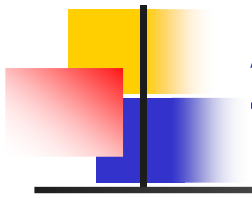
NooJ Conference 2013 – Saarbrücken

Adapting existing Japanese linguistic resources to build a NooJ dictionary to recognize honorific forms

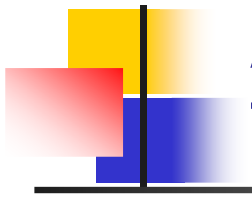
Valérie Collec-Clerc (**www.valtal.fr**)
(mail: valerie.clerccollec@yahoo.fr)

LABORATOIRE
D'INFORMATIQUE
FONDAMENTALE
de Marseille
www.lif.univ-mrs.fr



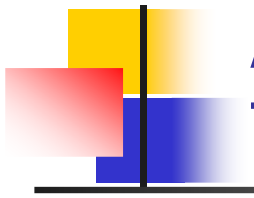
Adapting existing Japanese linguistic resources to build a NooJ dictionary to recognize honorific forms

- Context of study
- Overview of the Japanese language and its honorific system
- Creating linguistic resources for NooJ software
- Examples of graphs recognising honorific forms
- Developments in prospect



Adapting existing Japanese linguistic resources to build a NooJ dictionary to recognize honorific forms

- **Context of study**
- Overview of the Japanese language and its honorific system
- Creating linguistic resources for NooJ software
- Examples of graphs recognising honorific forms
- Developments in prospect



Adapting existing Japanese linguistic resources to build a NooJ dictionary to recognize honorific forms

Context of study

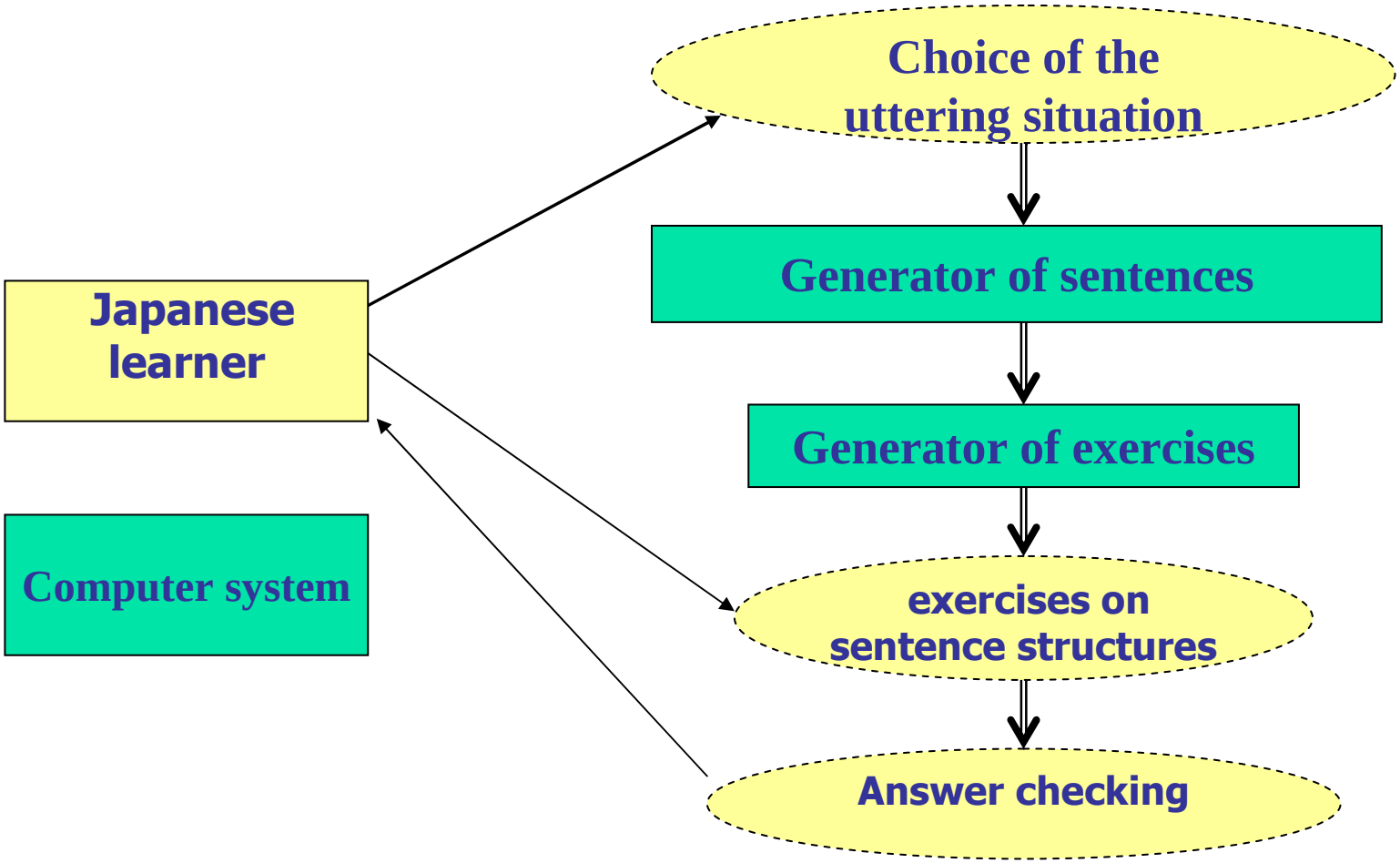
- Learning aids for the Japanese language through exercises on sentence structures.

