Crafting a lexicon of referential expressions for NLG applications

Ariel Gutman
relgu@google.com

Alexandros A. Chaaraoui
alexandrosc@google.com

Pascal Fleury
fleury@google.com

Google Research Europe, Switzerland

Introduction

NLG systems with a wide coverage, such as personal assistants, need access to a lexicon of **referential expressions** (e.g. *Paris, The Beatles* or *James Bond*) in order to produce fluent text. The lexicon should minimally contain:

- Surface form(s) of the expression; in case-inflecting languages several forms are needed.
- Grammatical properties: gender, number, case etc.
- Collocational properties: locative preposition, determiner etc.

Contrary to normal lexemes, referential expressions are **open-ended** (millions of entities).

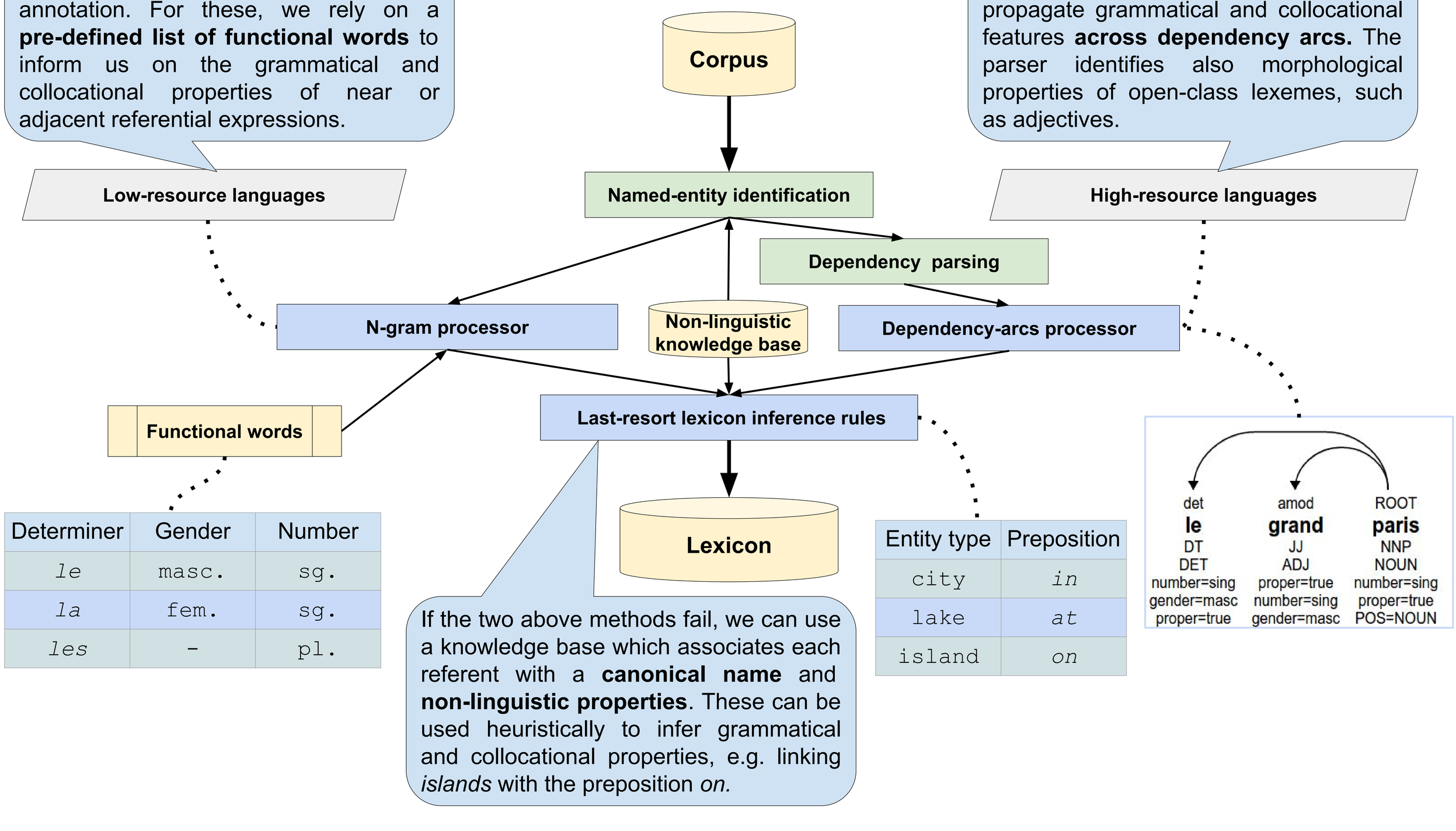
Name	Preposition	Determiner
Zurich	in	-
Lake Como	at	-
Isle of Man	on	the

It is sunny (Location).

in Zurich | at Lake Como | on the Isle of Man

For some languages we only have a corpus without any grammatical annotation. For these, we rely on a **pre-defined list of functional words** to inform us on the grammatical and collocational properties of near or adjacent referential expressions.

For other languages, we have also a dependency parser. In this case we can propagate grammatical and collocational features **across dependency arcs.** The parser identifies also morphological properties of open-class lexemes, such as adjectives.



Determiner	Gender	Number
le	masc.	sg.
la	fem.	sg.
les	-	pl.

Entity type	Preposition
city	*in*
lake	*at*
island	*on*

If the two above methods fail, we can use a knowledge base which associates each referent with a **canonical name** and **non-linguistic properties**. These can be used heuristically to infer grammatical and collocational properties, e.g. linking *islands* with the preposition *on.*

Results

The above techniques yield good precision results.

Preliminary evaluation on some European languages (French, Swedish and Czech) gives the following results:

- **Above 92%** precision for grammatical number (FR, SV)
- **66%-87%** precision for grammatical gender; more difficult since there are less grammatical cues for it (FR, SV)
- **Above 96%** precision for locative preposition (CS only)

Select references

- D. Andor et al. 2016. Globally normalized transition-based neural networks. *CoRR*.
- L. Clément et al. 2004. Morphology based automatic acquisition of large-coverage lexica. *LREC 2004*. pp. 1841–1844.
- A. Gutman et al. 2015. Bootstrapping the syntactic bootstrapper: Probabilistic labeling of prosodic phrases. *Language Acquisition* 22(3), pp. 285–309.
- N. Momchev. 2010. Annotating Web Documents With Wikipedia Entities. Master's thesis, Sofia University.
- B. Sagot. 2010. The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. *LREC 2010.*

Outlook

The NLP literature hardly refers to lexical properties of referential expressions used for NLG, yet these are needed to produce eloquent texts. While conceptually simple, crafting such a lexicon is challenging due to large scale of such expressions, and the varying amount of information available in different languages.

Future work

- Selection among various expressions referring to one entity.
- Reconciling several expressions to a grammatical paradigm.
- Annotation of various types of expressions (*official* vs. *colloquial*, *autonym* vs. *pseudonym* etc.).

We wish to acknowledge the help of Jana Strnadova and Ivan Korotkov.