UNIVERSITÉ DE STRASBOURG

Corpus et Outils en linguistique, langues et parole: Statuts, Usages et Mésusages

Using corpora to find relevant examples of Japanese honorific forms in order to generate exercises for intermediate learners of Japanese

Valérie Collec-Clerc (www.valtal.fr)

(mail: valerie.clerccollec@yahoo.fr)

LABORATOIRE D'INFORMATIQUE FONDAMENTALE de Marseille

www.lif.univ-mrs.fr

Using corpora to find relevant examples of Japanese honorific forms in order to generate exercises for intermediate learners of Japanese

- Research context
- Why do we need corpus analysis ?
- Corpora and tools
- Adopted methodology for the use of corpora
- Future research perspectives

- **Research context**
- Why do we need corpus analysis ?
- Corpora and tools
- Adopted methodology for the use of corpora
- Future research perspectives

- **Research context**
  - Goal: Generating sentences and exercises for intermediate learners of Japanese
  - Characteristics of Japanese
  - Japanese politeness language in a bird's eye view

- **Research context**
  - **Goal: Generating sentences and exercises for intermediate learners of Japanese**
  - Characteristics of Japanese
  - Japanese politeness language in a bird's eye view

Research context

Goal: Generating sentences and exercises for intermediate learners of Japanese

- Learning aids for the Japanese learners through exercises on sentence structures (like "phrasebook" approach already used for beginners) : **Zock M.** and **Lapalme G. 2010 -** ***A Generic Tool for Creating and Using Multilingual Phrasebooks.***
- Target learners : intermediate level ⇔syntactically complex sentences which rely on semantic and contextual elements (e.g.: Japanese formal politeness language, also called the "honorific" system).

➔ Identifying and formalising "rules" or standard patterns to generate realistic sentences **It does not only consist of syntactic rules but also relies on the semantic/pragmatic constraints of the situation of uttering.**

- **Research context**
  - Goal: Generating sentences and exercises for intermediate learners of Japanese
  - **Characteristics of Japanese**
  - Japanese politeness language in a bird's eye view

Research context
Characteristics of Japanese

- Writing systems :
  - Kanji :自然言語処理 (nouns, verb stems)
  - Syllabic scripts
    - Hiragana しぜんげんごしょり (furigana, grammatical words, inflections)
    - Katakana (フランス) (foreign words)
  - Latin script (romaji) (shizengengoshori)
- No space between words
- Word order (SOV) with case particles after the nouns
- No grammatical gender and number for nouns
- Examples :フランス人の学生が自然言語処理を勉強する。

➔ **Linguistic tools mainly designed for Western languages are not adapted to Japanese**

- **Research context**
  - Goal: Generation of sentences and exercises for intermediate learners of Japanese
  - Characteristics of Japanese
  - **Japanese politeness language in a bird's eye view**

Research context

Japanese politeness language in bird's eye view 1/5

- Focus = formal politeness system (generally regarded as complex for non native speakers)

- The sentence construction used reflects the location of speaker inside the Japanese inter-personal communication system Taiguhyougen 待遇表現)
  - Vertical relationship (jougekankei :上下関係)
    - Individual : superior/inferior (目上 (meue)/目下 (meshita))
    - Group (more important/ important)
  - Horizontal relationship
    - In-group (uchi : 内) (honne :本音)
    - Out-group (soto :外) (tatemae:建前)
    - Psychological distance (shinso : 親疎)

➔ **Contextual facts cannot be ignored to build up a correct sentence**

Research context
Japanese politeness language in a bird's eye view 2/5

Adressee-oriented politeness

- Plain form　(futsukei :普通形)
  - Neutral written language (日本語を学ぶ、きれいだ)
  - Familiar spoken language
- Polite form or Addressee honorifics (Teineikei :丁寧形)
  - Showing respect towards the addressee
    - (masu:　ます＝＞日本語を学びます) (desu:　です＝＞きれいです)
  - Exalted language : Emphasizing the speakers' willingness to show respect towards the addresee
    - (gozaimasu :ございます＝＞きれいでございあす);
    - (teorimasu : ております)

Research context
Japanese politeness language in bird's eye view 3/5

Referential politeness

Often referred as keigo：敬語

- **Subject honorifics, called *Sonkeigo*:** (尊敬語)

  The speakers elevate or show respect towards the subject of the utterance (appreciative towards the non speakers).