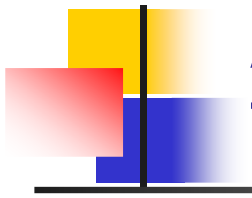
- Target learners : intermediate level ⇔syntactically complex sentences which rely on semantic and contextual elements (e.g.: Japanese honorific language).

- Formalising syntactic rules
  - Building up and analysing corpora
  - Corpora : Drawbacks : Difficulties in finding texts or existing corpora meeting our requirement.

Adapting existing Japanese linguistic resources to build a NooJ dictionary to recognize honorific forms

Context of study





Adapting existing Japanese linguistic resources to build a NooJ dictionary to recognize honorific forms

- Context of study
- **Overview of the Japanese language and its honorific system**
- Creating linguistic resources in Japanese
- Examples of graphs recognising honorific forms
- Developments in prospect

Overview of the Japanese language and its honorific system

- Brief description
- The honorific system
- Segmentation problems

Characteristics of the Japanese language (brief description)

- 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勉強する。

The honorific system

Taiguhyougen 待遇表現 = The Japanese inter-personal communication system

Two axes of relationship

- Vertical relationship (jougekankei :上下関係)
  - Individual : superior/inferior (目上 (meue)/目下 (meshita))
  - Group (important＜more important)
- Horizontal relationship
  - In-group (uchi : 内) (honne :本音)
  - Out-group (soto :外) (tatemae:建前)
  - Psychological distance (shinso : 親疎)

Neutral form : Superior=>Inferior, In-group=>In-group
Polite form : Inferior=>Superior, In-group=>Out-group

Neutral form + Polite form = Enunciative politeness

The honorific system

Enunciative politeness

- Neutral 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 listener
    - (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 listener
    - (gozaimasu :ございます＝＞きれいでございあす);
    - (teorimasu : ております)

The honorific system

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).

The honorific system

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)

The honorific system

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）

Segmentation problem

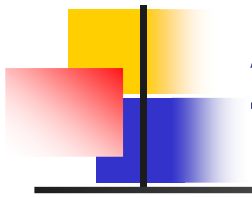
Recognizing syntactic forms in raw corpora (not annotated) is the major problem because of the absence of space between words.

- Common NLP approach = Starting with a phase of segmentation

  Texts are pre-processed with morpho-syntatic parsers like *Chasen* and *MeCab* in order to artificially insert separators.
  Drawbacks :These tools are not completely reliable (based on limited dictionaries and statistical data on sequence of POS for disambiguation).

