HG351 Corpus Linguistics

Case Studies: Pronouns and Classifiers

Francis Bond
Division of Linguistics and Multilingual Studies
http://www3.ntu.edu.sg/home/fcbond/
bond@ieee.org

Lecture 8
http://compling.hss.ntu.edu.sg/courses/hg3051/

HG3051 (2018)
Overview

➢ Revision of Lexical, Morphological and Syntactic Studies
  ➢ Lexical Studies
  ➢ Grammatical Studies
  ➢ Variation

➢ Case Studies
  ➢ Pronouns
  ➢ Classifiers
Revision of Lexical, Morphological and Syntactic Studies
Corpus Studies of Lexicography
Discussion *big, large, great*

- **big** mainly for concrete things
- **large** mainly for amounts and numbers
- **great** similar to **large** but many special senses
  - **great deal**
  - **great man**
  - **great burrow**
  - **great relative**

  also use as intensifier **great big, great importance**
Corpus Studies of Morphology
Discussion

- **-[ts]ion** more common in Academic (but common everywhere)
  basic use is to make an action non-agentive
  
  - *It provides a direct indication of fuel consumption.*

- **-ment** often used for mental states
  agreement, amazement, embarrassment
  
  - *Patrick shrugged in embarrassment.*

- **-ness** used for personal qualities
  bitterness, happiness, politeness
  
  - *The bitterness in his heart was mixed with . . . .*

  It would be good if we could automatically divide the words according to their semantic field (which we can approximate with WordNet, . . . )
Corpus Studies of Syntax
Discussion

Typically *start* is used to show the onset of a process, often with an adverb

- *The soil formation process may start again in the fresh material*
- *The train started down the hill*

*begin* is used with more concrete agents

- *Then I began to laugh a bit.*
- *The original mass of gas cooled and began to contract.*

Because the corpus doesn’t mark *animacy* or *concrete agent* these statements are weak: we can’t really make predictions or measure correlation.
**little vs small**: Interpretation

- Attributive much more common for both
  - Predicative relatively more common in conversation
  - Predicative relatively more common for *small* than *little*

- Collocation results:
  - *little*: concrete objects (*little boy*)
  - *small*: amounts (*small proportion*)

- But predicative *small* also for physical size:
  - *She’s small and really skinny*
  - *He’s really small isn’t he?*

- We still don’t really know why 😊
corpus linguistics gives us the *what*, but not the *why*
Where do we go from here?

- Corpora show clearly that even very similar words can show different behavior.

- But they still don’t explain why
  - Hand correction limits data sizes
  - Without semantic tags, we can’t generalize automatically

- Corpora with more mark-up (syntax and semantics) would help
  - But they are expensive, . . .
Case Study: Pronouns
Possessive Pronouns in Japanese contrasted with English

➤ Introduction

➤ Possessive Expressions in Japanese and English

(1) Kanji: 私は 舌を 噛んだ
Jap: watashi-wa shita-wo kanda
Gloss: I-TOP tongue-ACC bit
Eng: ‘I bit my tongue’

➤ Differences in Noun Phrase Structure

➤ Pragmatic Analysis

➤ Application to Machine Translation
   Proposed method for generating possessive pronouns
   Experimental Results

➤ Conclusion
Introduction

Possessive expressions

Possessive determinatives are often used as determiners in English when no equivalent would be used in a Japanese sentence with the same meaning.

Larger Problem

Japanese has no syntactic equivalent to determiners in English, no articles, and noun phrases are normally not marked for number.

Under-specified elements need to be deduced!
## Corpus-based Study of Distribution

<table>
<thead>
<tr>
<th>Type:</th>
<th>MT Test set</th>
<th>News reports</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No.</td>
<td>%</td>
</tr>
<tr>
<td>I English Idiomatic Possessive</td>
<td>105</td>
<td>16%</td>
</tr>
<tr>
<td>II Possessive Expression in Japanese</td>
<td>193</td>
<td>30%</td>
</tr>
<tr>
<td>III No Possessive in Japanese</td>
<td>359</td>
<td>54%</td>
</tr>
<tr>
<td>Total:</td>
<td>657</td>
<td></td>
</tr>
</tbody>
</table>

➢ Two Corpora

➢ NTT MT Test set (6,200 sentences, 15,000 NPs)
➢ Nikkei News Reports (1,382 sentences, 8000 NPs)

➢ Matched English:

Then hand checked Japanese for translation (on paper with colored pens!)

