Pluridisciplinary aspects of NLP and GIS

An Application to Itinerary Reconstruction

Ludovic Moncla

Naval Academy Research Institute, IRENav

ludovic.moncla@ecole-navale.fr

Overview

One of the main challenge of my work is to **connect text with geographic space** and to provide a map-based representation of itineraries described in textual documents. The main objectives are:

- data mining for **Geographic Information Retrieval** (GIR),
- toponym resolution and disambiguation,
- extract and retrieve displacement from **textual documents**.

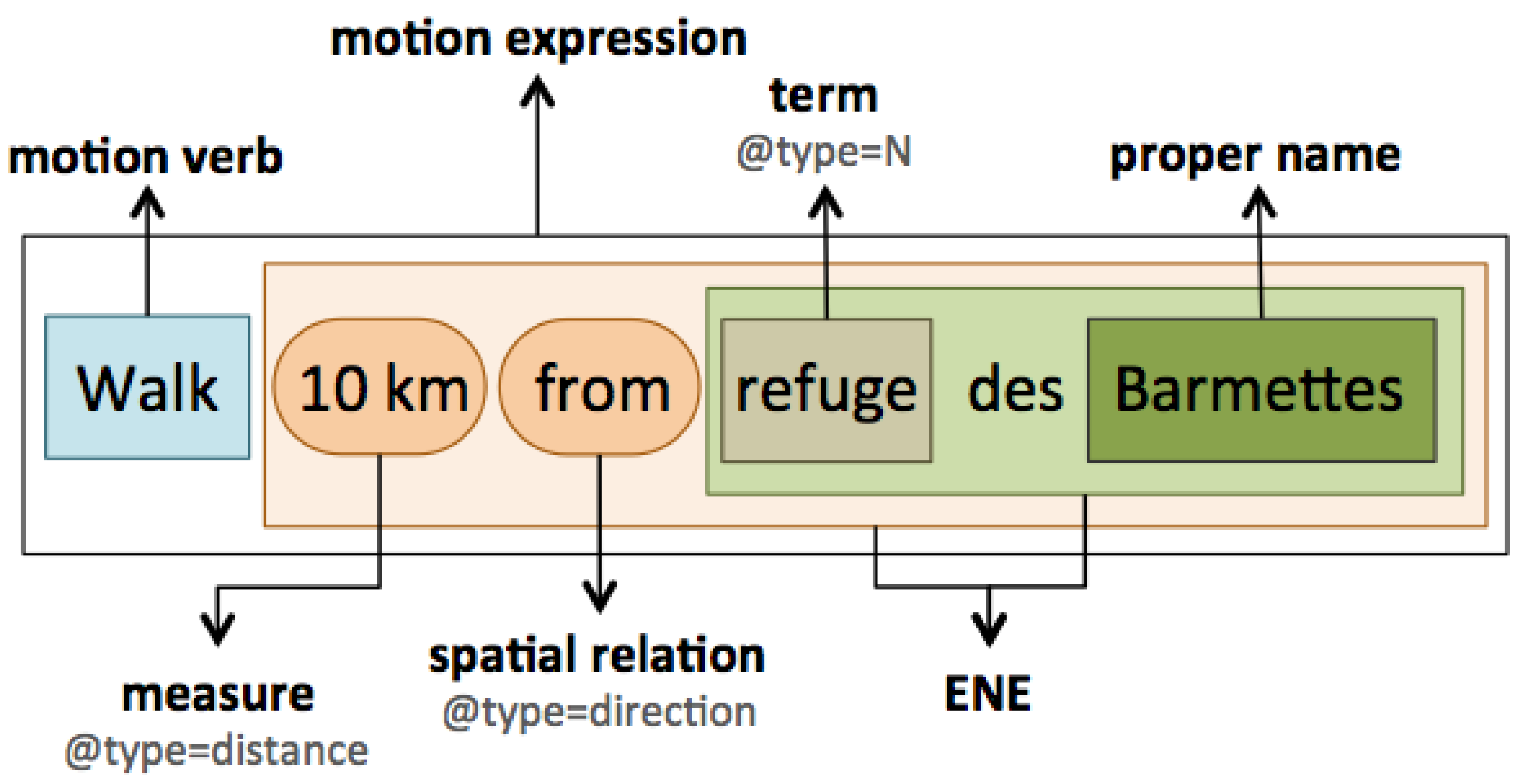
1. Geoparsing Places in a Dynamic Space Context

