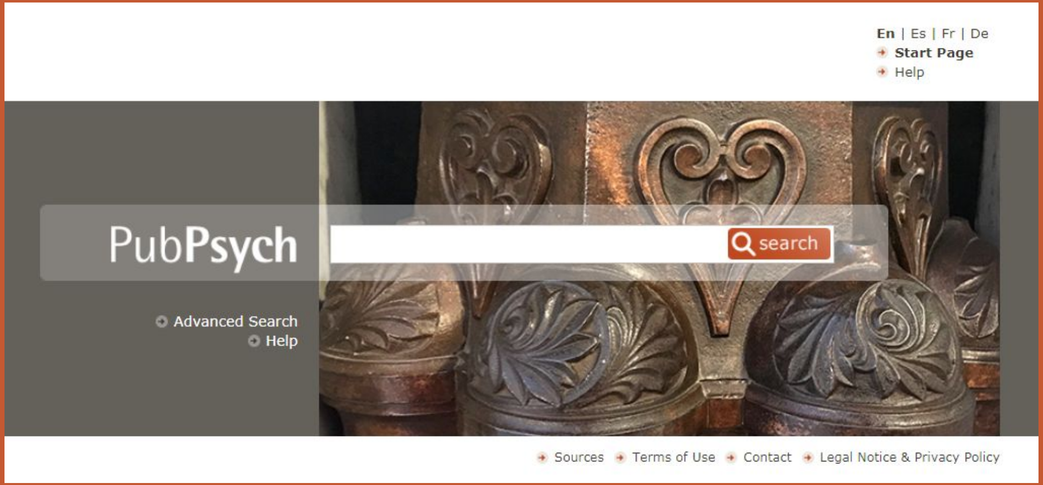
The CLuBS Use Case: PubPsych



Andreas Lüschow, Roland Ramthun

CLuBS Final Project Workshop
June 7th, 2019, DFKI, Saarbrücken

leibniz-psychology.org

Frontend changes

- Current productive system shows only original titles at top level
- Human translated titles only in detailed record view
- Abstracts are already shown in the interface language, if available

- We implemented multilinguality for title display
  - Original title is always shown
  - Title translations are shown in the user‘s interface language
  - Machine translations are marked as such

Frontend changes

- Same mechanism could be used for abstract translation
  - Not implemented yet
  - Important aspect: adequacy vs. fluency
  - Might make full document view more confusing
  - Possible solution: hide translations/show additional information only on user request

Backend

- Currently: only human generated translations available in the index
- We added machine translations in appropriate fields (cf. Cristinas talk earlier this day)
    - e.g.
        - TI // TI_E // TI_D_from_E // TI_F_from_E // TI_S_from_E
        - AB // ABE // ABHR_E // ABHR_D_from_E // ABHR_F_from_E // ABHR_S_from_E
        - CT // CTE // CTEL // CTDL_from_E // CTFL_from_E // CTSL_from_E
- Second approach: online query translation

Backend - Online query translation

- Search query is analyzed
- Dictionaries help with translation
- Appropriate fields are then used with translations (e.g. TI_D_from_E)
- No effect on frontend display, only on result list generation
- Translation process in recorded in logfiles, statistics are also saved for further analysis

Logging

```
[...]
[…] QueryFieldRewriter Translating string "music" , which is the value of field text
[…] QueryFieldRewriter The whole string is contained in the MeSh dictionary.
[...] QueryFieldRewriter translateWholeString: "music" counts towards the statistics.
[...]
[...] QueryFieldRewriter translate: "processing" counts towards the statistics.
[...] QueryFieldRewriter Translating string "processing" , which is the value of field text
[...] QueryFieldRewriter The whole string is contained in the mixed dictionary.
[...] QueryFieldRewriter translateWholeString: "processing" counts towards the statistics.
```

Dictionaries

MeSH

```
97662: music|||de:musik|||es:musica|||fr:musique
202708: musik|||en:music|||es:musica|||fr:musique
274102: musique|||de:musik|||en:music|||es:musica
342875: musica|||de:musik|||en:music|||fr:musique
342876: musico|||de:musik|||en:music|||fr:musique
```

Mixed dictionary

```
62749:processing|||de:processing|||es:processing|||fr:processing
```

Final query (in background)

```
+(+(text:music | (SW:music)^2.0 | (AU:music)^1.1 | (TI:music)^2.0 | text:musik | text:musique | text:musica
| (SW:musik)^2.0 | (SW:musique)^2.0 | (SW:musica)^2.0 | (TI_D:musik)^2.0 | (TI_D_from_E:musik)^2.0 |
(TI_D_from_F:musik)^2.0 | (TI_D_from_S:musik)^2.0 | (TI_F:musique)^2.0 | (TI_F_from_D:musique)^2.0 |
(TI_F_from_E:musique)^2.0 | (TI_F_from_S:musique)^2.0 | (TI_S:musica)^2.0 |
(TI_S_from_D:musica)^2.0 | (TI_S_from_E:musica)^2.0 | (TI_S_from_F:musica)^2.0) +(text:processing |
(SW:processing)^2.0 | (AU:processing)^1.1 | (TI:processing)^2.0 | text:processing | text:processing |
text:processing | (SW:processing)^2.0 | (SW:processing)^2.0 | (SW:processing)^2.0 |
(TI_D:processing)^2.0 | (TI_D_from_E:processing)^2.0 | (TI_D_from_F:processing)^2.0 |
(TI_D_from_S:processing)^2.0 | (TI_F:processing)^2.0 | (TI_F_from_D:processing)^2.0 |
(TI_F_from_E:processing)^2.0 | (TI_F_from_S:processing)^2.0 | (TI_S:processing)^2.0 |
(TI_S_from_D:processing)^2.0 | (TI_S_from_E:processing)^2.0 | (TI_S_from_F:processing)^2.0))
```

Statistics

Mesh usage word level: 14
Mesh usage multi-token level: 2
Mesh usage query level: 10
Backoff usage word level: 6
Backoff usage multi-token level: 2
Backoff usage query level: 2
Number of copies at query level: 1
Number of copies at multi-token level: 1
Number of copies at word level: 2
Singular usage word level: 3
Singular usage multi-token level: 3
Singular usage query level: 0

Final system

- Final backend implementation depending on evaluation results
    - online query translation vs. content translation
- Frontend changes necessary to reflect cross-lingual retrieval
    - show (machine) translations, show query expansion
    - depending on user/interface language
    - could be personalized (e.g., a user speaks multiple languages)
- Frontend changes should be evaluated using A/B testing
    - not the focus of the project, but important aspect!
    - ask users for feedback on translations?

Conclusion

Project outcomes

- Translation Pipeline
  - GitHub: https://github.com/clubs-project/DBtranslator
- Machine Translation Models
- Retrieval Assessment Tool
  - can be used for other purposes as well, e.g., classification of documents
  - will have a module for A/B tests of websites/designs
  - published open source when finished (Q3 / 2019)

Project outcomes

- Manually translated parallel data (abstracts, titles, and queries)
  - 800 abstracts, 7,195 sentences, 145,538 words
  - in 4 languages; GER, FRE, and SPA each translated twice and double-checked
  - 261 queries in 4 language
- Publications and project documentation
  - see project website: https://www.clubs-project.eu/
- Improved PubPsych search engine

Thank you for your attention

HUMBOLDT-UNIVERSITÄT · ZU BERLIN ·

leibniz-psychology.org

UNIVERSITÄT
DES
SAARLANDES