Case Studies: Pronouns and Classifiers
Examples

Type I: English Idiomatic Possessive (16%–19%)

(2) Kanji: 彼女は 知恵を 絞った。
Jap: kanojo-wa chie-wo shibotta
Gloss: she-TOP knowledge-ACC squeezed
Eng: ‘She racked her brains’

Type II: Possessive expression in Japanese (30%–3%)

(3) MT test set is not a corpus of natural text

Kanji: 彼女は 彼の 顔を 見た。
Jap: kanojo-wa kare-no kao-wo mita
Gloss: she-TOP he-ADN face-ACC saw
Eng: ‘She saw his face’
Type III: No possessive expression in Japanese (54%–78%)

(4) Kanji: 彼女は財布をなくした。
Jap:  kanojo-wa saifu-wo nakushita
Gloss: she-TOP wallet-ACC lost
Eng: ‘She lost her wallet’

(5) NTT はメンバーネットの名処で今年2月から常に
NTT-wa ‘member-netto’-no meesho-de kotoshi nigatsu-kara tsune-ni
NTT-TOP ‘member-net’-ADN name-by this year February-from already
サービスを開始している
sa–bisu-o kaishi-shite-iru
service-ACC start-is

“NTT began its VPN services in February.”
Distribution of possessives in English

- Possessive determinatives used relatively frequently
  — of POSSESSIVE PRONOUN rare

- Generally not used after verbs of possession or acquisition, except for emphasis
  \( I \text{ have a car} \) vs \( I \text{ have my car} \)

- Typically referential use, not generic or ascriptive

  In particular, words which denote work, body parts, personal possessions, attributes and relational nouns such as kin and people defined by their relation to another person (such as assailant, subordinate) are often modified by possessive determinatives in English.
Distribution of possessives in Japanese

- Normally only if ‘possessor’ is not subject

(5) `watashi-wa saifu-o otoshita`
    I-TOP wallet-ACC dropped
    *I dropped my wallet*

(6) `watashi-wa jibun-no saifu-o otoshita`
    I-TOP self-ADN wallet-ACC dropped
    *I dropped my own wallet*

(7) `watashi-wa kare-no saifu-o otoshita`
    I-TOP he-ADN wallet-ACC dropped
    *I dropped his wallet*

- Use of any pronouns is rare
  All 5 uses in the newspaper corpus are common nouns (pronominalized in translation)
Two examples:

(8) インドネシア政府は三千五億ドルの資金をインフラ整備として投入する計画だ。

*Indonesia* is planning to invest *300.5 billion dollars* to expand *its infrastructure*
ムバラク大統領の来日時に表明する考えだ。

"the decision will be conveyed to President Muhammad Hosni Mubarak during his visit to Tokyo"
English NP Structure

1. NP → Det (Mod)* Noun (Det is specifier)
2. Possessive determinative functions as central determiner
3. Unique
4. Contrasts with a closed set (+ integers):
   - **articles**  ZERO,  *a/an*,  *some*,  *the*,  NULL
   - **possessive phrases**  e.g.  *the man’s*
   - **demonstratives**  *this*,  *these*,  *that*,  *those*
   - **pronouns**  *we*,  *you*,  *us*
   - **quantifiers**  *each*,  *enough*,  *much*,  *more*,  *most*,  *less*,  *a few*,  *a little* . . .
   - **wh-words**  *which*,  *what* (interrogative or relative)
   - **determinatives**  *some*,  *any*,  *no*,  *either*,  *neither*,  *another*

NP’s headed by count nouns must have an article
Japanese NP Structure

1. NP → (Mod)* Noun (no specifier)

2. Possessive expression functions as modifier

3. Can be multiple modifiers: (rare)

   (10)  watashi-no kono hon
         me-ADN this book
         Lit: “my this book”

4. Is a member of an open set, including:

   none (most common)
   genitive noun phrases  Tarou-no “Taro’s”, nihon-no “Japanese” . . .
   demonstratives  kono “this”, sono “that”, that over there “ano"
   quantifiers  koko-no “each”, kaku “each” . . .
   wh-words  dono “which, what” . . .
Analysis

Explain the differences with Grice’s Conversational Maxims.

