Using Information Structure to Improve Transfer-based MT

Sanghoun Song (UW)
Emily M. Bender (UW)

HPSG2011
Workshop on Information Structure and Formal Grammar
08-23-2011

Outline

Hypothesis

- The essential part of translation is reshaping the means of conveying information, instead of simply changing the words or reordering phrases.

Working Hypothesis

- Transfer-based MT systems can be improved
  - by encoding information structure in both the source and target grammars
  - by preserving information structure in the transfer stage

Goal

- How information structure can be represented in HPSG/MRS
- How information structure can help refine multilingual MT
- A sample translation between English & Japanese/Korean

Outline

Allosentences

- close paraphrases which share truth conditions (Lambrecht, 1996)
- They are not always in the same felicity conditions.

(1) a. I am Kim.

b. watashi-ga/wa Kim desu.
I-NOM/TOP Kim COP [jpn]

(2) Q: Who are you?

A: watashi-#ga/wa Kim desu.

(3) Q: Who is Kim?

A: watashi-ga/#wa Kim desu.

Active/Passive Pairs

(4) a. Kim tore the book.

b. The book was torn by Kim.

(5) a. Kim-ga sono hon-o yabut-ta.
Kim-NOM DET book-ACC tear-PST

'Kim tore the book.'

b. ?sono hon-ga Kim-ni yabu-rare-ta.
DET book-NOM Kim-DAT tear-PASS-PST

'The book was torn by Kim.' [jpn]

Other Issues

- (4b) has at least eight allosentences in Japanese (2×2×2)
  - actives *vs.* passives
  - case markers (e.g. *ga*) *vs.* topic markers (e.g. *wa*)
  - scrambling: SOV *vs.* OSV (Choi, 1999; Ishihara, 2001)

(6) a. Kim-ga/wa sono hon-o/wa yabut-ta.
Kim-NOM/TOP DET book-ACC/TOP tear-PST

b. sono hon-o/wa Kim-ga/wa yabut-ta.
DET book-ACC/TOP Kim-NOM/TOP tear-PST

- two more options: 32 allosentences, in total (8×2×2)
  - pro-drop
  - null NP markings
- not felicitous in the same contexts.

Information Structure

- In orde to solve the mismatch and refine the translations, we need to use Information Structure.
  - The difference in felicity conditions between allosentences is the subject of study of information structure.
  - Information structure is hypothesized to be universal.
  - All languages have some way to mark topics and foci.
    - pitch accent (e.g. A/B-accent)
    - specific word order (e.g. scrambling)
    - morphological marking (e.g. *wa*)
    - some combination of them
- This study looks at the particular case of translating English passives into Japanese/Korean.

Outline

Basic Assumptions

1. Sentences always have at least one focus, but they do not always have a topic; further, constituents may be 'background' (i.e. neither topic nor focus).
2. 'contrast' as a cross-cutting information structure category (Molnár, 2002)

(7) Kim thì di Hanoi
Kim CT go Hanoi

'Kim goes to Hanoi(, but nobody else).' [vie]

Basic Assumptions (cont'd)

3. Semantically empty categories (e.g. complementizers, expletives) are informatively empty as well (i.e. assigned no category).

   (8) a. ~~It is~~ Kim ~~who~~ tore the book.

   b. The book ~~was~~ torn ~~by~~ Kim.

4. It is assumed that the canonical position of topics is sentence-initial at least in our sample of languages (English, Japanese, and Korean).

Information Structure in HPSG/MRS

- INFO-STR: a semantic feature in the MRS
- MKG: a categorical feature encoding the lexical marking
- SFORM: a set of constraints on phrase structure rules.