1.1 Extended Named Entity

- Construction grammars adapted for French, Spanish and Italian.
- Implemented with a cascade of finite-state transducers (Unitex).

$$\text{ENEA}\begin{bmatrix} \text{level} & 1 \\ \text{type} & \textit{location} \\ \text{cat.} & \textit{descriptive} \\ \text{comp.} & \text{NN,IN,} \\ \text{ENEA} & \begin{bmatrix} \text{level} & 0 \\ \text{type} & \textit{location} \\ \text{cat.} & \textit{pure} \\ \text{comp.} & \text{NNP} \\ \text{lex.} & \textit{Montréal} \end{bmatrix} \\ \text{lex.} & \textit{île de Montréal} \end{bmatrix}$$

1.2 Motions Expressions

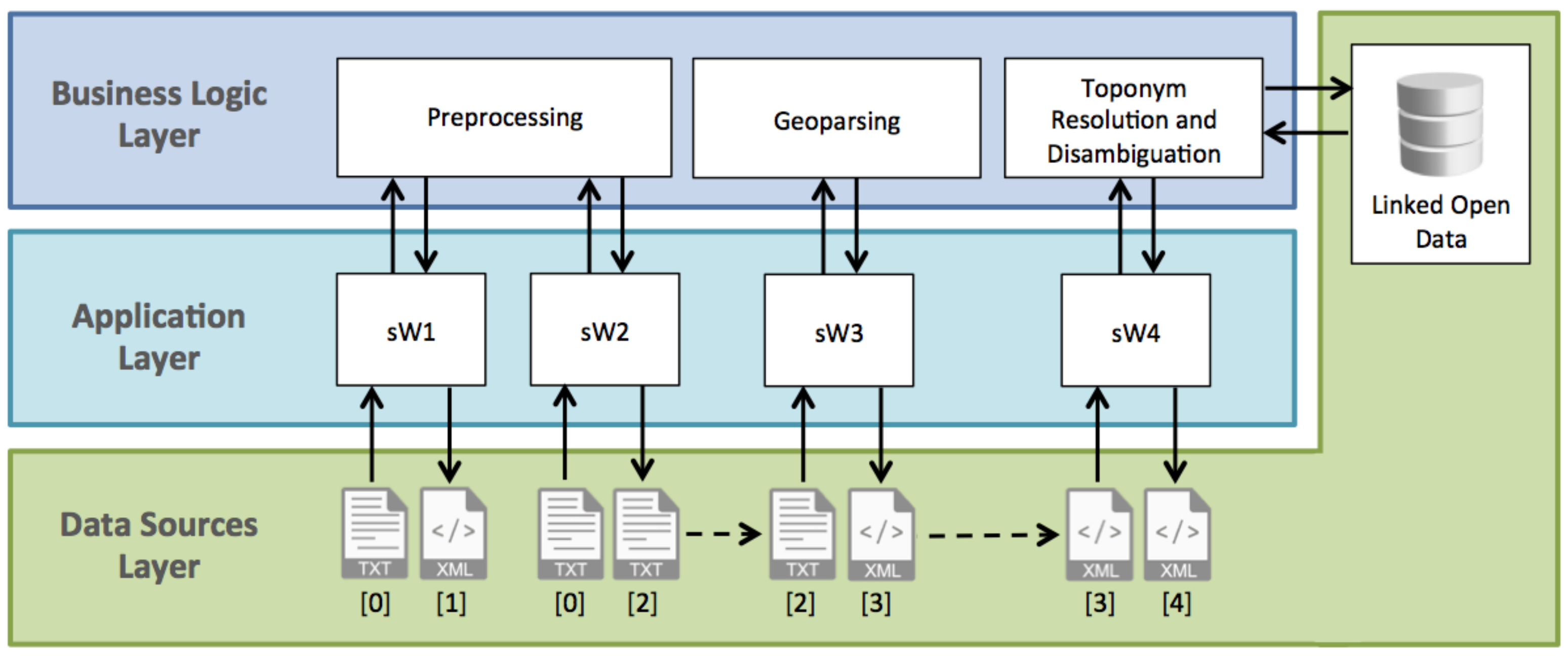


1.3 Extended Named Entity

- XML-TEI output format following the standard guidelines for encoding of texts in digital form
- Feature types from the ontologies

