From built examples to attested examples: a syntax-based query system for non-specialists

Ilaine Wang[1,2] Sylvain Kahane[1] Isabelle Tellier[2]

[1]MoDyCo (UMR 7114) - CNRS
Université Paris Nanterre

[2]LaTTiCe (UMR 8094) - CNRS
ENS Paris, PSL .University, USPC

Abstract

Using queries to explore corpora is today part of the routine of not only researchers of various fields with an empirical approach to discourse, but also of non-specialists who use search engines or concordancers for language learning purposes. If keyword-based queries are quite common, non-specialists still seem to be less likely to explore syntactic constructions. Indeed, syntax-based queries usually require the use of regular expressions with grammatical words combined with morphosyntactic tags, which imply that users master both the query language of the tool and the tagset of the annotated corpus. However, non-specialists like language learners might want to focus on the output rather than spend time and efforts on mastering a query language. To address this shortcoming, we propose a methodology including a POS-tagger or a syntactic parser and using common similarity measures to compare sequences of morphosyntactic tags automatically provided.

For whom?

› University students or self-directed learners who have no (or little) access to the target language, beginner or advanced-level (Boulton 2008)

› Language teachers or anyone involved in making teaching materials

/!\ Complexity of current tools

Requirements:
- *language*-related knowledge: how to identify and characterise a specific construction (grammar)?
- *corpus*-related knowledge: how is it encoded (corpus tagset)?
- *computer*-related knowledge: how to make a query (query language)?

› effort and time-consuming

What needs?

More exposure to authentic data

› Direct confrontation to **native corpora** , displaying actual use of the target language (Kennedy & Miceli 2001, Bernardini 2002, Chambers 2005)

› Observation phase of the target language (Holec 1990)

› **Data-Driven Learning** (Johns 1991) : enhancing active learning, "learner as researcher"

Methodology

Illustration of the processing chain on Korean with -(으)로 *-(u)lo* as the instrumental case particle and -(으)ㄹ지도 모르다 *-(u)lcito moluta* as a construction with an epistemic value

User input ①
Natural language

김치는 배추로 만듭니다.

Kimchi is made of cabbage.

내일은 맑을지도 모릅니다.

I am not sure if the weather will be clear tomorrow.

Selection of relevant word(s) or morpheme(s) ④

김치 는 배추 로 만들 ㅂ니다 .
○ ○ ○ ● ○ ○ ○

내일 은 맑 을지 도 모르 ㅂ니다 .
○ ○ ○ ● ● ● ○ ○

Selection on first output ⑥
Query refining

(1) 나는 다시 경어로 물었다.
I again honorifics.INS asked

(2) 남자는 형에게 묻는다.
Man older brother.DAT ask

(1) 어쩌면 그럴지도 모르겠습니다.
perhaps to.be.like.this.PRSP.whether not know

(2) 오히려 잘 되었는지도 모르겠다.
rather well turn.out.PST.NPRSP.whether not know

② Automatic Syntactic Analysis
Parsing or POS-tagging*

KKMA POS-tagger from KoNLPy

김치/NNG 는/JX 배추/NNG **로/JKB** 만들/VV ㅂ니다/EF ./SF
kimchi TOP cabbage **INS** make AH-IND-DECL .

내일/NNG 은/JX 맑/VA **을지/EC** **도/JX 모르/VV** ㅂ니다/EF ./SF
tomorrow TOP clear **PRS**-whether also not know AH-IND-DECL .

③ Query simplification (optional)
Suppressing non-relevant lexical items to use POS tag only instead*

NNG 는/JX NNG **로/JKB** VV ㅂ니다/EF ./SF
TOP **INS** AH-IND-DECL .

NNG 은/JX VA **을지/EC** **도/JX 모르/VV** ㅂ니다/EF ./SF
TOP **PRP**-whether also not know AH-IND-DECL .

POS-tagged corpus from Sejong Project

⑤ Comparison with tagged corpus
Similarity measures computation: jaccard/dice with unigrams/bigram or edit distance*

	Wordform(s)?	Context?	
(1)	Same	Similar	-> Concordancing + similar context
(2)	Different	Similar	-> Distributional analysis or similar constructions context

⑦ Final output

(1)나는 다시 경어로 물었다.
I asked again in honorific form.
모래로 만든 벽.
A wall (made) of sand.
<희수>의 한자를 <喜囚>로 만들면 어떨까?
How about using <喜囚> for <희수>?

(2)남자는 형에게 묻는다.
The man asked his older brother.
나는 속으로 웃었다.
I laughed up my sleeve (lit. inside).
나는 벤치에서 일어섰다.
I stood up from the bench.

(1)어쩌면 그럴지도 모르겠습니다.
(Perhaps) It might be like this.
혹시 차가 올지도 모르니까요.
Because a car might come by any chance.
그건 사실일지도 모른다.
This might be true.

(2)오히려 잘 되었는지도 모르겠다.
Maybe it is better this way.
"고향이 북쪽인지도 모르지."
"Not sure if my hometown is up in North."
하지만 그것은 역설이 아니었는지도 모른다.
But maybe this was not a paradox.

A work in progress...

Configuration on:
- number of sentences in input?
- efficiency and relevancy of modes?
- use of lexical units?
- similarity measures?
- relevancy of text genres?

Tests using a syntactic treebank

Application on other languages

Qualitative Evaluation + Clustering

*Needs to be configured for each language

PACLIC 2016, Oct 28th-30th