MRS: *info-str*

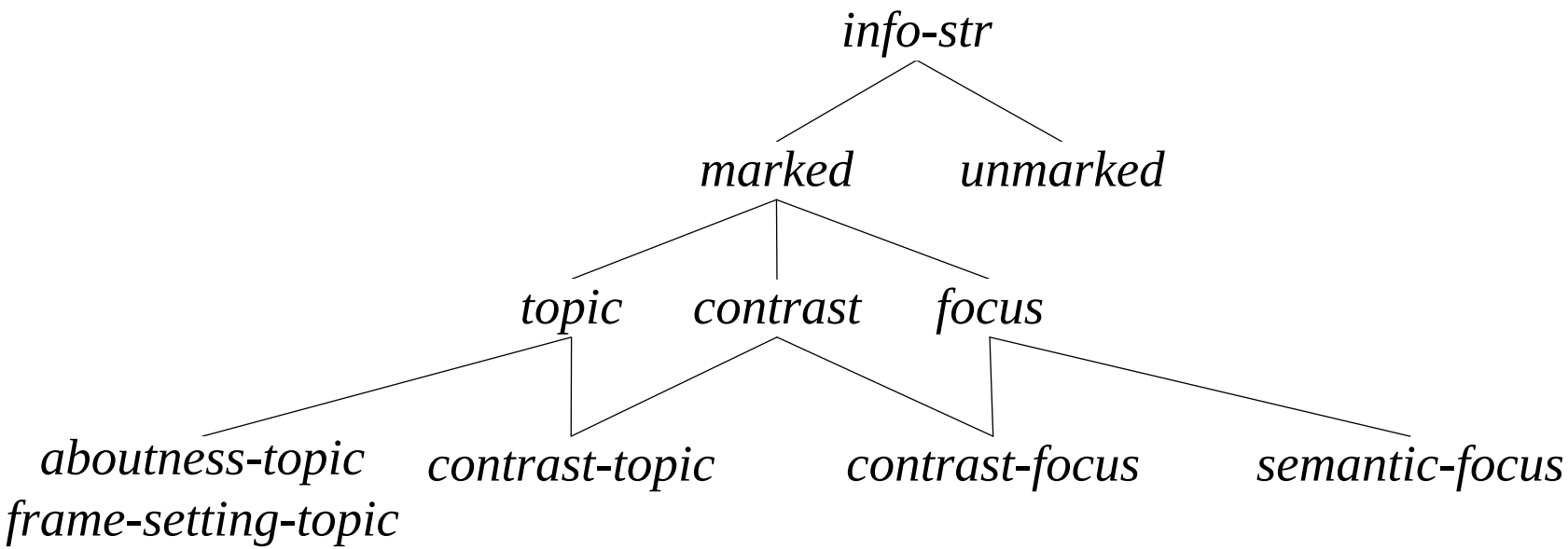


Figure: Type Hierarchy of *info-str*

- representing information structure with a feature on indices directly in the MRS

Marking: *mkg*

$$\begin{bmatrix} \text{MKG} & \begin{bmatrix} \text{TP} & \textit{bool} \\ \text{FC} & \textit{bool} \end{bmatrix} \end{bmatrix}$$