- **Non-subject honorifics, called Humility, or *Kenjougo*:** (謙譲語)

  The speakers humble themselves by showing respect to the non-subject referent, generally the object of the utterance (depreciative towards the speakers).

Japanese politeness language in bird's eye view 4/5

Subject honorifics – Main verbal construction

- Passive form
  - (RARERU/RERU) : 今朝の新聞をもう読まれましたか。
    - kesanoshimbun wo mou yomaremashitaka ?
- Honorific construction
  - O +[STEM FORM]+ NI NARU
  - 今朝の新聞をもうお読みになりましたか。
    - Kesanoshimbun wo mou oyomi ni narimashita ka?
- Specific honorific verbs
  - For instance : the use of the verb いらっしゃる (IRASSHARU) (come, go) instead of 来る *kuru* (come) or 行く *iku* (go)

Research context
Japanese politeness language in bird's eye view 5/5

Non-subject honorifics : Main constructions

- ageru （上げる）: as a completive auxiliary
  - 申し上げる （言う : say） 差し上げる (上げる: give）
- o-verbal stem + SURU/ITASU (お読みするoyomi suru : read) (お返し致すokaeshi itasu : return)
- go-verbal noun + SURU (ご案内する goannai suru : show around, guide）
- 拝 Hai (worship) + sino-japanese verb reading + suru
  - 拝見 (haiken) する（見る miru : see）、拝借(haishaku)する（借りる kariru : borrow）

- Research context
- **Why do we need corpus analysis ?**
- Corpora and tools
- Adopted methodology for the use of corpora
- Future research perspectives

Why do we need corpus analysis ?

- Existing formalising approaches = dedicated to a specific approach
  - The speaker
  - The hearer (the addressee)
  - The context of utterance

- Deeper study of the honorific forms taking in account the most relevant documents