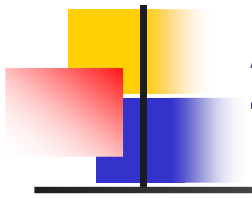
- For a finer analysis, we need a tool that works on non segmented text and deal with syntactic expression. ➔ **use of NooJ**
  Drawbacks : **No Japanese resources available in the standard NooJ repository.**

- A first approach was made by Claire OLIVIER (2009) to create a Japanese NooJ dictionary. It consisted in extracting syntactic forms from corpora according to formal syntactic rules.
  Drawbacks : Time-consuming, generation of a small-sized dictionary.



Adapting existing Japanese linguistic resources to build a NooJ dictionary to recognize honorific forms

- Context of study
- Overview of the Japanese language and its honorific system
- **Creating linguistic resources in Japanese**
- Examples of graphs recognising honorific forms
- Developments in prospect



Adapting existing Japanese linguistic resources to build a NooJ dictionary to recognize honorific forms

Creating linguistics resources in Japanese

- Dictionary
- Inflections/Derivations
- Examples

NooJ Dictionary

- Based on existing resources from JDIC/EDICT Project
  - JDIC/EDICT is a set of reference of Japanese dictionaries (bilingual, monolingual on a given domain, etc.) of several formats (on line, XML, text) and, coordinated by **Jim Breen** (Monach University Australia) which is linked to other international projects (Wordnet, …)
  - Full resources are made up of more than 150 000 entries
  - We use a medium-sized Japanese/German/English lexicon (13000 entries) with kanji / kana equivalents and lexical JDIC tags (parts of speech, types of verbs, semantic domain, ..)
- PERL program to filter and generate a "NooJ" format
  - ➔ 15000 NooJ .dic entries

NooJ Dictionary

- **Filtering principles adopted to help disambiguation**
  - Reject of entries with archaic forms.
  - Standard Kanji/Kana form is the unique lemma (no entry for equivalent phonetic kana (furigana) or romaji).
  - No named entities in the basic dictionary (additional dictionaries if needed).
  - The tags "prefix" and "suffix" as parts of speech are limited to words which cannot be used differently (e.g. nouns also used as prefixes are only categorised as nouns).
  - Small number of basic POS with subclasses (adjective ending in i=> A+ADJI).

Adapting existing Japanese linguistic resources to build a NooJ dictionary to recognize honorific forms

NooJ Dictionary

Main classes (Part of Speech)

- **N** noun
- **ENAM** named entity / proper noun
- **A** adjective
- **V** verb
- **AV** adverb
- **CONJ** conjunction
- **INTJ** interjection

subclasses for nouns (N)

- Abbrev noun entry is an abbreviation
- CPREF noun also used as a prefix
- CSUF noun also used as a suffix
- VS base to suru verb:
- ANO base to no adjective-no: a

subclasses for adjectives (A)

- ADJI i-adjective
- ADJNA na-adjective
- ADJF adjective pronominally
- ADJPN pre-noun adjectival (rentaishi)
- ADJTARU adjective taru
- ADJAUX auxiliary adjective
- ADJS adjective like ookii
- CPREF adjective also used as a prefix
- CSUF adjective also used as a suffix

- **NUM** numeral
- **CTR** counter
- **PART** particle
- **PN** pronoun
- **PREF** prefix (**of no other class**)
- **SUF** suffix (**of no other class**)
- **EXP** complete expression

subclasses for verbs (V)

- VT transitive verb
- VI intransitive verb
- VAUX auxiliary verb
- Copula to identify だ copula
- CPREF verb also used as a prefix
- CSUF verb also used as a suffix

subclasses for adverbs (AV)

- ADVTO adverb to
- VS base to suru-verb:
- CPREF adverb also used as a prefix
- CSUF adverb also used as a suffix

subclasses for particles (PART)

- PCAS casual particle
- PREL relational particle
- PADV adverbial particle
- PCOORD coordinate particle
- PEND final particle

NooJ Dictionary

Additional tags for entries	
+Onoma	Onomatopoeia
+Ateji	The kanji that compose this lemma are only used for their phonetic values
+Hira	The lemma is rarely used with its kanji form
+Anat +Archi +Astron +Biol + …	Semantic field
FLX=	NooJ standard for inflection for verbs and i-adjectives
DRV=	NooJ standard for suru derivation and adjectives in no derivation
KANA=	Kana form
DE=	Suggested equivalence in German
EN=	Suggested equivalence in English