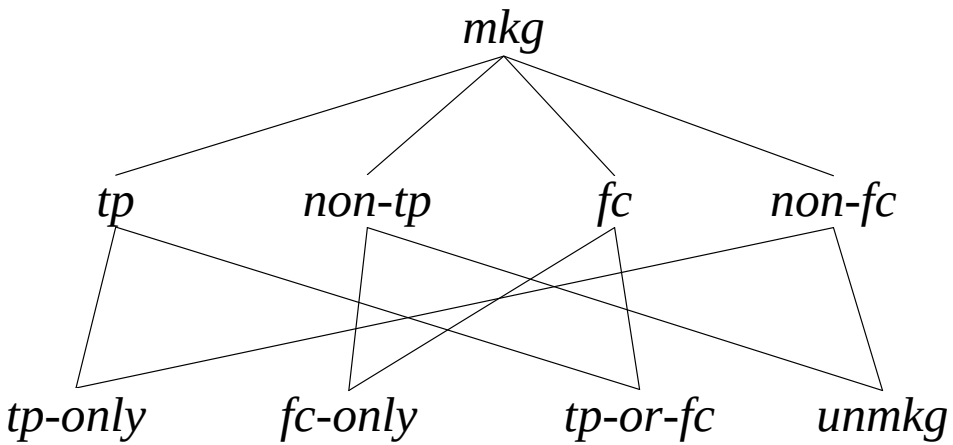


Figure: Type Hierarchy of *mkg*

Sentential Forms: *sform*

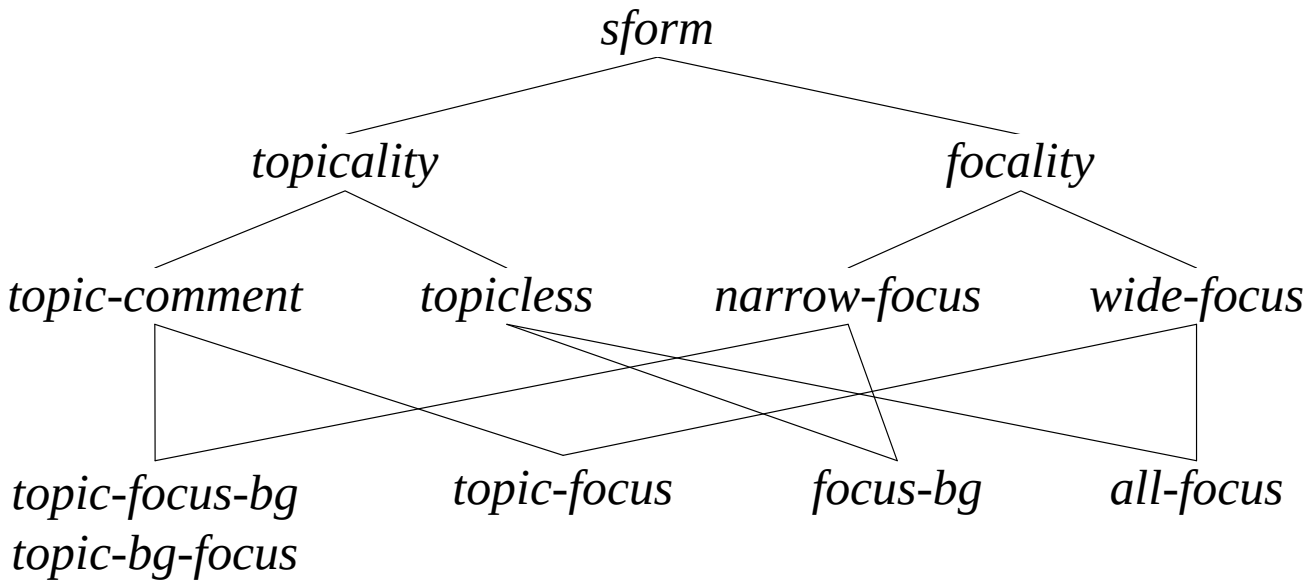


Figure: Type Hierarchy of *sform*: adapted from Paggio (2009), etc.

Sentential Forms: *sform* (cont'd)

- *topicality*
  - *topic-comment*

    (9) As for **the book**, KIM tore it.
  - *topicless*

    (10) ~~It is~~ KIM ~~who~~ tore the book.

Sentential Forms: *sform* (cont'd)

- *focality*
  - *narrow-focus*
  - *wide-focus*

(11) a. ínam-**é** yimi
boy-F came

'THE BOY came.'

b. ínam **á**-yimi
boy F-came

'The boy CAME.' (Rendille [red] (Lecarme, 1999))

Sentential Forms: *sform* (cont'd)