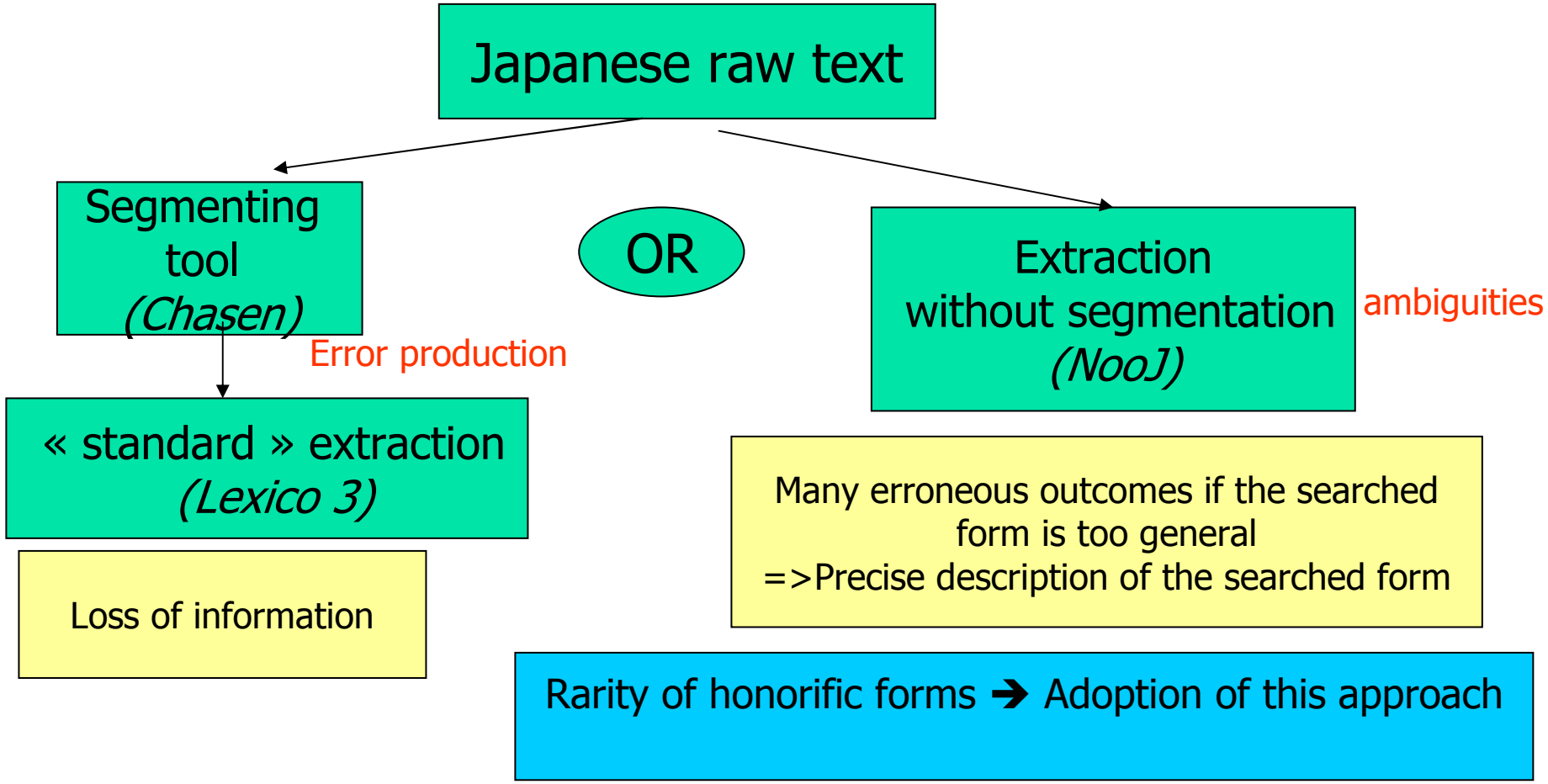
- Identification of the most frequently used honorific forms within the same document gender

- Research context
- Why do we need corpus analysis ?
- **Corpora and tools**
- Adopted methodology for the use of corpora
- Future research perspectives

- **Corpora and tools**
  - Segmentation problem
  - Selected Corpora and tools
  - Open search on the web

- **Corpora and tools**
  - **Segmentation problem**
  - Selected Corpora and tools
  - Open search on the web

Corpora and tools: segmentation problem



- **Corpora and tools**
  - Segmentation problem
  - **Selected corpora and tools**
    - **BCCWJ (corpus)**
    - **SAGACE (corpus + tool)**
    - **NooJ (tool)**
  - Open search on the web

Corpora and tools
BCCWJ: Balanced Corpus of Contemporary Written Japanese

- BCCWJ is a corpus computed between 2005 and 2011, as a contemporary part of the KOTONOHA corpus of written Japanese. It contains more than 100 million words.
- Contents has various origins: but the main part was generated by a systematic sampling of books, newspapers, and official texts,
- It is organized in sub-corpora (texts 2001-2005, books from 1985, best-selling books, web, governmental papers, laws, …)
- On line access is available, to search KWiC. XML CDROM can also be obtained.

Corpora and tools
SAGACE

- SAGACE : a tool to extract morphological forms from non-segmented texts for Asian languages.

