Preliminary Thoughts on Issues of Modeling Japanese Dictionaries Using the OntoLex Model

Louis Lecailliez

RASLAN 2017

December 1st

Self introduction

- Master degree of NLP (2015)
- Mater degree in Japanese studies (2016)
- Bachelor of Informatics; bachelor of Japanese studies

- Here (Masaryk University, FI) with a national scholarship
- PhD next year in Japan (probably)

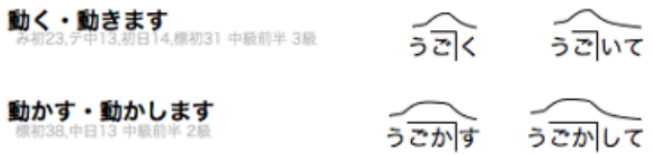
What this talk is about?

- Dictionaries of Japan (there's a lot!)
- Graph dictionaries, more specifically the OntoLex model
- => issues in encoding Japanese dictionaries with OntoLex

Dictionaries of the Chinese cultural sphere

Dictionary type	*Japanese name*	*Also exists*	*Entry example*
Chinese Character Dictionary	漢字辞典	As bilingual dict. e.g. Chinese-French	和、天、車、龍
Chinese Compound Dictionary	漢語辞典	Korea (?) => Unilingual in Chinese	天下、中国人
“Four Character Compound Dictionary”	四字熟語辞典	Chinese world, Korea	七転八倒、兎起鶻落

Dictionaries exclusive to Japan

Dictionary type	*Japanese name*	*Entry example*
Accent Dictionary	アクセント辞典	
Classical Language Dictionary	古語辞典	く、ふううんのおもひ
Dictionary of katakana words	カタカナ語辞典	コーヒー、チェコ、フランス

Graph dictionary landscape

- Dictionary are traditionally (paper, xml files) tree-based
  - Focus on human users or human facing applications
  - “So far, the full potential of computers has not been exploited for the benefit of digital dictionary users because […] the approach used is still that of printed dictionaries with electronic access.” [1]
- Related resources exist, structured as graph: wordnets, ontologies
  - Mostly target NLP applications
- OntoLex: ontology lexicalization
  - “rich linguistic grounding for ontologies” [2]
  - Detourned for implementing dictionaries

Unilingual dictionary

In Czech, French, etc.

Word : Definition

In Japanese:

Word :	Kanji	Definition
せんせい	先生	人を教えたり…
せんせい	宣誓	みんなの前で…

Kanji (Chinese character) dictionary

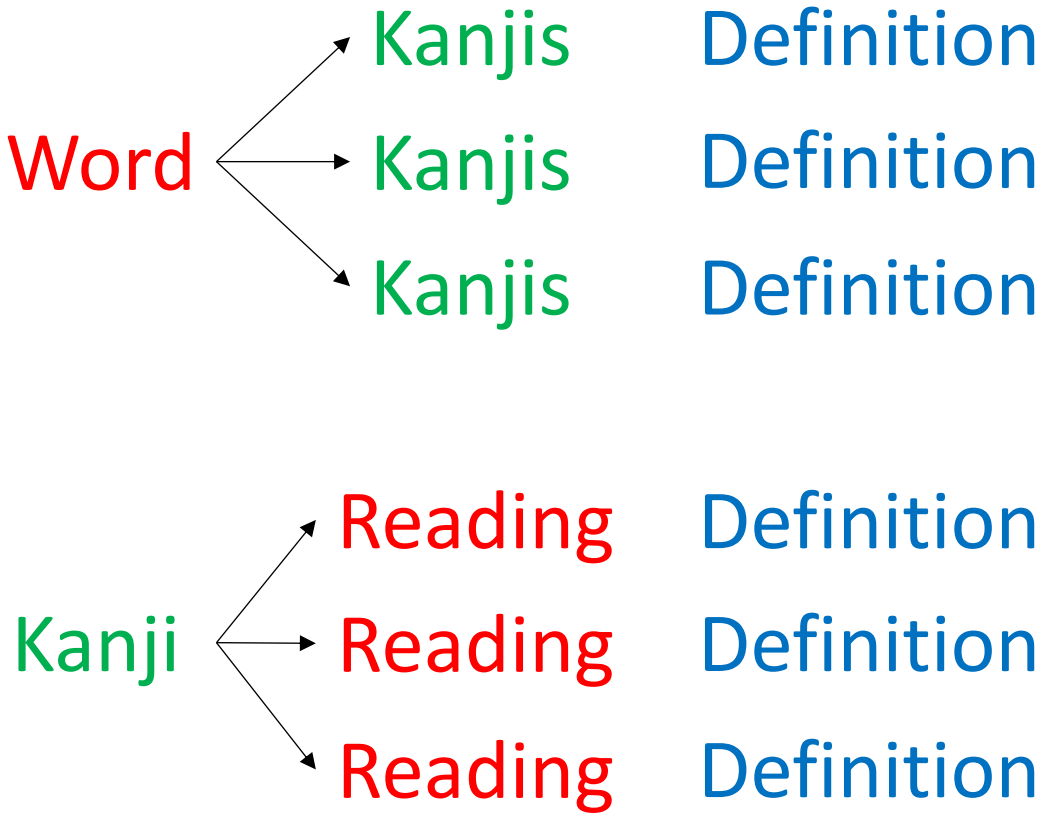
In a kanji dictionary:

Kanji	Reading	Definition
鬱	ウチ	*definition*
鬱	しげる	*definition*

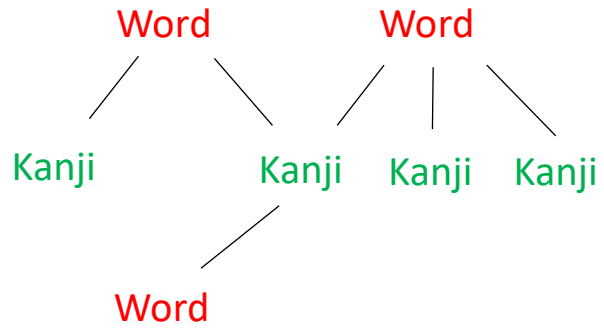
In a unilingual dictionary:

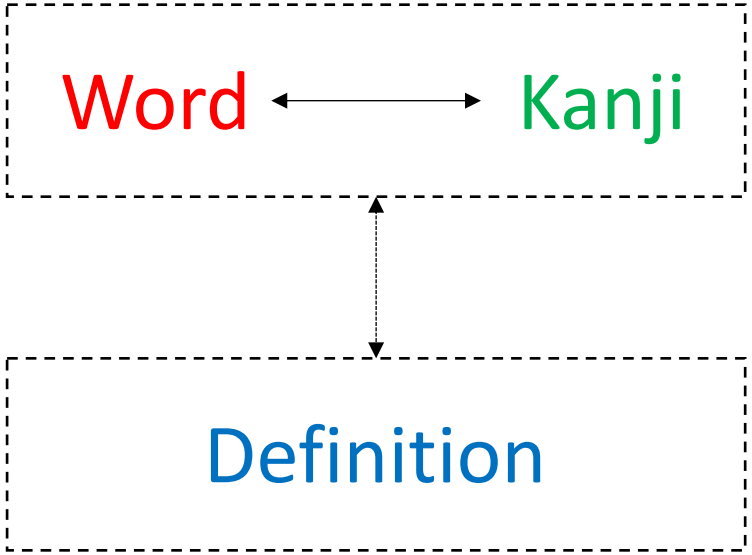
Word :	Kanji	Definition
せんせい	先生	人を教えたり…
せんせい	宣誓	みんなの前で…

