



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



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



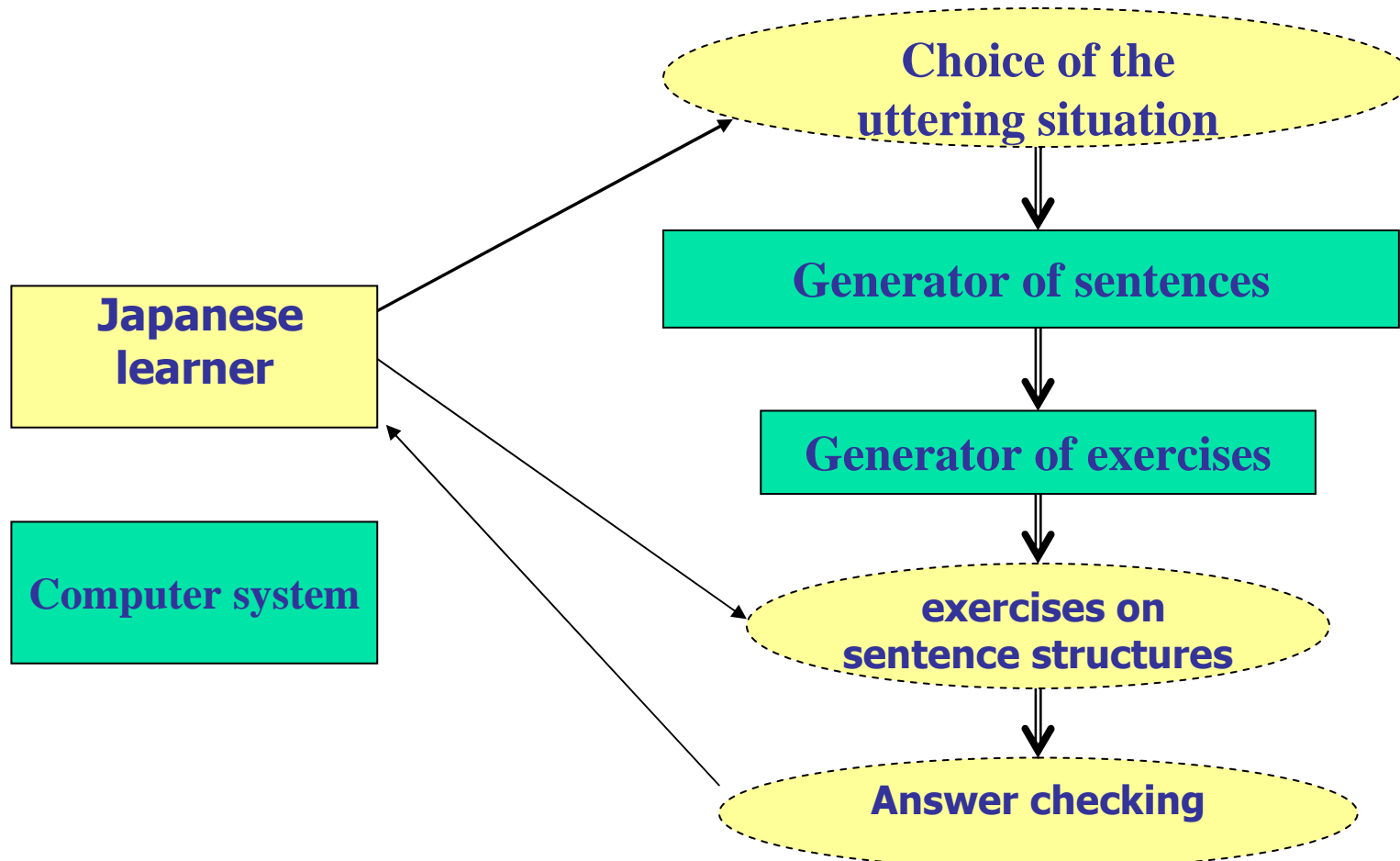
Adapting existing Japanese linguistic resources to build a NooJ dictionary to recognize honorific forms

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



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

Overview of the Japanese language and its honorific system

- Brief description
- The honorific system
- Segmentation problems



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

Characteristics of the Japanese language (brief description)

- Writing systems :
 - Kanji : 自然言語処理 (nouns, verb stems)
 - Syllabic scripts
 - Hiragana しぜんげんごしより (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。