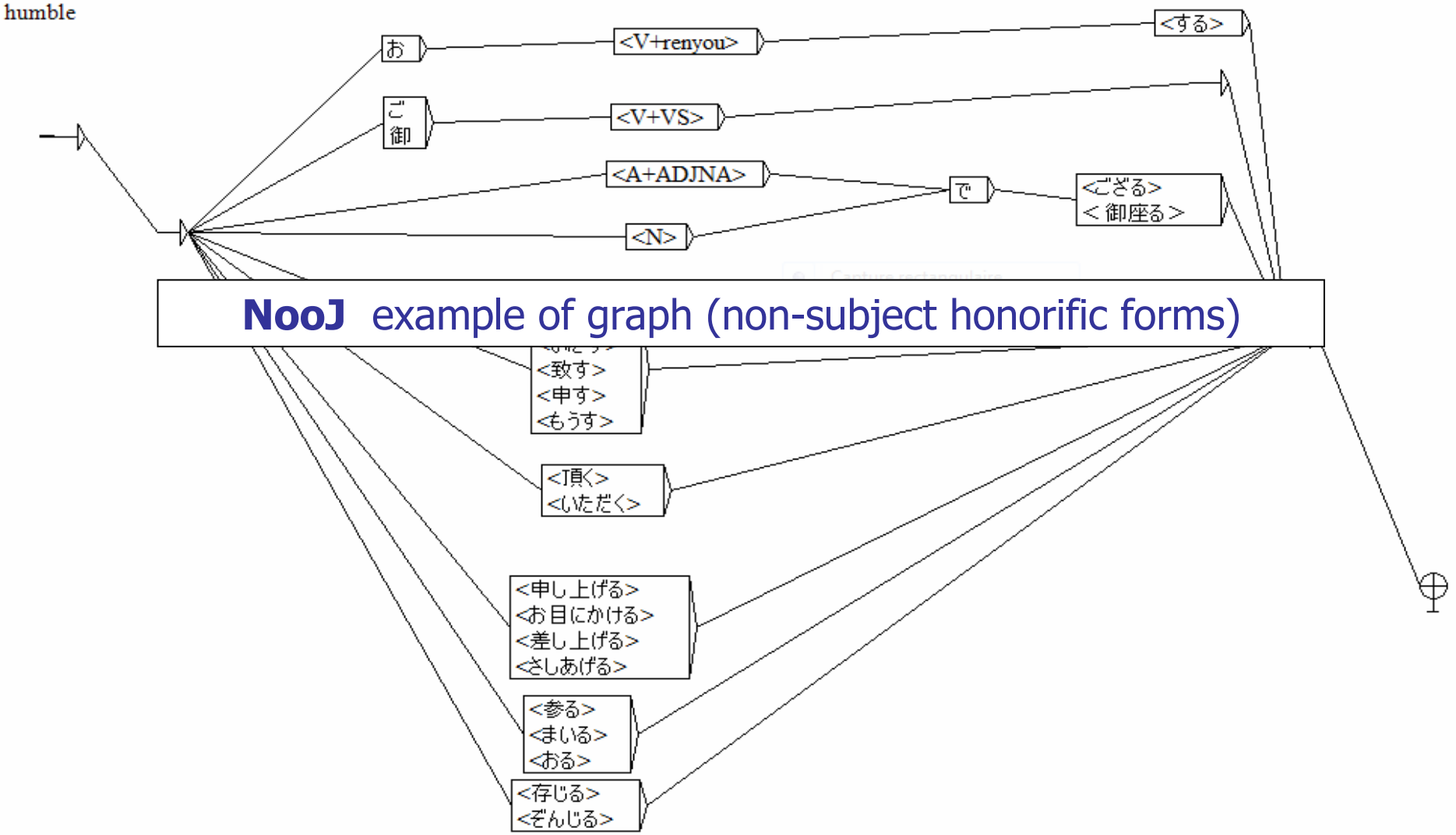
- On line tool is directly usable with Japanese corpora from several sources and genres.

