Lo Congrès
permanent de la lenga occitana

AitalvivemBot, a tool to pour (occitan) lexemes in Wikidata

Aure Séguier, NLP project manager at Lo Congrès

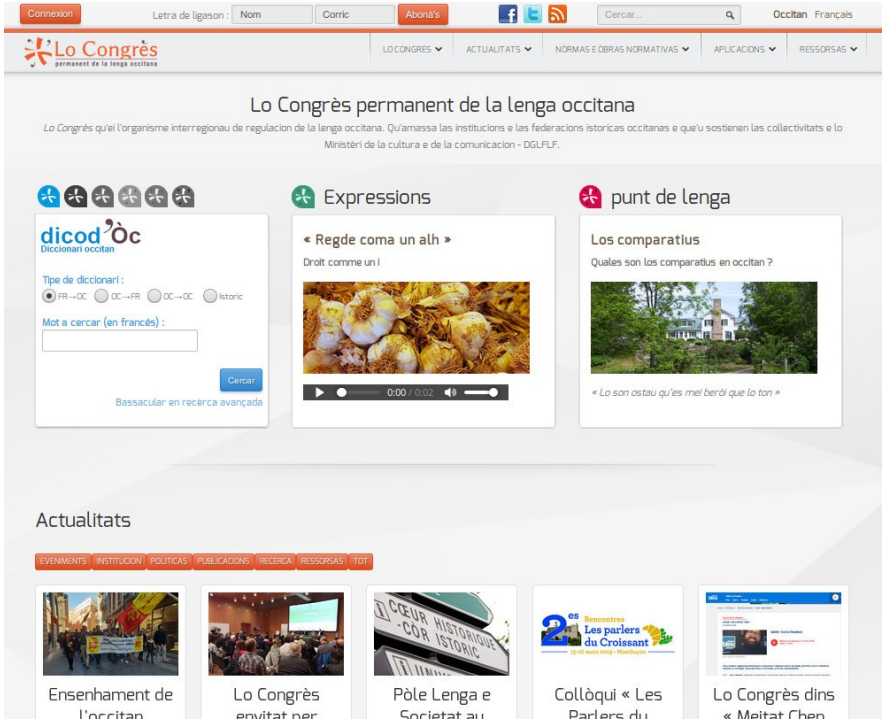
Vincent Gleizes, computer science student

Occitan, a second language in south of France

Lo Congrès permanent de la lenga occitana

Lo Congrès permanent de la lenga occitana

- Interregional organization for the regulation of occitan language
- Production of referentials linguistical tools
- Regulation using numeric tools :
  - Online dictionaries
  - Verb conjugators
  - Predictive keyboard
  - Spell checker
  - Automatic translation…



Lo Congrès permanent de la lenga occitana

Pouring occitan lexemes in Wikidata

- Lo Congrès has a lot of lexicographical datas from digital dictionaries, verbs conjugators…
- Occitan has'nt enough money to build a private knowledge base (and why doing this?)
- Linking occitan lexemes to concept will permit to build NLP tools in occitan using semantical analysis :
  - Question answering
  - Text production
  - Documents classification
  - Personal assistants…

Lo Congrès
permanent de la lenga occitana

First step studies and ground question

- Finding the correspondances between the datas organization model used by Wikidata and Lo Congrès
- Finding characteristics to determine if 2 lexemes are identical
- Studying the API json request and answers
- Listing the functions the script will use


```
{
    "searchinfo": {
        "search": "ostal"
    },
    "search": [
        {
            "repository": "",
            "id": "L12",
            "concepturi": "http://test.wikidata.org/entity/L12",
            "title": "Lexeme:L12",
            "pageid": 233953,
            "url": "//test.wikidata.org/wiki/Lexeme:L12",
            "label": "test",
            "description": "Unknown language, Unknown",
            "match": {
                "type": "alias",
                "language": "und",
                "text": "ostal"
            },
            "aliases": [
```


JSON


```
{
    "action": "wbsearchentities",
    "format": "json",
    "search": "ostal",
    "language": "oc",
    "type": "lexeme"
}
```


Copy

Lo Congrès permanent de la lenga occitana

Writing and testing

- Writing and testing each functions separately
- Writing and testing the main algorithm
- Real test run in wikidata
- Correction following the advises and feebacks given by the Wikidata community

```
def chercheLex(info, lg, catLex, declaLex):

        """
        This function search for a lexeme in the wikidata database
                If the function find the lexeme it return the id
                else the function call the createLex function
        """

        S = requests.Session()
        URL = "https://www.wikidata.org/w/api.php"

        PARAMS = {
                'action':'wbsearchentities',
                'language' : lg['libLg'],
                'type' : 'lexeme',
                'search': info,
                'format':'json'
        }

        R = S.get(url=URL, params=PARAMS)
        DATA = R.json()

        lexExists = False

        try:
                for item in DATA['search']:
                        # if the lexeme exists
                        if item['match']['text'] == info and item['match']['language'] == lg['libLg']:
                                # I check the grammaticals features
                                lexQid = getQidCat(item['id'])
                                if lexQid==1:
                                        return 0, 3

                                if catLex == lexQid:
```

Difficulties

- How to deal with a category used by Lo Congres if it has no match in Wikidata ?
  → finding another category which includes the one we were looking for

- Impossible to create a lexeme and its forms sequentially
  → running the program 2 times, the first one in creating only the lexemes

- How to deal with dialects ?
  → asking for the creation of a new property, in the meantime, use a geogaphical property (P276)

- How to indicate if a lexeme without plural or without feminine is complete ?
  → using existing items (ex. : plurale tantum - Q138246)

- How to make the bot reusable for other langages ?
  → using a simple and international XML format which works with every property-item pair

Lo Congrès
permanent de la lenga occitana

The result

https://github.com/aitalvivem/AitalvivemBot

- A Bot reading an XML file with lexemes, forms and their properties
- A Bot checking if a lexeme exists and creating a new one if not
- A Bot checking if a form exists for a lexeme and creating it if not
- A Bot adding pairs of properties – items to lexemes and forms

Lo Congrès permanent de la lenga occitana

Lo Congrès
permanent de la lenga occitana

Mercés plan de vòstra atencion