➢ **The Maximum of Quantity:**
   (i) make your contribution as informative as is required for the current purposes of the exchange
   (ii) do not make your contribution more informative than is required

➢ **The Maximum of Relevance:**
   Make your contributions relevant

   The kind of information encoded by determinatives such as quantifiers and demonstratives is generally encoded in both Japanese and English. The Maxim of Relevance requires its presence if relevant.
1. Possessive determinative contrasts with articles — equivalent effort

2. Use of indefinite article implicates not owned — unless ‘possession’ predicated by verb

3. Use of definite article implicates more restricted reference

4. ⇒ Use possessive determinative if relevant — unless ‘possession’ predicated by verb (don’t be more informative than is required)
Japanese:

1. Possessive expression requires extra effort

2. Don’t use by default
   — interpretation is that subject is antecedent

3. ⇒ Use possessive expression to contradict default

4. ⇒ Use possessive expression to emphasize default
A complicated example

The word 経常利益 keijourieki “pretax profit” appeared 29 times. In Japanese it was only pre-modified by time expressions (12 times).

The English equivalents were more varied:

<table>
<thead>
<tr>
<th>Det</th>
<th>Freq</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>φ</td>
<td>12</td>
<td>Prepositional phrase</td>
</tr>
<tr>
<td>φ</td>
<td>4</td>
<td>Direct Object (3 x post, 1 x expect)</td>
</tr>
<tr>
<td>φ</td>
<td>4</td>
<td>Subject</td>
</tr>
<tr>
<td>its</td>
<td>1</td>
<td>Subject</td>
</tr>
<tr>
<td>its</td>
<td>4</td>
<td>COMPANY said/announced that its . . .</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>A one billion yen pretax profit</td>
</tr>
<tr>
<td>both</td>
<td>1</td>
<td>very free translation</td>
</tr>
<tr>
<td>Toyobo’s</td>
<td>1</td>
<td>Subject (Toyobo from other sentence)</td>
</tr>
<tr>
<td>their</td>
<td>1</td>
<td>Direct Object of (post)</td>
</tr>
</tbody>
</table>

Subject is many companies
A complicated example (cont)

(11) \textit{COMPANY}_i \textit{ announced Wednesday it}_i \textit{ has posted } \phi_i \textit{ pretax profits of ...}

(12) \textit{COMPANY}_i \textit{ announced Tuesday that } \textit{its}_i \textit{ pretax profit rose ...}

(13) \textit{COMPANY\textquoteright}s 11 [...] subsidiaries\textsubscript{i} \textit{ are expected to post } \textit{their}_i \textit{ first-ever combined pretax profits of ...}

(14) \textit{COMPANY}_i \textit{ will post a rise of 6\% in } \phi_i \textit{ pretax profits ...}

(15) \textit{COMPANY}_i \textit{ will post 28 billion yen in } \phi_i \textit{ pretax profits ...}

The direct object of \textit{post} implies ‘possession’ by its subject, the direct object of \textit{announce} doesn’t. But what about the PPs?

Should we put this in the lexicon?
Application to Machine Translation

- Mark nouns that head English noun phrases with possessive determinatives where there is no possessive expression in the Japanese in the lexicon (possessed-nouns)
  - 205 different possessed-nouns (MT test set)
  - heading 825 noun phrases
  - 359 (44%) translated with possessive pronouns

- Mainly nouns that denote **kin, body parts, work, personal possessions, attributes and people defined by their relation to another person**

- Which nouns need to be marked is language specific, and probably register and domain specific as well.

Taking what we have learned and using it to make predictions.
Translating NPs headed by possessed-nouns

1. A noun phrase that fulfills all of the following conditions will be generated with a default possessive determinative with deictic reference determined by the modality of the sentence it appears in*.

(a) The noun phrase is headed by a possessed-noun that denotes *kin* or *body parts*
(b) The noun phrase is the subject of the sentence
(c) The noun phrase is referential
(d) The noun phrase has no other determiner

* First person for declarative, second person for imperative or interrogative.
2. A noun phrase that fulfills all of the following conditions will be generated with a default possessive determinative whose antecedent is the subject of the sentence the noun phrase appears in.

(a) The noun phrase is headed by a **possessed-noun**
(b) The noun phrase is not the subject of the sentence
(c) The noun phrase is referential
(d) The noun phrase has no other determiner
(e) The noun phrase is not the direct object of a verb of **possession** or **acquisition**
Effects of noun phrase referentiality

Only for **Referential NPs:**

(16) Kanji: 鼻が かゆい。
Jap:  *hana-ga* *kayui*
Gloss: nose-NOM itch
Eng: ‘*My* nose itches’
MT-93  *A* nose itches
MT-94  *My* nose itches
Not for Generic NPs:

(17)  
Kanji: 鼻は 感覚器官 だ。
Jap:  hana-wa kankakukikan da
Gloss: nose-TOP sensory organ is
Eng:  ‘The nose is a sensory organ’
MT-93: A nose is a sensory organ
MT-94: φ Noses are sensory organs
If a noun phrase headed by a possessed-noun is the direct object of a verb of possession or acquisition then do not generate a possessive pronoun.