Corpora and tools
NooJ

- Nooj enables word extractions from non-segmented texts and displays the results in a KWiC (Key Word in Context) environment.
- NooJ makes it possible the creation of graphs which describe parts of speech, inflectional and derivational forms (with the help of lexicons).
- Disadvantages : Necessity to implement these resources if not already existing

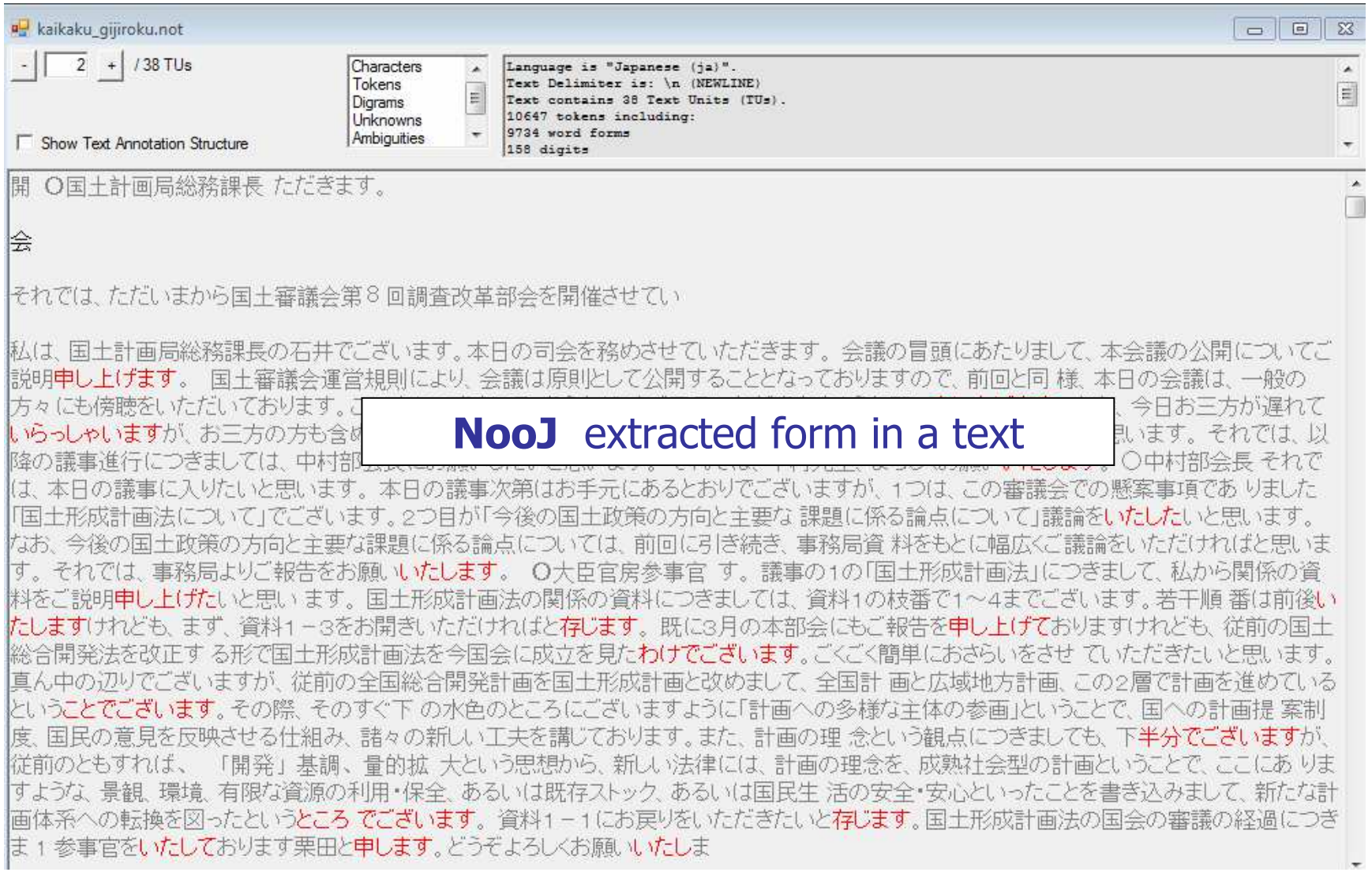
  ➔ Creation of linguistic resources for NooJ (Lexicon, inflectional and derivational rules)