```
<placeName>
  <geogName type="R" subtype="ST">
    <geogFeat>
      <w lemma="rue" type="N">rue</w>
    </geogFeat>
    <w lemma="de" type="PREP">de</w>
    <name>
      <w lemma="Rivoli" type="NPr">Rivoli</w>
    </name>
  </geogName>
</placeName>
```

1.4 Web Services



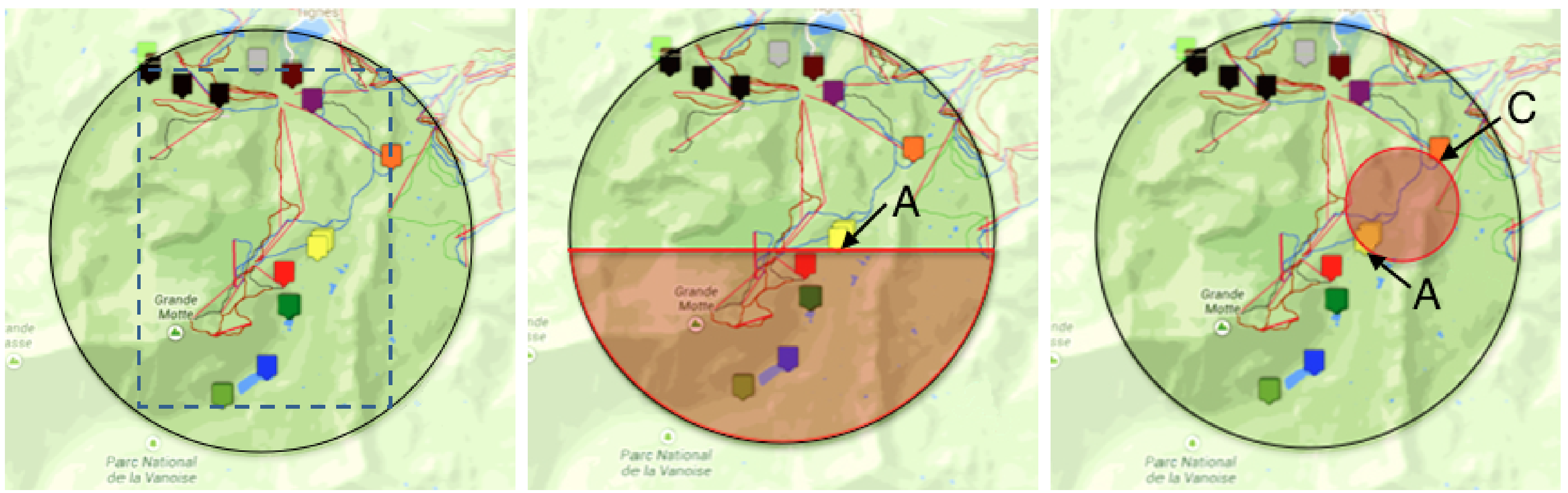
2. Toponym Resolution and Disambiguation

2.1 Geographic Data

- Official national geographic databases.
- Geographic gazetteers from Linked Open Data (GeoNames, OpenStreetMap).

2.2 Toponym Disambiguation

1. Subtyping of place named entities:
   - querying metadata from gazetteers to match feature types.
2. Density-based spatial clustering (DBSCAN).
3. Geocoding for unreferenced ambiguity:
   - automatic gazetteers and data enrichment.