(18) Kanji: 車を持って いますか。
Jap: kuruma-wo motteimasu-ka
Gloss: car-OBJ have-Q
Eng: ‘Do you have a car?’
MT-93: Do you have a car?
MT-94: Do you have your car?
MT-94: Do you have a car?
Results of the generation of all noun phrases headed by **possessed-nouns** in the MT test set (Total 752 noun phrases).

<table>
<thead>
<tr>
<th>Result</th>
<th>Not generated</th>
<th>Generated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>I hit him in <em>the</em> face</td>
<td>I hid <em>my</em> face</td>
</tr>
<tr>
<td>Bad</td>
<td>I scratched <em>a</em> face</td>
<td>I lost <em>my</em> face</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Result</th>
<th>Possessive determinative</th>
<th>MT-93</th>
<th>MT-94</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NPs</td>
<td>%</td>
<td>NPs</td>
</tr>
<tr>
<td>Good</td>
<td>Not generated</td>
<td>429</td>
<td>57%</td>
</tr>
<tr>
<td></td>
<td>Generated</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>429</td>
<td>57%</td>
</tr>
<tr>
<td>Bad</td>
<td>Not generated</td>
<td>323</td>
<td>43%</td>
</tr>
<tr>
<td></td>
<td>Generated</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>323</td>
<td>43%</td>
</tr>
</tbody>
</table>
Over All Results

323 NPs required possessive determinatives
  Appropriately generated: 263
  Inappropriately generated: 83

<table>
<thead>
<tr>
<th></th>
<th>MT-93</th>
<th>MT-94</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>57%</td>
<td>81%</td>
</tr>
<tr>
<td>Precision</td>
<td>—</td>
<td>88%</td>
</tr>
</tbody>
</table>

Improve accuracy by:
  improving parsing and transfer stages
correctly identifying all possessess-nouns (use parsed aligned corpora)

Improve precision by:
  improving determination of referentiality
  add explicit semantic constraints:
  only for possessess-nouns that denote clothes if the antecedent is human
Conclusions

1. Possessive determinatives are used in English even when there is no equivalent possessive expression used in Japanese.

2. This can be explained by the fact that in English possessive determinatives function as determiners, while in Japanese the possessive construction is an optional modifier phrase.

3. ‘possessed-nouns’ can be identified in English that act (imperfectly) as cues.

4. Implementing an algorithm that uses possessed-nouns in the Japanese-to-English MT system ALT-J/E generated possessive pronouns with an accuracy of 81% (up from 57%) and precision of 88%.

5. Should also be applicable to other under-specified generation: AAC.
1. Satoru Ikehara calls our approach meaning analysis as opposed to meaning understanding. We attempt to solve problems, even if not perfectly, by stepwise refinement.

2. Generally, a brute-force approach of adding information to the lexicon (which may mean checking 70,000+ common nouns . . .) and adding new rules takes 3-6 months and gets an 80% solution.

3. I did this once for number/countability and articles (which took three years), then possessive pronouns, and then numeral classifiers.

4. By this stage, determiners and number were good enough that problems with prepositions and tense/aspect became more pressing.

5. The hope is that any work done will still be useful in the next version/refinement of the problem: this has proved to be the case so far.
Conclusions

1. Possessive pronouns are used in English even when there is no equivalent possessive expression used in Japanese

2. ‘possessed-nouns’ can be identified in English that act as cues

3. An algorithm is proposed that uses possessed-nouns to appropriately generate possessive pronouns in a Japanese-to-English MT system

4. Implementing the algorithm in ALT-J/E generated possessive pronouns with an accuracy of 81% (↑ 57%) and precision of 88%.
Annotation of Pronouns in a Multilingual Corpus of Mandarin Chinese, English and Japanese
Motivation and Overview

- Attempting to model lexical and structural semantics
  - For multiple languages — identify cross-lingual differences
  - Exploit them to learn meaning (make the implicit explicit)

- Started by annotating content words (with wordnets)

- But nouns were often translated as pronouns — so tag them
  1. Identify pronouns used in the corpus
  2. Analyze in terms of components — aids matching
    - Extended wordnet gives full decompositional analysis
  3. Annotate the pronouns monolingually in each language
    - Link to extended wordnet for analysis
  4. Annotate their correspondences across languages
  5. Analyze the distribution cross-lingually
Examined words tagged as pronouns in (Mandarin) Chinese, English, Japanese (and later Indonesian) parts of the NTU Multilingual Corpus (NTU-MC) — used the POS tags