Using corpora to find relevant examples of Japanese honorific forms in order to generate exercises for intermediate learners of Japanese

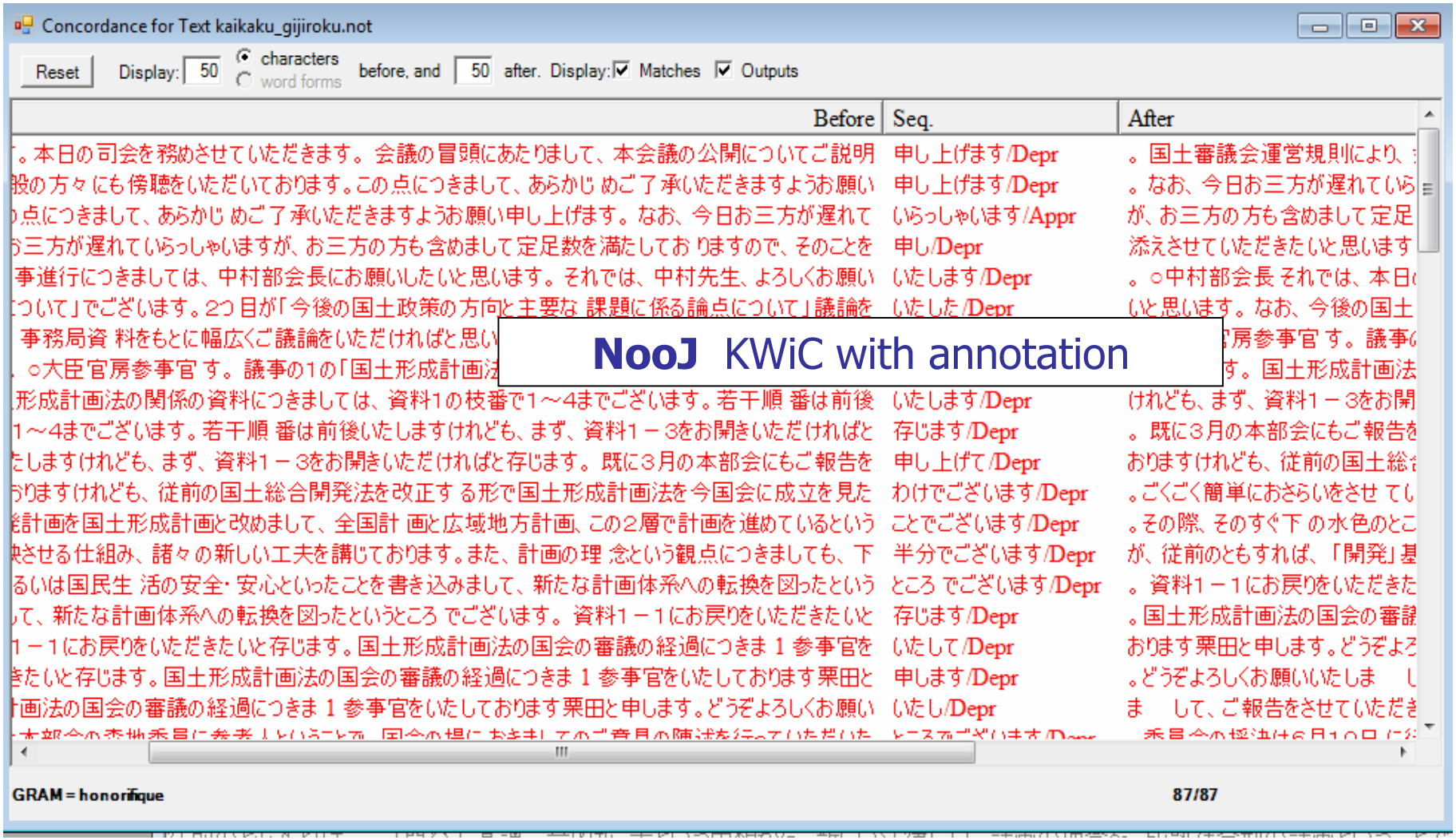


NooJ example of graph (non-subject honorific forms)

