES
  - Some tools for direct lexical matching, e.g. BioPortal [Jonquet et al.]
Automatic word segmentation and applications

Named Entity Recognition in Chinese [Lei et al. JAMIA 2014], [Xu et al. JAMIA 2014]
  – Word segmentation (vs. character) performs better
  – Joint segmentation+NER yields 1-% improvement for both
  – F-measure of 90+% for 4 entity types: performance comparable to English, with specific features

  – Addresses the lack of spacing
  – A probabilistic model of word segmentation using dictionaries
  – Successfully applied to English – can be useful for OCR
Transliteration issues

• Expansion of English Abbreviations in Japanese
  [Shinohara et al. Methods Inf Med 2013]
  – code-switching
  – Pilot study on 8 short forms associated to 2 or more long forms
  – Character segmentation (vs. word) performs better

• Word segmentation in Hebrew [Cohen et al. Methods Inf Med 2010]
  – Identification of transliterated words improves medical term extraction by 29%
Lexicons and terminology development

- **Term translation**
  - Using automatic translation systems [van Mulligen et al. 2016]

- **Mapping of terminologies to the UMLS**
  - French consumer vocabulary [Tapi Nzali et al. 2017]
### Corpora and annotations (for Romance Languages)

#### Monolingual Corpora

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Text type</th>
<th>Annotations</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARR</td>
<td>Literature, EHR</td>
<td>Abbreviations, Entities (ADR)</td>
<td>Open</td>
</tr>
<tr>
<td>Oronoz et al.</td>
<td></td>
<td>Negation</td>
<td>Restricted</td>
</tr>
<tr>
<td>IULA</td>
<td></td>
<td></td>
<td>Open</td>
</tr>
<tr>
<td>French</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CépiDC</td>
<td>Death certificate, EHR</td>
<td>ICD10 Concepts, E+R+M, PoS,</td>
<td>under DUA,</td>
</tr>
<tr>
<td>QUAERO</td>
<td>Literature, EHR</td>
<td>Sentiment</td>
<td>Open</td>
</tr>
<tr>
<td>MERLOT</td>
<td>Drug inserts</td>
<td></td>
<td>Restricted</td>
</tr>
<tr>
<td>Sequoia</td>
<td>Social Media</td>
<td></td>
<td>Open</td>
</tr>
<tr>
<td>Tapi Nzali et al.</td>
<td></td>
<td></td>
<td>Restricted</td>
</tr>
<tr>
<td>Portugese</td>
<td>Aluisio et al.</td>
<td>Patient speech classification</td>
<td>Restricted</td>
</tr>
<tr>
<td>Italian</td>
<td>Attardi et al.</td>
<td>EHR Silver entities</td>
<td>From authors</td>
</tr>
<tr>
<td>Romanian</td>
<td>BioRo</td>
<td>Literature, lecture notes</td>
<td>[Mitrofan P16, Wed 4:35]</td>
</tr>
</tbody>
</table>

#### Parallel Corpora

- Used in the WMT campaigns: Scielo, EDP, UFAL
Adapting NLP architectures developed for English
Rule-based systems

- **Negation**
  - Adaptation of NegEx to French [Chapman et al. 2013], Swedish [Skeppstedt 2011], German [Cotik et al. 2016], Dutch [Afzal et al. 2014] and Spanish [Costumero et al. 2014] [Cotik et al. 2016]

  \[
  \text{Absence of [evidence to suggest acute cardiac process]}
  \]

  \[
  \text{Absence de [ganglions métastasiques]}
  \]

- **De-identification**
  - Adaptation of De-ID to French [Grouin et al. 2009]

- **Temporal analysis**
  - Heideltime adapted to French [Tapi Nzali et al. 2015] and Swedish [Vellupilai et al. 2014]
Temporal relation extraction: Clinical TempEval task, THYME corpus

• Temporal “container” relations [Bethard et al. 2016]
  - Between pairs of events
  - Between events and temporal expressions
  - Between events and document creation time