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

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



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

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 => きれいでございます);
 - (teorimasu : ております)



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

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



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

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)



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

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 (haihaku) する (借りる kariru : borrow)



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

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-syntactic 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



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

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



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

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	▪ NUM	numeral
▪ ENAM	named entity / proper noun	▪ CTR	counter
▪ A	adjective	▪ PART	particle
▪ V	verb	▪ PN	pronoun
▪ AV	adverb	▪ PREF	prefix (of no other class)
▪ CONJ	conjunction	▪ SUF	suffix (of no other class)
▪ INTJ	interjection	▪ EXP	complete expression

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

subclasses for verbs (V)

▪ VT	transitive verb
▪ VI	intransitive verb
▪ VAUX	auxiliary verb
▪ Copula	to identify ㇿ 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



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

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



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

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



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



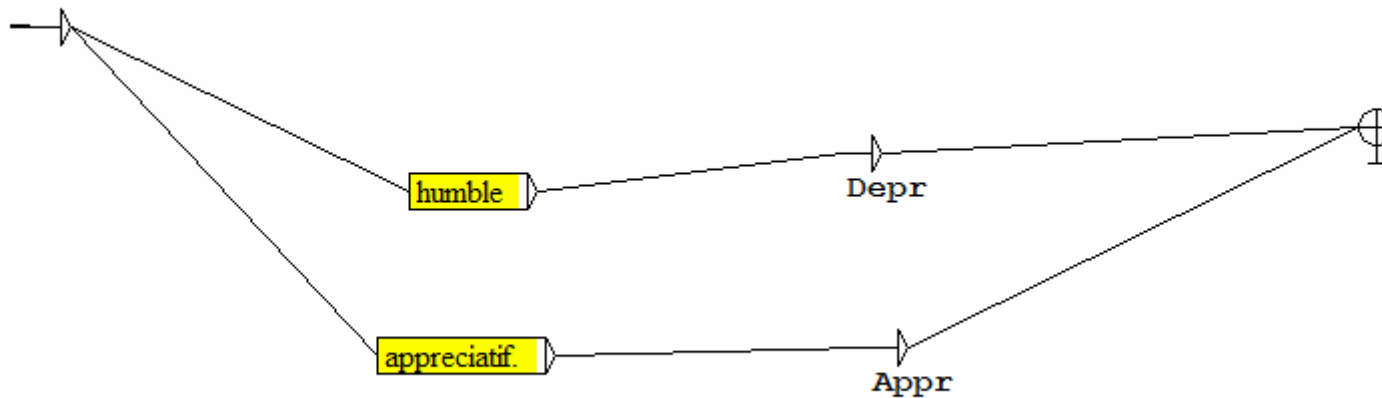
Adapting existing Japanese linguistic resources to build a NooJ dictionary to recognize honorific forms