Inflection/Derivations

- Derivations
  - Generation of additional verbs from the light verb するsuru (do, make) for some specific nouns or adjectives.
    勉強 benkyou (studies) => 勉強する benkyousuru (to study)
    - File "derivation.nof"
- Inflections
  - Tenses for Japanese adjectives (i-adjectives)
    - File "adj_inflection.nof" (5 forms)
  - Tenses and modes for the different categories of verbs and their endings (up to 20 types of verbs, including irregular verbs)
    - File "verb_inflection.nof" (45 verbal forms / 20 types of verbs)
- From 15 000 entries of .dic ➔ 210 000 recognized forms

Examples

JDIC source

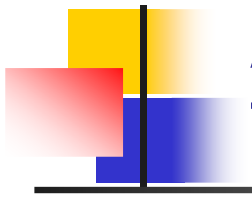
輸出 ゆしゅつ 3 (n,vs,adj-no,adj-na) export,efferent (medical)
Ausfuhr,Export

習う ならう 5 (v5u,vt) to take lessons in, to be taught,
to learn (from a teacher), to study (under a teacher), to get training in
lernen, studieren, Unterricht nehmen, rel. üben

Generated .dic

輸出,**N**+ANO+VS+DRV=ADJNO+DRV=VSURU:SURU+KANA=ゆしゅつ
+EN=export+DE=Ausfuhr
輸出,**A**+ADJNA+KANA=ゆしゅつ+EN=export+DE=Ausfuhr

習う,**V**+VT+FLX=AU+KANA=ならう+EN=to take lessons in+DE=lernen

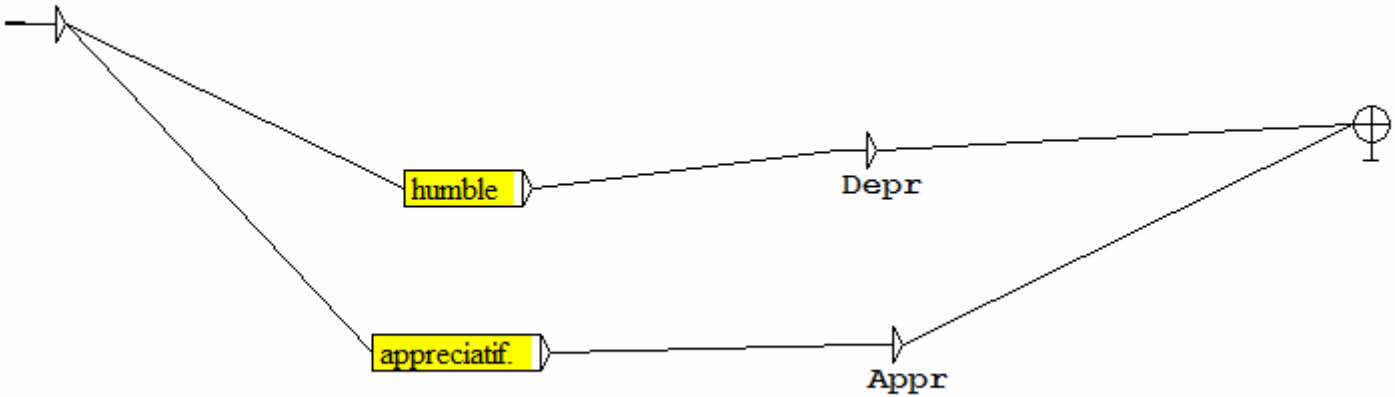


Adapting existing Japanese linguistic resources to build a NooJ dictionary to recognize honorific forms