As Trees

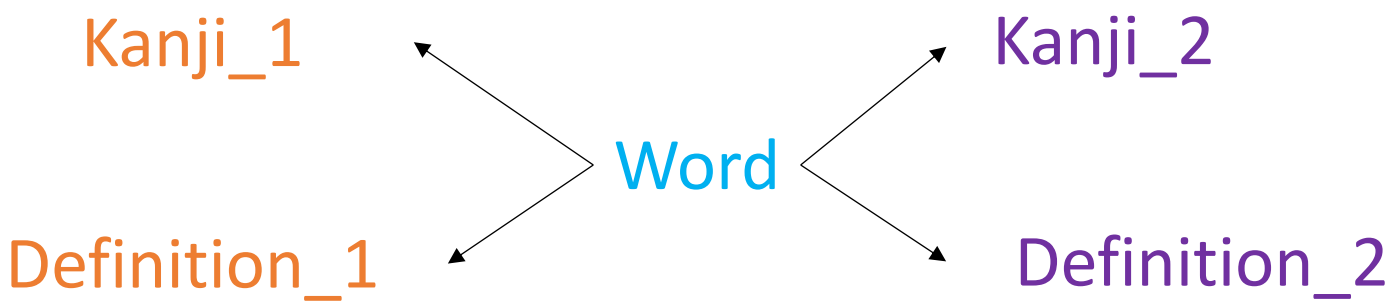


As Graph



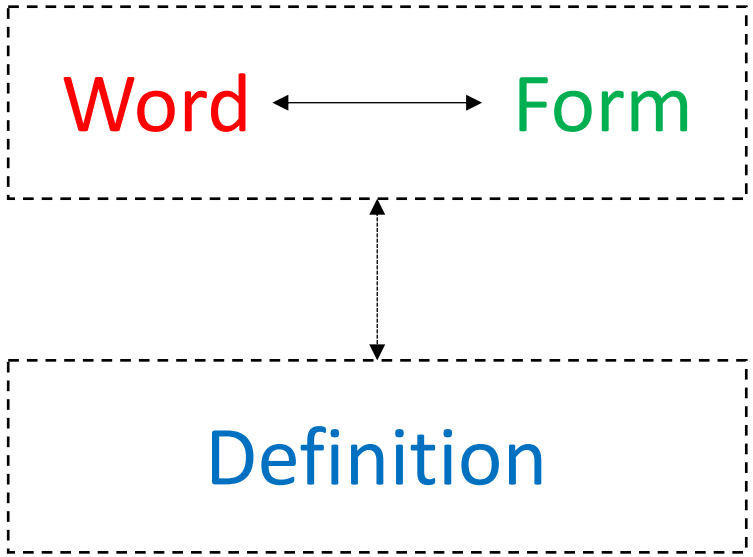


The wrong inferences issue

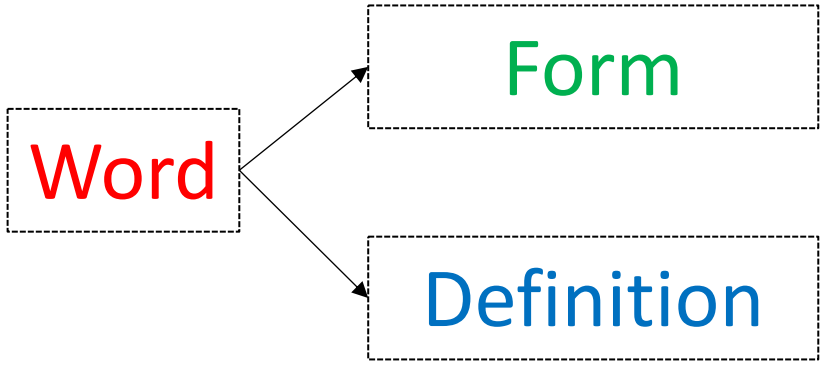


Right Interferences	*Wrong Inferences*
Word, Kanji_1, Definition_1	Word, Kanji_1, Definition_2
Word, Kanji_2, Definition_2	Word, Kanji_2, Definition_1