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



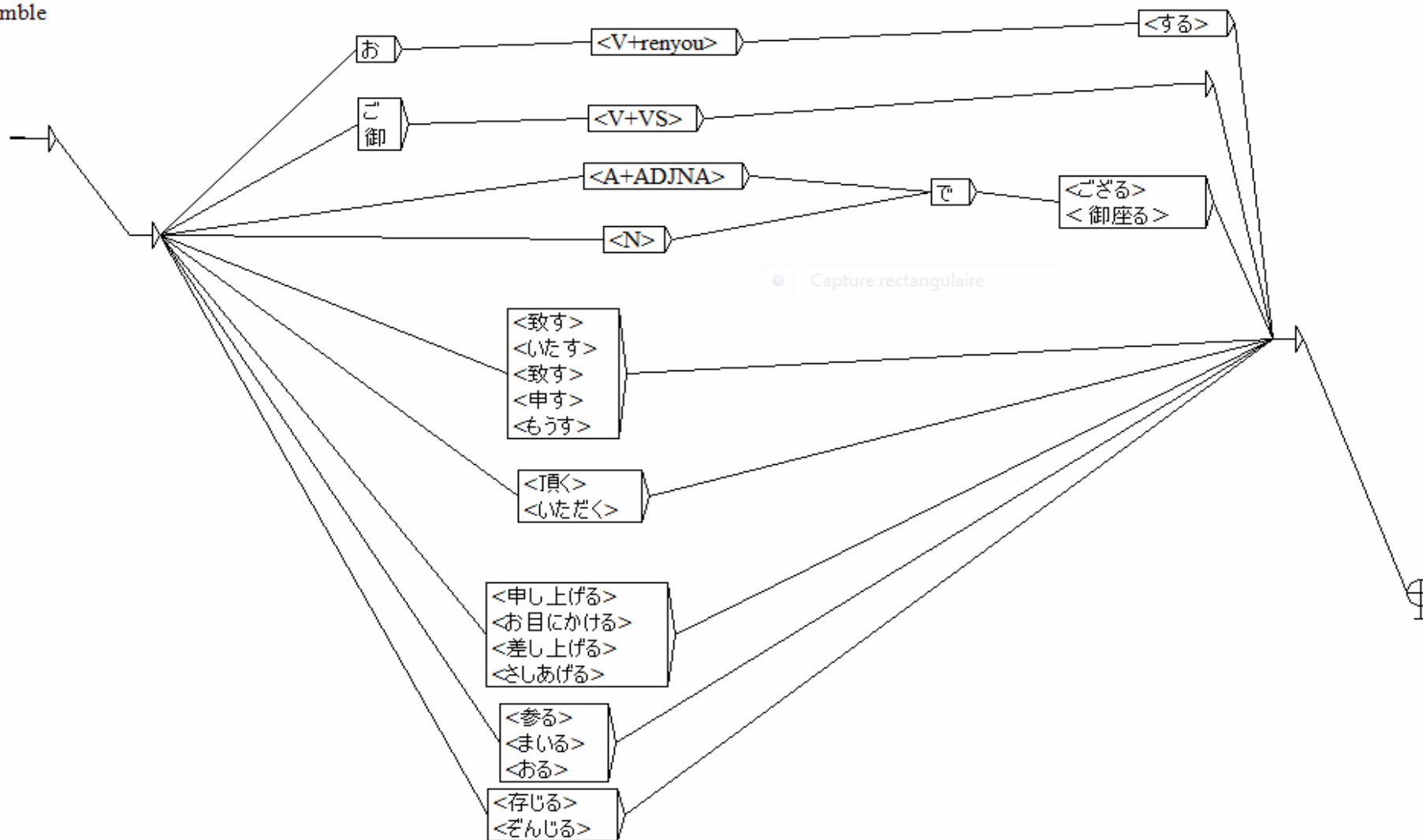
Adapting existing Japanese linguistic resources to build a NooJ dictionary to recognize honorific forms