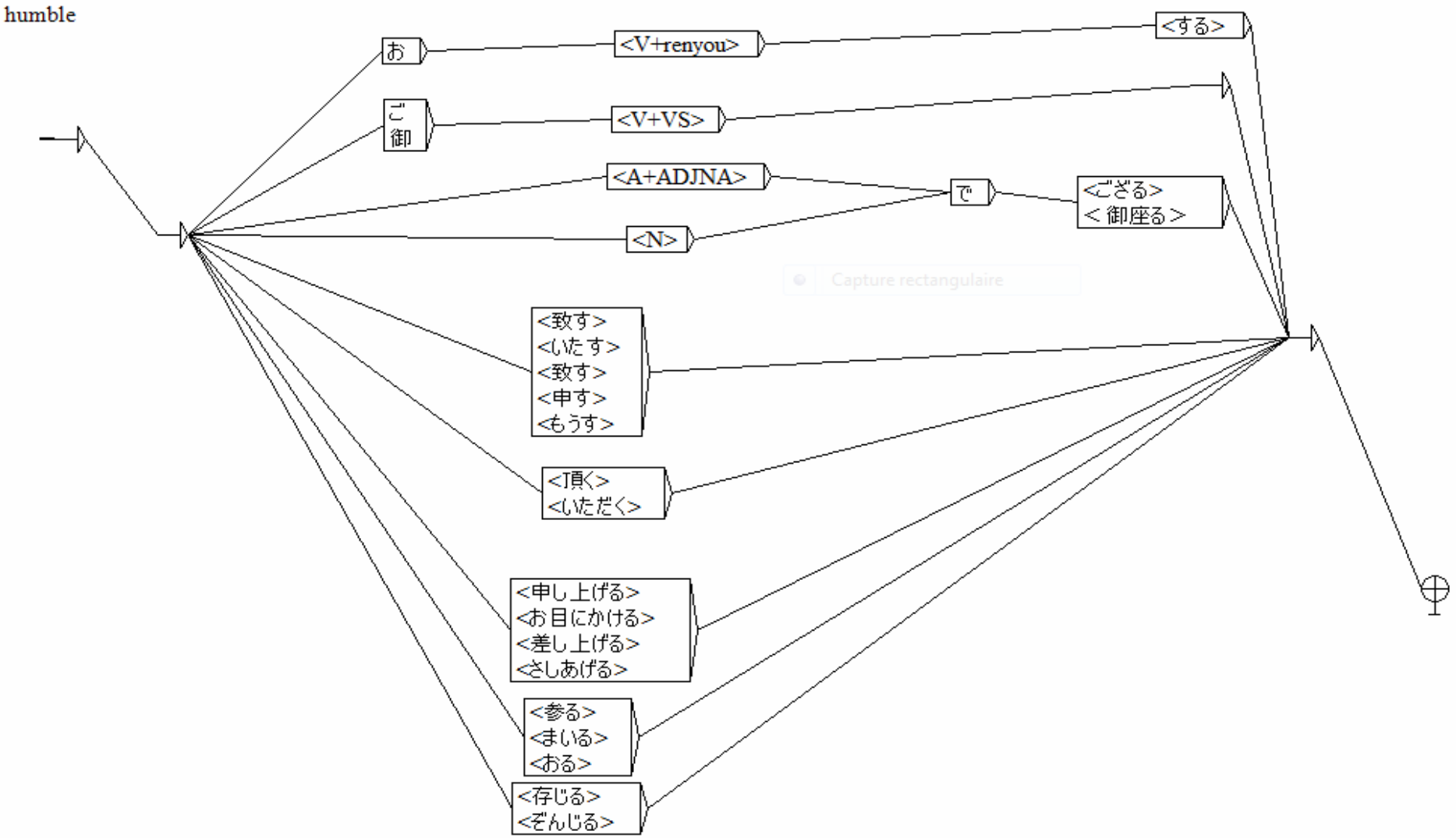
- Context of study
- Overwiew of the Japanese language and its honorific system
- Creating linguistic resources in Japanese
- **Examples of graphs recognising honorific forms**
- Development in prospects

Adapting existing Japanese linguistic resources to build a NooJ dictionary to recognize honorific forms