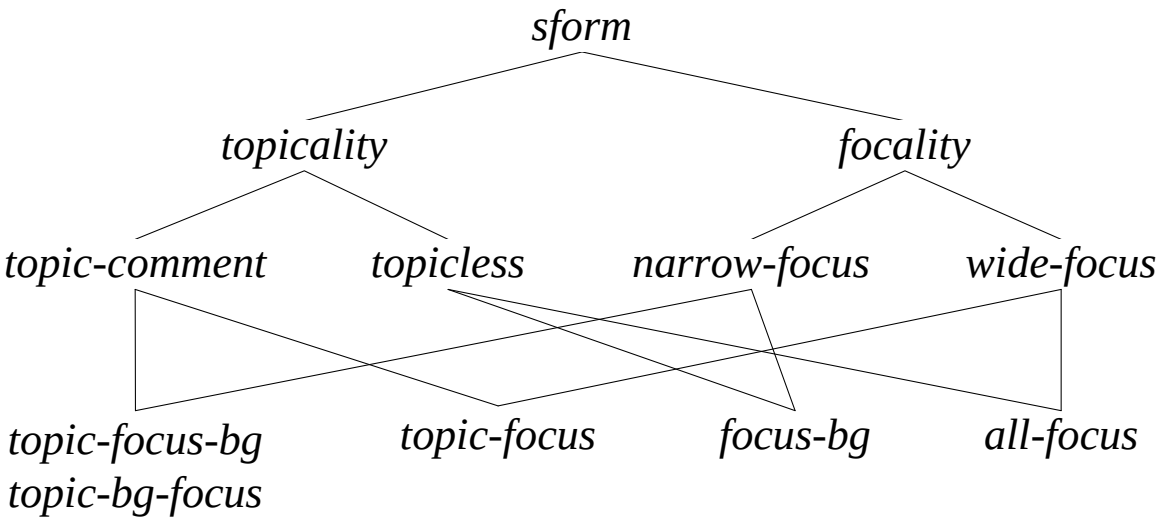


Figure: Type Hierarchy of *sform*

(12) a. **The book** was torn by [$_f$ KIM] (*topic-bg-focus*).
b. **The book** [$_f$ was torn] by Kim (*topic-focus-bg*).
c. **The book** [$_f$ was torn by Kim] (*topic-focus*).
d. [$_f$ THE BOOK] was torn by Kim (*focus-bg*).
e. [$_f$ The book was torn by Kim] (*all-focus*).

Outline

Information Structure in English

- Information structure is normally constrained by pitch accents.
  - A-accented phrases (H*) take *focus*

$$\textit{fp-lex-rule} \rightarrow \begin{bmatrix} \text{PROSODY} & \textsc{a}\textit{-accent} \\ \text{INFO-STR} & \textit{focus} \end{bmatrix}$$

  - B-accented phrases (L+H*) take *topic*

$$\textit{tp-lex-rule} \rightarrow \begin{bmatrix} \text{PROSODY} & \textsc{b}\textit{-accent} \\ \text{INFO-STR} & \textit{topic} \end{bmatrix}$$

Information Structure in Japanese/Korean

- topic markers: aboutness *vs.* contrastiveness

(13) Q: Who is a student?

A: Kim-ga/wa gakusei desu.
Kim-NOM/TOP student COP
'Kim is a student.' [jpn]

Information Structure in Japanese/Korean (cont'd)

- scrambling

(14) a. Kim-wa sono hon-o yabut-ta.
Kim-TOP DET book-ACC tear-PST
(topic)

b. sono hon-o Kim-wa yabut-ta.
DET book-ACC Kim-TOP tear-PST
(contrastive-focus)

c. Kim-ga sono hon-wa yabut-ta.
Kim-NOM DET book-TOP tear-PST
(contrastive-focus)

d. sono hon-wa Kim-ga yabut-ta.
DET book-TOP Kim-NOM tear-PST
(contrastive-topic) [jpn]

Information Structure in Japanese/Korean (cont'd)

Table: IS of topic-marked NP (adapted from Choi (1999))

	in-situ	**scrambling**
subject	*topic*	*contrast-focus*
non-subject	*contrast-focus*	*contrast-topic*

$$nom\text{-}marker \rightarrow \begin{bmatrix} \text{ORTH} \left\langle ga \right\rangle \\ \text{MKG } unmkg \\ \text{CASE } nom \end{bmatrix} \qquad topic\text{-}marker \rightarrow \begin{bmatrix} \text{ORTH} \left\langle wa \right\rangle \\ \text{MKG } tp \\ \text{CASE } case \end{bmatrix}$$

Information Structure in Japanese/Korean (cont'd)

$$\begin{bmatrix} \textit{topic-comment} & \\ \text{MKG} & \textit{tp} \\ \text{HD} \mid \text{MKG} & \textit{fc} \\ \text{NON-HD} \mid \text{MKG} & \textit{tp} \end{bmatrix}$$

- INFO-STR in Japanese and Korean is specified at the phrasal level (i.e. each grammatical rule, such as *subj-head* and *comp-head*) unlike English

(15) sono hon-wa Kim-ga yabut-ta.
DET book-TOP Kim-NOM tear-PST
'**The book** was torn by [$_f$ KIM].' [jpn]

$$\begin{bmatrix} \textit{top-scr-subj-head} \\ \text{HD} \mid \text{VAL} \mid \text{COMPS} \left\langle [] \right\rangle \\ \text{NON-HD} \mid \text{INFO-STR} \ \textit{contrast-focus} \end{bmatrix} \qquad \begin{bmatrix} \textit{top-scr-comp-head} \\ \text{HD} \mid \text{VAL} \mid \text{COMPS} \left\langle \right\rangle \\ \text{NON-HD} \mid \text{INFO-STR} \ \textit{contrast-topic} \end{bmatrix}$$

Outline

A Sample Translation

The book was torn by Kim.