Different tag-sets identified quite different collections

We took the union, and filled in missing entries by hand

also referred to reference grammars
not complete, but getting there

117 different types; 249 tokens: Chinese 57
English 68
Indonesian 40 (in progress)
Japanese 84

numbers out of date
We include related determiners (demonstratives and quantifiers)
### Components

<table>
<thead>
<tr>
<th>Head</th>
<th>Person</th>
<th>Number</th>
<th>Gender</th>
<th>Case</th>
<th>Q/Type</th>
<th>Formality</th>
<th>Proximity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantifier</td>
<td>First</td>
<td>Dual</td>
<td>Feminine</td>
<td>Objective</td>
<td>Assertive</td>
<td>Formal</td>
<td>Proximal</td>
</tr>
<tr>
<td>Entity</td>
<td>First (I)</td>
<td>Plural</td>
<td>Masculine</td>
<td>Possessive</td>
<td>Elective</td>
<td>Formal</td>
<td>Distal</td>
</tr>
<tr>
<td>Time</td>
<td>First (E)</td>
<td>Singular</td>
<td>Neuter</td>
<td>Subjective</td>
<td>Negative</td>
<td>Informal</td>
<td>Medial</td>
</tr>
<tr>
<td>Manner</td>
<td>Second</td>
<td></td>
<td></td>
<td></td>
<td>Other</td>
<td>Politeness</td>
<td>Remote</td>
</tr>
<tr>
<td>Person</td>
<td>Third</td>
<td></td>
<td></td>
<td></td>
<td>Reciprocal</td>
<td>Polite</td>
<td></td>
</tr>
<tr>
<td>Place</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Universal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reason</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Interrogative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reflexive</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Similative** are treated as +Manner, +Proximity
Analyzing Pronouns Mono-lingually

➢ Decompose into:

➢ head (HYPONYM)
➢ quantifier (QUANTIFIER: new relation)
➢ features (DOMAIN-USAGE)

➢ Also mark as INSTANCE of pronoun$_{n:1}$ or its hyponyms

➢ E.g. there$_{n:1}$: HYPONYM location$_{n:1}$; DOMAIN-USAGE distal$_{a:1}$; INSTANCE demonstrative pronoun$_{n:1}$
## Components: Place

<table>
<thead>
<tr>
<th>Head</th>
<th>Type/Proximity</th>
<th>English</th>
<th>Japanese</th>
<th>Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Place</td>
<td>Interrogative</td>
<td><em>where</em></td>
<td>何処, どこ <em>doko</em></td>
<td>哪里 <em>nǎlǐ</em></td>
</tr>
<tr>
<td></td>
<td>Proximal</td>
<td><em>here</em></td>
<td>此処, ここ <em>koko</em></td>
<td>这里 <em>zhèlǐ</em></td>
</tr>
<tr>
<td></td>
<td>Distal</td>
<td><em>there</em></td>
<td></td>
<td>那里 <em>nàlǐ</em></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>其処, そこ <em>soko</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Medial</td>
<td></td>
<td>彼処, あそこ <em>asoko</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Remote</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Universal</td>
<td><em>everywhere</em></td>
<td>どこも <em>doko mo</em></td>
<td>到处 <em>dàochù</em></td>
</tr>
<tr>
<td></td>
<td>Existential</td>
<td></td>
<td>どこか <em>doko ka</em></td>
<td>某处 <em>mǒuchù</em></td>
</tr>
<tr>
<td></td>
<td>Assertive</td>
<td><em>somewhere</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Elective</td>
<td><em>anywhere</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td><em>elsewhere</em></td>
<td>よそ <em>yoso</em></td>
<td>别处 <em>biéchù</em></td>
</tr>
</tbody>
</table>

Not all lemmas shown

Already refined 45
Tagging Pronouns Mono-lingually

- Tagged one document by hand *The Adventure of the Speckled Band*

<table>
<thead>
<tr>
<th>Language</th>
<th>English</th>
<th>Chinese</th>
<th>Japanese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contentful</td>
<td>1,370</td>
<td>1,177</td>
<td>463</td>
</tr>
<tr>
<td>Other</td>
<td>75</td>
<td>19</td>
<td>51</td>
</tr>
<tr>
<td>Total</td>
<td>1,445</td>
<td>1,196</td>
<td>514</td>
</tr>
<tr>
<td>Sentences</td>
<td>599</td>
<td>620</td>
<td>702</td>
</tr>
<tr>
<td>Words</td>
<td>11,628</td>
<td>12,433</td>
<td>13,902</td>
</tr>
</tbody>
</table>