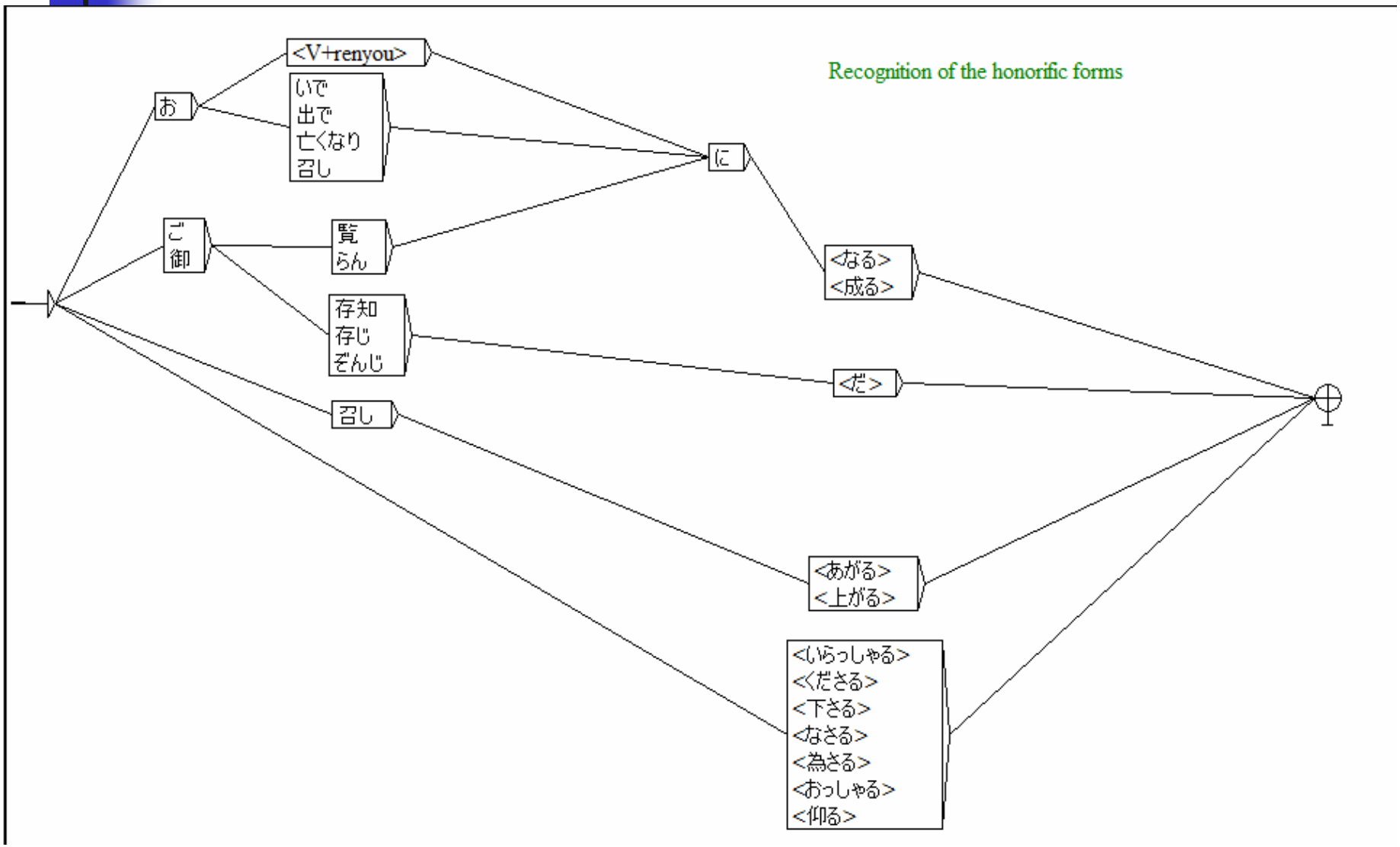
Recognition graphs



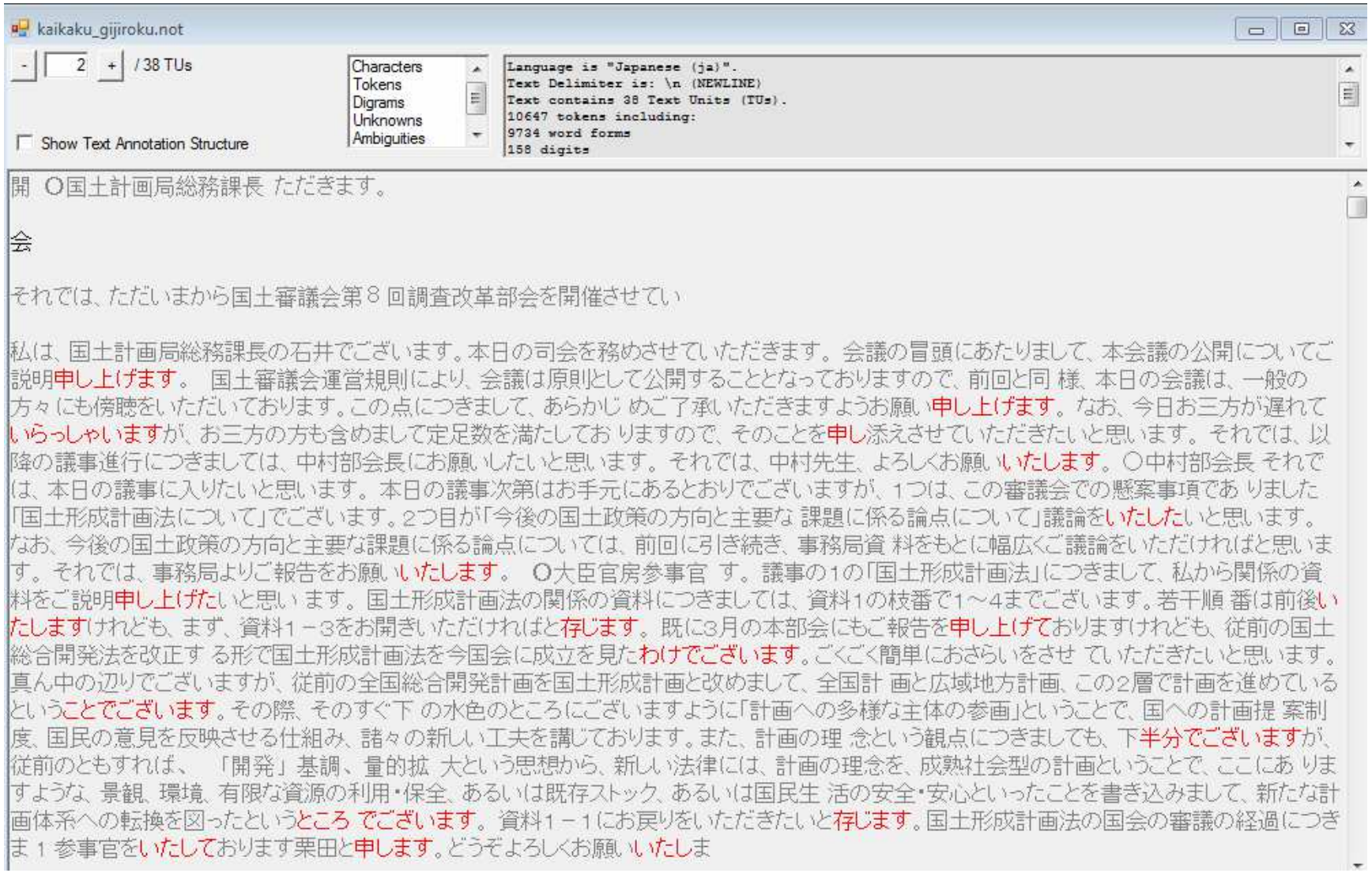
Adapting existing Japanese linguistic resources to build a NooJ dictionary to recognize honorific forms