1. Kim-ga sono hon-o yabut-ta.
2. sono hon-o Kim-ga yabut-ta.
3. Kim-wa sono hon-o yabut-ta.
4. sono hon-o Kim-wa yabut-ta.
5. Kim-ga sono hon-wa yabut-ta.
6. sono hon-wa Kim-ga yabut-ta.
7. Kim-wa sono hon-wa yabut-ta.
8. sono hon-wa Kim-wa yabut-ta.

A Sample Translation (cont'd)

The book was torn by [$_f$ KIM].

1. Kim-ga sono hon-o yabut-ta.
2. sono hon-o Kim-ga yabut-ta.
3. Kim-wa sono hon-o yabut-ta.
4. sono hon-o Kim-wa yabut-ta.
5. Kim-ga sono hon-wa yabut-ta.
6. sono hon-wa Kim-ga yabut-ta.
7. Kim-wa sono hon-wa yabut-ta.
8. sono hon-wa Kim-wa yabut-ta.

A Sample Translation (cont'd)

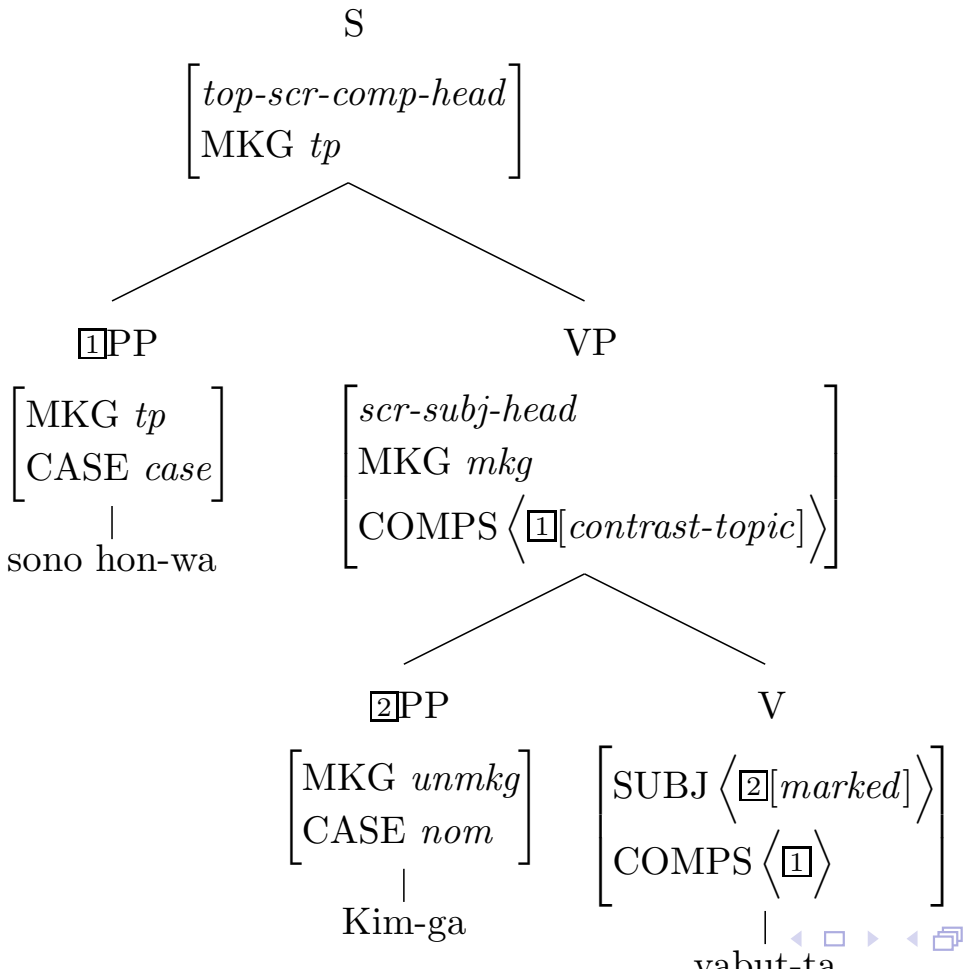
The book was torn by [$_f$ KIM].