- Distinguished existential *there* (but not dummy *it*) with POS tags

- *other* includes relative pronouns, dummy *it*, idioms and segmentation errors
Tagging and Analyzing Pronouns Cross-lingually

- Automatically linked by matching features

- Hand corrected:

<table>
<thead>
<tr>
<th># Matching Features</th>
<th>Linked Pronouns</th>
<th>Non-linked Pronouns</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td># Chinese</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td># Japanese</td>
<td>15</td>
<td>120</td>
</tr>
</tbody>
</table>

- Case and politeness mismatches common

- A surprising number of non-linked pronouns in Chinese and Japanese
Interesting Cross-Linguistics Differences

(19) She\textsubscript{i} shot him\textsubscript{j} and then herself\textsubscript{i};

a. 妻-さん が 旦那-さん を 撃って 、 それから 自分 も 撃った

Wife\textsubscript{i} shot husband\textsubscript{j} and then shot self\textsubscript{i} too

b. 她 拿 枪 先 打 夫 , 然后 打 自己

She\textsubscript{i} took the gun to first shoot husband\textsubscript{j}, and then shot self\textsubscript{i}
(20) [many (cases) strange] . . . but none commonplace . . .

a. 但 是 却 没 有 一 例 是 平 淡 无 奇 的
Dan4shi4 que4 mei2you3 yi1li4 shi4 ping2dan4wu2qi2 de
‘But, there is not one case that is featureless.’

b. ど れ も 尊 常 で は な い 事 件 で ある
Dore mo jinjode wa nai jiken dearu
‘Everything is a case which is not usual.’

(21) It is a swamp adder!

a. 这 是 一 条 沼 地 蟒 蛇 ！
Zhe4 shi4 yi1tiao2 zhao3di4 kui2she2 ！
‘This is a swamp adder!’

b. 沼 蛇 だ ！
numahebi da ！
‘φ is a swamp snake’
Discussion

➢ A new way of annotation that links wordnets to corpora

➢ Unresolved issues (possible ideas for project 2)
   ➢ Further analysis of unlinked pronouns: which and why? In particular how and why are Japanese and Chinese different?
   ➢ Tag more corpora (ongoing); Extend to more languages;
   ➢ Integrate to HPSGs: ERG, Jacy, MCG, IndoGram
Classifiers
How do we count Email in Japanese?

- Japanese has two classifiers for counting messages:
  - 通 tsuu: used for letters
  - 件 ken: used for incidents
  - 本 hon: used for phone calls

- See how they are used to count Email and SMS


Change with familiarity

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Email Usage</td>
<td>5%</td>
<td>11%</td>
<td>34%</td>
<td>81%</td>
<td>86%</td>
</tr>
<tr>
<td>Classifier</td>
<td>通, 本</td>
<td>通</td>
<td>通</td>
<td>通, 件</td>
<td>通, 件</td>
</tr>
<tr>
<td>SMS Usage</td>
<td>—</td>
<td>39%</td>
<td>45%</td>
<td>67%</td>
<td>76%</td>
</tr>
<tr>
<td>Classifier</td>
<td>通, コール</td>
<td>通</td>
<td>通</td>
<td>件, 通</td>
<td></td>
</tr>
</tbody>
</table>

Change in Classifier use with increased familiarity
Classifiers listed in frequency order

通 has more of a one-way feeling, while 件 is more of a conversation.

Sometimes depends on the tool (which classifier does it use).
Conclusions

➢ Different questions require different resources

➢ Good corpora are useful for multiple tasks
More SQL

➤ Find animals

```sql
select * from synset where synset in
(select synset from xlink where resource='lexnames' and xref=5)
limit 125
```

➤ Find sentences with animals

```sql
attach 'eng.db' as as 'e';
select sent from e.sent join e.concept
on e.sent.sid=e.concept.sid
where tag in (
select synset from synset where synset in
(select synset from xlink where resource='lexnames' and xref=5)
limit 50
```
Another way (without duplicate sentences)

```sql
select sent from e.sent
where sid in
  (select distinct sid from concept
  where tag in
    (select synset
     from synset
     where synset in
       (select synset from xlink
        where resource='lexnames' and xref=5)))
limit 50
```