Using corpora to find relevant examples of Japanese honorific forms in order to generate exercises for intermediate learners of Japanese

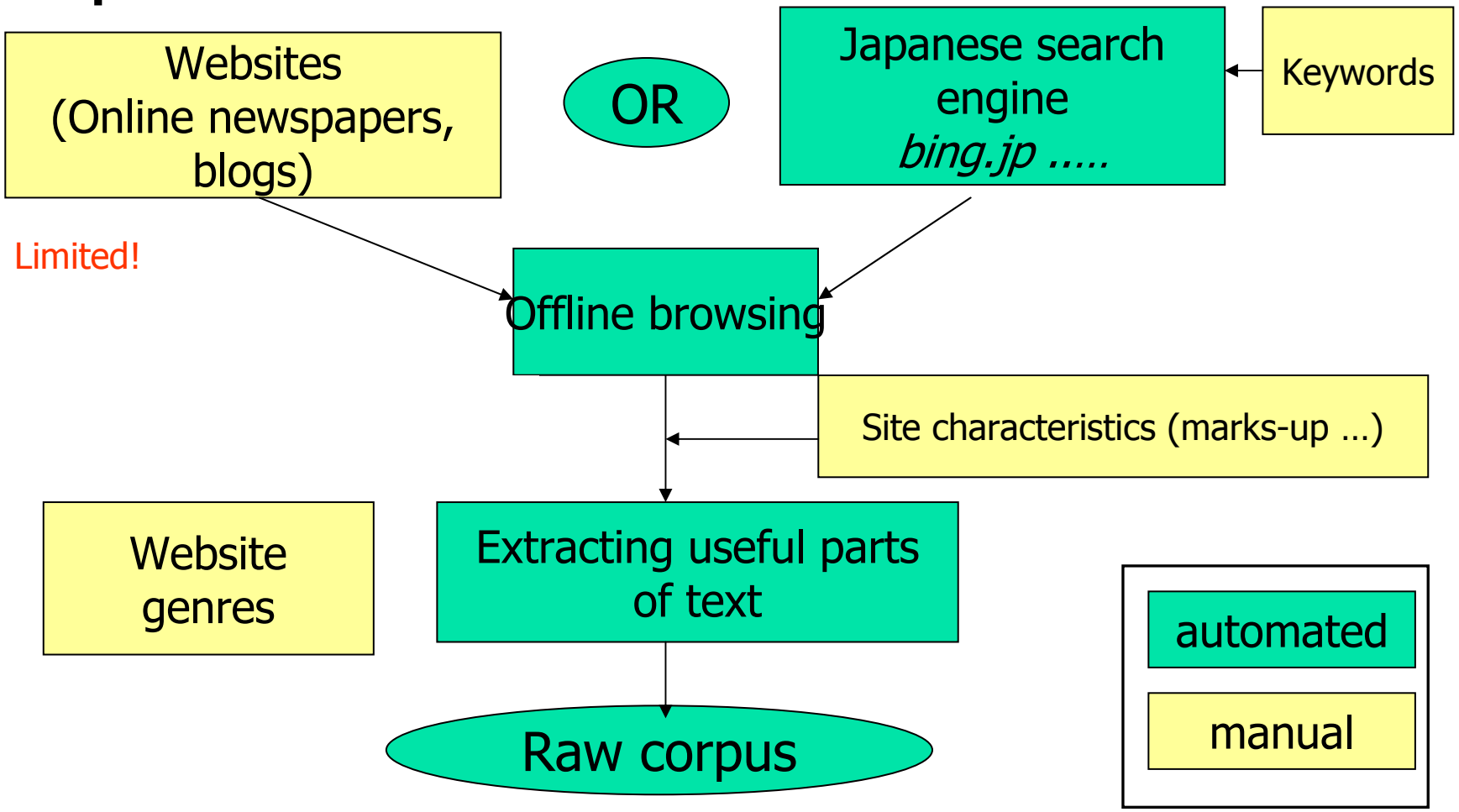


Using corpora to find relevant examples of Japanese honorific forms in order to generate exercises for intermediate learners of Japanese