• Participation in the challenge with a neural system [Tourille et al. 2016]
Temporal relation extraction: English to French Language Adaptation

- Use of the MERLOT French clinical corpus, with entity and relation annotations [Campillos et al. 2017]

- Need to convert temporal relations
  - From TimeML to “container” relations
  - Also include related relations, e.g. “reveals”

\[
\text{Au cours de son \textit{séjour}, ce patient a été \textit{tranfusé}.}
\]

- - - - PROC - - - - PROC
- - - - OVER. - - - - OVER.
Experiments: English to French Language Adaptation

Ce *traitement* par *Forlax* et *Motilium* a été *prescrit* pour *15 jours*.

- EVENT
- AFTER
- EVENT
- AFTER
- EVENT
- AFTER
- EVENT
- OVER.
- TIMEX3
- CONTAINER
- CONTAINER

- Generic framework with language dependent features
Results: Temporal relation extraction in English and French

1. DCT relations

<table>
<thead>
<tr>
<th></th>
<th>French (MERLOT)</th>
<th>English (THYME)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>baseline</td>
<td>0.67</td>
<td>0.67</td>
</tr>
<tr>
<td>bef./over.</td>
<td>0.68</td>
<td>0.69</td>
</tr>
<tr>
<td>before</td>
<td>0.81</td>
<td>0.60</td>
</tr>
<tr>
<td>after</td>
<td>0.79</td>
<td>0.69</td>
</tr>
<tr>
<td>overlap</td>
<td>0.88</td>
<td>0.92</td>
</tr>
<tr>
<td>micro-average</td>
<td>0.83</td>
<td>0.84</td>
</tr>
</tbody>
</table>

2. Contains relations

<table>
<thead>
<tr>
<th></th>
<th>French (MERLOT)</th>
<th>English (THYME)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>baseline</td>
<td>0.43</td>
<td>0.15</td>
</tr>
<tr>
<td>no-relation</td>
<td>0.99</td>
<td>1.00</td>
</tr>
<tr>
<td>contains</td>
<td>0.75</td>
<td>0.57</td>
</tr>
<tr>
<td>micro-average</td>
<td>0.98</td>
<td>0.98</td>
</tr>
</tbody>
</table>

- In spite of lower resources for French, performance is similar.
- Performance is comparable to inter-annotator agreement.

[Tourille et al. EACL 2017]
Applications
Biomedical NLP tasks addressed

- **Text classification**
  - Healthcare associated infections in Swedish patient records [Jacobson and Dalianis 2016]
  - Multiple Myeloma in German records [Löpprich et al. 2016]

- **Information extraction used for computing clinical scores**
  - Cardiovascular score (French) [Grouin et al. 2012]
  - Memory scores (Japanese) [Takano et al. 2017]
Multilingual Corpora

Improve access to medical information
- Off-the-shelf automatic translation, e.g. Google translate, Babelfish [Zeng-Treitler et al. 2010] [Wu et al. 2011]
- Medical Speech translation [Bouillon et al. 2007]

Crosslingual Information Retrieval by query translation
- French, knowledge-based [Thirion et al. 2010]
- French/Czech/German, MT based [Pecina et al. 2014]

Study of clinical cultural differences
- Breast cancer information in Germany vs. UK [Weissenberger et al. 2004]
- Clinical records [Wu et al. 2013] and doctor reviews [Hao et al. 2017] in China vs. US
Challenges and opportunities
How can we advance clinical NLP in languages other than English?

- **Mapping NLP efforts:**
  - Track progress through literature review
- **Resource development**
  - Terminologies: increase coverage of the UMLS
  - Annotated corpus
- **Clinical Shared Tasks**
  - CLEF 2018: ICD10 coding for French, Hungarian, and Italian
  - BARR 2017: abbreviation resolution in Spanish
- **Support the creation of modular, multilingual NLP suites**
Thank you!

CABeRneT ANR-13-JS02-0009-01

EU H2020 Marie Sklodowska-Curie grant No 676207