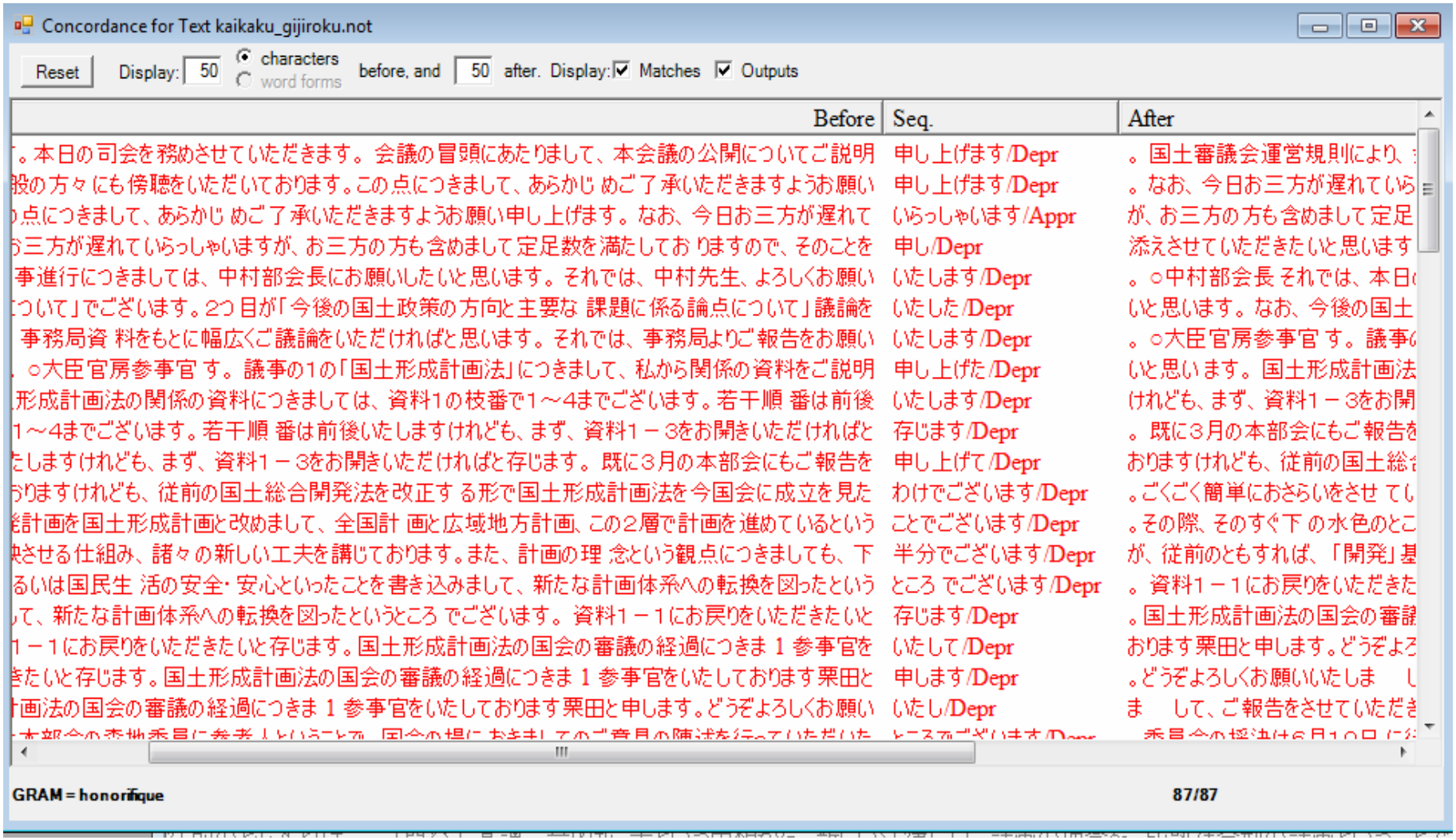
Adapting existing Japanese linguistic resources to build a NooJ dictionary to recognize honorific forms

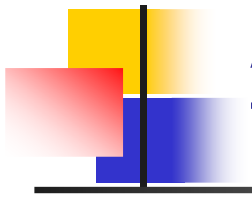


Adapting existing Japanese linguistic resources to build a NooJ dictionary to recognize honorific forms



Adapting existing Japanese linguistic resources to build a NooJ dictionary to recognize honorific forms





Adapting existing Japanese linguistic resources to build a NooJ dictionary to recognize honorific forms

- Context of study
- Overview of the Japanese language and its honorific system
- Creating linguistic resources in Japanese
- Examples of graphs recognising honorific forms
- **Developments in prospect**

Developments in prospect

- NooJ community
  - Share these resources to obtain feedback from other users
  - Graphs for disambiguation

- Resources
  - Increasing the number of entries (full EDICT resource ?)
  - Improving translation filters (German, Wordnet link)
  - Increasing the number of Corpora: use of our approach in our context of study (on going work)

Vielen Dank für Ihre Aufmerksamkeit