- **Corpora and tools**
  - Segmentation problem
  - Selected corpora and tools
    - BCCWJ (corpus)
    - SAGACE (corpus + tool)
    - NooJ (tool)
  - **Open search on the web**

- **Open search on the Web**



- Research context
- Why do we need corpus analysis ?
- Corpora and tools
- **Adopted methodology for the use of corpora**
- Future research perspectives

Adopted methodology

- Identifying characteristic forms from other studies on honorifics
- Extracting these forms from Japanese corpus incorporated in parsing tools (Sagace, BCCWJ)
- Identifying the most frequently used honorific forms
- Doing further research on the Web (updates)
- Identifying text-genders and contexts of uttering
- Creating annotated corpus from the raw elements obtained
- Formalizing the rules of generation
- Validating of the generated sentences by native Japanese speakers

- Research context
- Why do we need corpus analysis ?
- Corpora and tools
- Adopted methodology for the use of corpora
- **Future research perspectives**

Future research perspective

- Analysing other texts to increase the corpus size (identification and selection of contexts =a significant time-consuming manual phase)
- Classifying contexts of use and defining the data representation for context of the utterance (relational Database / ontology?) (Current prototyping in Prolog)
- Implementing the generated sentences in a e-learning environment
- Application to other linguistic contexts

Bibliography

- **Blin, R. 2001**. *Traitement automatique du japonais et études linguistiques*. Histoire Epistémologie Langage vol 23 n° 1 p. 33-48.
- **Kishimoto, H. 2010.** *Possessor Raising and Two Types of Honorification*, NANZAN Linguistics No.6, p 1-12.
- **Maekawa, K. 2007** "Design of a Balanced Corpus of Contemporary Written Japanese." *Proceedings of Symposium on Large-Scale Knowledge Resources (LKR2007)*, pp.55-58, 2007:03.
- **Rastier, F. (dir.) , 1995** *L'analyse thématique des données textuelles*, Paris : Didier, 1995 p. 223-249.
  Texte légèrement remanié.)
- **Schikowski, R. 2008** *Skript zum Strukturkurs Japanisch*. LMU Munich, Institute of General and Typological Linguistics, winter term 2007/08
- **Shimamori, R. 2001** *Grammaire japonaise systématique*, volume II - Edition Jean Maisonneuve Paris 12ème édition, 2001

Bibliography

- **Siegel M. 2000.** *Japanese Honorification in an HPSG Framework,* Proceedings of the 14th Pacific Asia Conference on Language, Information and Computation. p. 289-300.
- **Sugimura, R. 1986.** *Japanese honorifics and Situation Semantics* International Conference on Computational Linguistics COLONG. p. 507-510.
- **Terrya, K. 2007**. *Interpersonal grammar of Japanese* in A Systemic functional grammar of Japanese**,** Vol 2, Ch 4**,** p135-205.
- **Trusty, A. / Truong, K 2011** *Augmenting the Web for Second Language Vocabulary Learning Conference on Human Factors in Computing System (CH)., pp 3179-3188*
- **Wlodarczyk, A. 1996**. *Politesse et Personne – Le japonais face aux langues occidentales*. Editions L'harmattan.
- **Wlodarczyk, A. 2007**. *Towards a Unified Treatment of Linguistic - Person and Respect - Identification* Japanese Linguistics – European Chapter Tokyo.
- **Zock M. / Lapalme G. 2010** *A Generic Tool for Creating and Using Multilingual Phrasebooks.* Natural Language Processing and Cognitive Science.

Merci vielmol fers zuhere