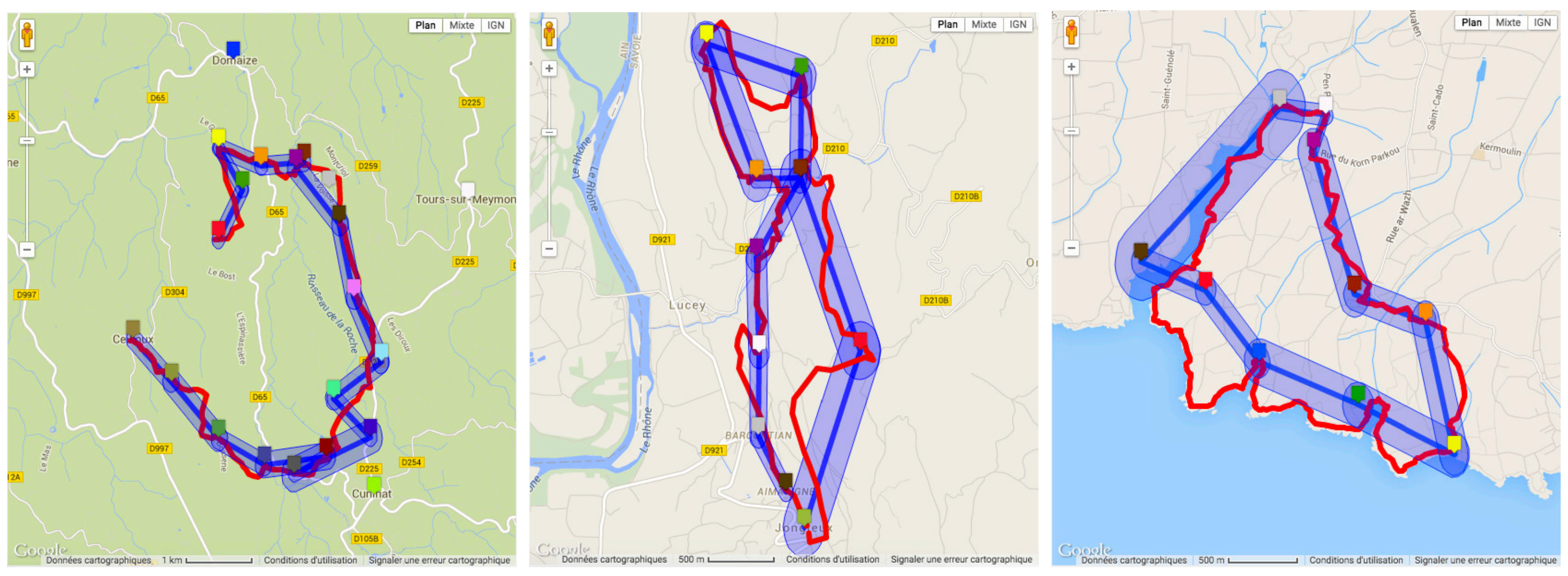
3. Use Case and Results

3.1 Extended Named Entity Recognition and Classification

- 90 hiking descriptions
- 82% of ENE are correctly detected,
- 38% of ENE are associated with motion verbs,
- **54%** of ENE are associated with a **feature type** (level > 0),
- Almost **25%** of place names are **not found in geographical databases**.

Toponyms	#	%
manually annotated	1523	100%
automatically annotated	1249	82%
located by gazetteers	719	57%
located by inferences	**402**	**32%**
unlocated	128	10%

3.2 Automatic reconstruction of itineraries



Conclusions

- Automatic **geoparsing** and **geocoding** process combining textual information referring to motion and space with data from external geographical resources.
- **Toponym disambiguation** methods adapted to places in a dynamic space context.
- **Automatic itinerary reconstruction** combining quantitative and qualitative criteria, based on data extracted from the text and data extracted from external geographic databases.

Acknowledgements

I would like to gratefully acknowledge the RDA EU Early Career support Programme for the travel grant to attend this meeting.