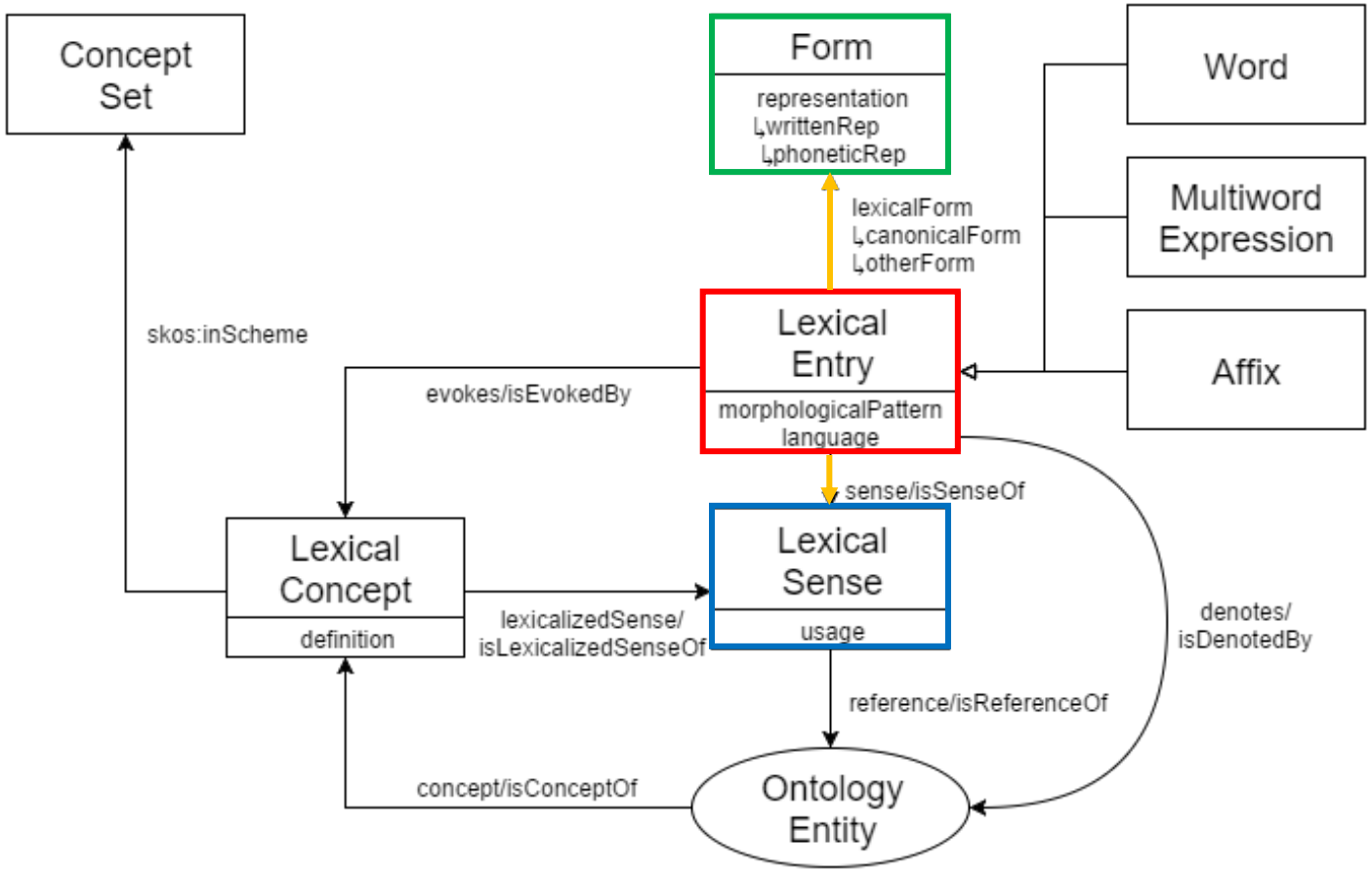
modeling



modeling



OntoLex model



OntoLex also inherits all issues of RDF

- :nihongo lemon:canonicalForm
  [lemon:writtenRep "日本語"@ja-Jpan ;
  isocat:transliteration "にほんご"@ja-Hira ;
  isocat:transliteration "nihongo"@ja-Latn] .

  Example 15 from the Lemon Cookbook

- Blank nodes [ ] are not addressable
- Literals " " cannot be the source of a link
- Node edge annotations

Conclusion

- Chinese character representation must be tackled first
  - On it depend almost every kind of Japanese dictionary
- The N-N relationship between kanji and readings need special modeling
  - Otherwise information may be lost
  - A similar problem arise at word level
  - A good solution should handle both of these problem
- The current OntoLex model needs extension to deal with Japanese
  - The current “Form” entity does not fit the bill as it unrelated to the meaning

Thanks for your attention

References

[1] L'Homme & Cormier. 2014. Dictionaries and the digital revolution: a focus on users and lexical databases. *International Journal of Lexicography*, Vol. 27 No. 4, pp. 331-340. doi:10.1093/ijl/ecu023

[2] W3C. 2017. Final Model Specification. https://www.w3.org/community/ontolex/wiki/Final_Model_Specification

Full references in the proceedings.