Recognition graphs

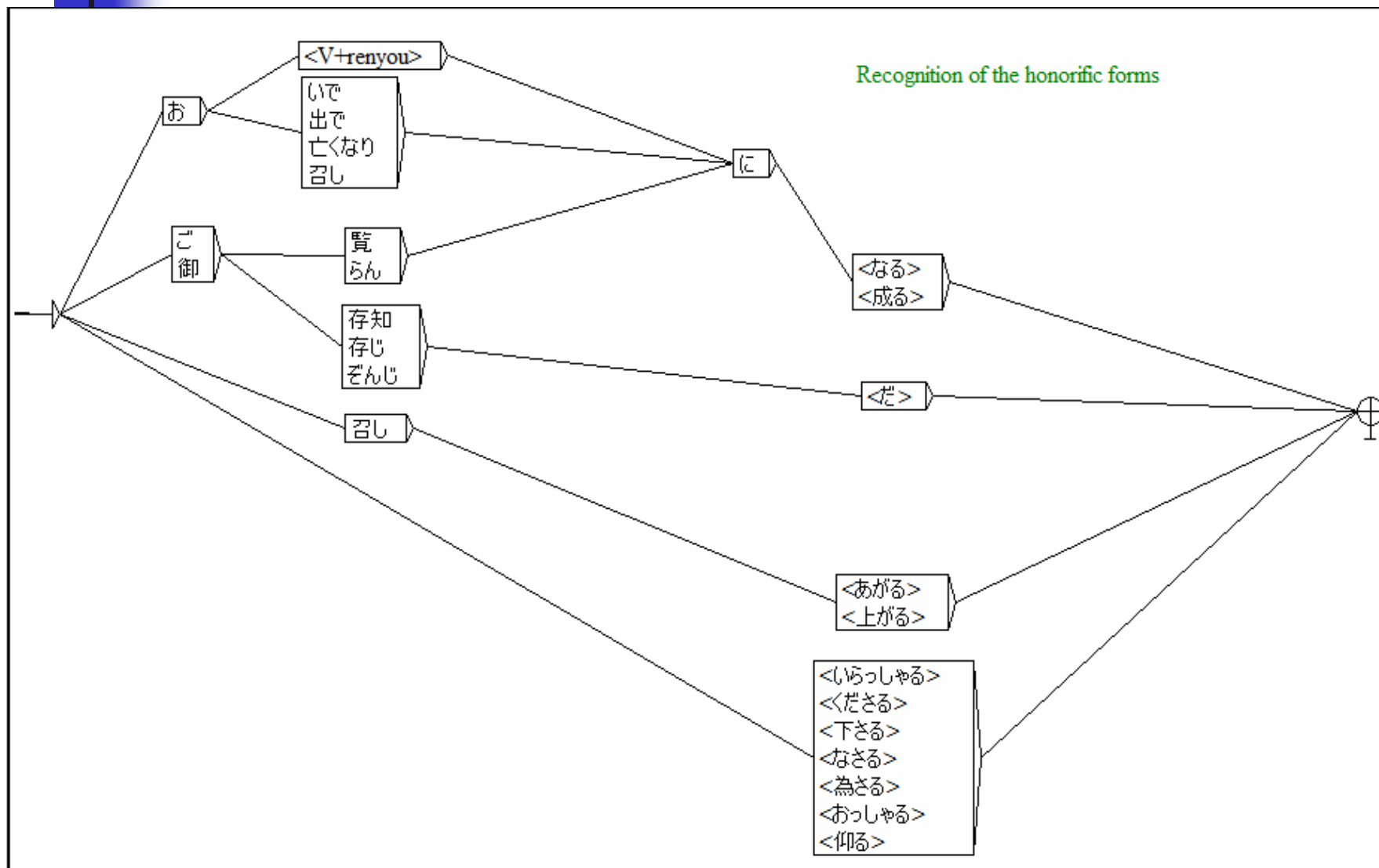


Adapting existing Japanese linguistic resources to build a NooJ dictionary to recognize honorific forms

humble



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

開 ○国土計画局総務課長 いただきます。

会

それでは、ただいまから国土審議会第8回調査改革部会を開催させてい

私は、国土計画局総務課長の石井でございます。本日の司会を務めさせていただきます。会議の冒頭にあたりまして、本会議の公開についてご説明申し上げます。国土審議会運営規則により、会議は原則として公開することとなっておりますので、前回と同様、本日の会議は、一般の方々にも傍聴をいただいております。この点につきまして、あらかじめご了承くださいませようお願いします。なお、今日お三方が遅れていらっしゃいますが、お三方の方も含めまして定足数を満たしておりますので、そのことを申し添えさせていただきますと思います。それでは、以降の議事進行につきましては、中村部会長にお願いしたいと思います。それでは、中村先生、よろしくお願いいたします。○中村部会長 それでは、本日の議事に入りたいと思います。本日の議事次第はお手元にあるとおりでございますが、1つは、この審議会での懸案事項でありました「国土形成計画法について」でございます。2つ目が「今後の国土政策の方向と主要な課題に係る論点について」議論をいたしたいと思います。なお、今後の国土政策の方向と主要な課題に係る論点については、前回に引き続き、事務局資料をもとに幅広くご議論をいただければと思います。それでは、事務局よりご報告をお願いいたします。○大臣官房参事官 す。議事の1の「国土形成計画法」につきまして、私から関係の資料をご説明申し上げます。国土形成計画法の関係の資料につきましては、資料1の枝番で1~4まででございます。若干順番は前後いたしますけれども、まず、資料1-3をお開きいただければと存じます。既に3月の本部会にもご報告を申し上げますけれども、従前の国土総合開発法を改正する形で国土形成計画法を今国会に成立を見たわけでございます。ごくごく簡単におさらいをさせていただきますと思います。真ん中の辺りでございますが、従前の全国総合開発計画を国土形成計画と改めまして、全国計画と広域地方計画、この2層で計画を進めているということでございます。その際、そのすぐ下の水色のところがございますように「計画への多様な主体の参画」ということで、国への計画提案制度、国民の意見を反映させる仕組み、諸々の新しい工夫を講じております。また、計画の理念という観点につきましても、下半分でございますが、従前のとすれば、「開発」基調、量的拡大という思想から、新しい法律には、計画の理念を、成熟社会型の計画ということで、ここにありませうような、景観、環境、有限な資源の利用・保全、あるいは既存ストック、あるいは国民生活の安全・安心といったことを書き込みまして、新たな計画体系への転換を図ったところでございます。資料1-1にお戻りをいただきたいと思います。国土形成計画法の国会の審議の経過につきまして1参事官をいたしております栗田と申します。どうぞよろしくお願いいたします。

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

Concordance for Text kaikaku_gijiroku.not

Reset Display: 50 characters before, and 50 after. Display: Matches Outputs

Before	Seq.	After
。本日の司会を務めさせていただきます。会議の冒頭にあたりまして、本会議の公開についてご説明 般の方々にも傍聴をいただいております。この点につきまして、あらかじめご了承くださいませようお願い の点につきまして、あらかじめご了承くださいませようお願い申し上げます。なお、今日お三方が遅れて お三方が遅れていらっしゃいますが、お三方の方も含めまして定足数を満たしておりますので、そのことを 事進行につきましては、中村部会長にお願いしたいと思ひます。それでは、中村先生、よろしくお願ひ について」でございます。2つ目が「今後の国土政策の方向と主要な課題に係る論点について」議論を 事務局資料をもとに幅広くご議論をいただければと思ひます。それでは、事務局よりご報告をお願ひ 。大臣官房参事官 ず。議事の1の「国土形成計画法」につきまして、私から関係の資料をご説明 形成計画法の関係の資料につきましては、資料1の枝番で1～4まででございます。若干順番は前後 1～4まででございます。若干順番は前後いたしますけれども、まず、資料1-3をお開きいただければと いたしますけれども、まず、資料1-3をお開きいただければと存じます。既に3月の本部会にもご報告を おりますけれども、従前の国土総合開発法を改正する形で国土形成計画法を今国会に成立を見た る計画を国土形成計画と改めまして、全国計画と広域地方計画、この2層で計画を進めているという 決させる仕組み、諸々の新しい工夫を講じております。また、計画の理念という観点につきましても、下 るいは国民生活の安全・安心といったことを書き込みまして、新たな計画体系への転換を図ったという て、新たな計画体系への転換を図ったというところでございます。資料1-1にお戻りをいただきたいと 1-1にお戻りをいただきたいと存じます。国土形成計画法の国会の審議の経過につきま 1 参事官を きたいと存じます。国土形成計画法の国会の審議の経過につきま 1 参事官をいたしております栗田と 計画法の国会の審議の経過につきま 1 参事官をいたしております栗田と申します。どうぞよろしくお願ひ 本部会の審議委員に参考とということも、国会の場におきましての御意見の陳述を行っていただいた と存じます/Depr	申し上げます/Depr 申し上げます/Depr いらっしゃいます/Appr 申し/Depr いたします/Depr いたしました/Depr いたします/Depr 申し上げた/Depr いたします/Depr 存じます/Depr 申し上げて/Depr わけでございます/Depr ことでございます/Depr 半分でございます/Depr ところでございます/Depr 存じます/Depr いたして/Depr 申します/Depr いたし/Depr	。国土審議会運営規則により、 。なお、今日お三方が遅れていら が、お三方の方も含めまして定足 添えさせていただきたいと思ひます 。○中村部会長 それでは、本日 いと思ひます。なお、今後の国土 。○大臣官房参事官 ず。議事の いと思ひます。国土形成計画法 けれども、まず、資料1-3をお開 。既に3月の本部会にもご報告を おりますけれども、従前の国土総 。ごくごく簡単におさらいをさせ 。その際、そのすぐ下の水色のと が、従前のともすれば、「開発」基 。資料1-1にお戻りをいただき 。国土形成計画法の国会の審議 おります栗田と申します。どうぞよ 。どうぞよろしくお願ひいたしま まして、ご報告をさせていただき 委員会の採決は6月10日に決

GRAM = honorifique 87/87



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

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)



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

Vielen Dank für Ihre Aufmerksamkeit