~~1. Kim-ga sono hon-o yabut-ta.~~
~~2. sono hon-o Kim-ga yabut-ta.~~
~~3. Kim-wa sono hon-o yabut-ta.~~
~~4. sono hon-o Kim-wa yabut-ta.~~
~~5. Kim-ga sono hon-wa yabut-ta.~~
6. sono hon-wa Kim-ga yabut-ta.
~~7. Kim-wa sono hon-wa yabut-ta.~~
~~8. sono hon-wa Kim-wa yabut-ta.~~

A Sample Translation (cont'd)

(16) a. **The book** was torn by [$_f$ KIM].

b. sono hon-wa Kim-ga yabut-ta.
DET book-TOP Kim-NOM tear-PST



An Input MRS (English)

$$
\left[\begin{array}{ll}
\textit{mrs} & \\
\text{LTOP} & \boxed{h1}\ h \\
\text{INDEX} & \boxed{e2}\ e \\
\text{RELS} & \left\langle \left[\begin{array}{ll}
\textit{exist\_q\_rel} & \\
\text{LBL} & \boxed{h3}\ h \\
\text{ARG0} & \boxed{x4}\left[\begin{array}{ll} x & \\ \text{INFO-STR} & \textit{topic} \\ \text{COG-ST} & \textit{uniq+fam+act} \end{array}\right] \\
\text{RSTR} & \boxed{h5}\ h \\
\text{BODY} & \boxed{h6}\ h
\end{array}\right], \left[\textit{\_book\_n\_rel}\right], \right. \\
 & \left. \left[\begin{array}{ll}
\textit{\_tear\_v\_rel} & \\
\text{LBL} & \boxed{h1} \\
\text{ARG0} & \boxed{e2} \\
\text{ARG1} & \boxed{x8}\left[\begin{array}{ll} x & \\ \text{INFO-STR} & \textit{focus} \\ \text{COG-ST} & \textit{type-id} \end{array}\right] \\
\text{ARG2} & \boxed{x4}
\end{array}\right], \left[\textit{proper\_q\_rel}\right], \left[\begin{array}{ll}
\textit{named\_rel} & \\
\text{LBL} & \boxed{h9}\ h \\
\text{ARG0} & \boxed{x8} \\
\text{CARG} & \text{Kim}
\end{array}\right] \right\rangle \\
\text{HCONS} & \left\langle \left[\textit{qeq}\right], \left[\textit{qeq}\right] \right\rangle
\end{array}\right]
$$

An Output MRS (Japanese)

$$
\begin{bmatrix}
\textit{mrs} & \\
\text{LTOP} & \boxed{h1}\ h \\
\text{INDEX} & \boxed{e2}\ e \\
\text{RELS} & \left\langle
\begin{bmatrix}
\textit{exist\_q\_rel} & \\
\text{LBL} & \boxed{h3}\ h \\
\text{ARG0} & \boxed{x4}\begin{bmatrix} x & \\ \text{INFO-STR} & \textit{contrast-topic} \\ \text{COG-ST} & \textit{uniq+fam+act} \end{bmatrix} \\
\text{RSTR} & \boxed{h5}\ h \\
\text{BODY} & \boxed{h6}\ h
\end{bmatrix},
\begin{bmatrix} \textit{\_book\_n\_rel} \end{bmatrix}, \right. \\
& \left.
\begin{bmatrix} \textit{proper\_q\_rel} \end{bmatrix},
\begin{bmatrix}
\textit{named\_rel} & \\
\text{LBL} & \boxed{h8}\ h \\
\text{ARG0} & \boxed{x9}\begin{bmatrix} x & \\ \text{INFO-STR} & \textit{focus} \\ \text{COG-ST} & \textit{type-id} \end{bmatrix} \\
\text{CARG} & \text{Kim}
\end{bmatrix},
\begin{bmatrix} \textit{\_tear\_v\_rel} \end{bmatrix}
\right\rangle \\
\text{HCONS} & \left\langle \begin{bmatrix} \textit{qeq} \end{bmatrix}, \begin{bmatrix} \textit{qeq} \end{bmatrix} \right\rangle
\end{bmatrix}
$$

Experiment: Translating Passives

- process
  1. constructing toy grammars for English, Japanese, and Korean
  2. adding other rules to produce allosentences
  3. implementing information structure into each grammar
  4. creating the mapping between internal and external features of indices (semi.vpm)

Experiment: Translating Passives (cont'd)

- eight types of allosentences in English, for each of the three verbal types (i.e. 24 input sentences)
  - 'tear'
  - 'chase'
  - 'hit'

Experiment: Translating Passives (cont'd)

- hypothetical suffixes
  - '-FP': A-accent for foci
  - '-TP': B-accent for topic

(17) a. **The book** was torn by [$_f$ KIM].

b. The book-TP was torn by Kim-FP.

Results

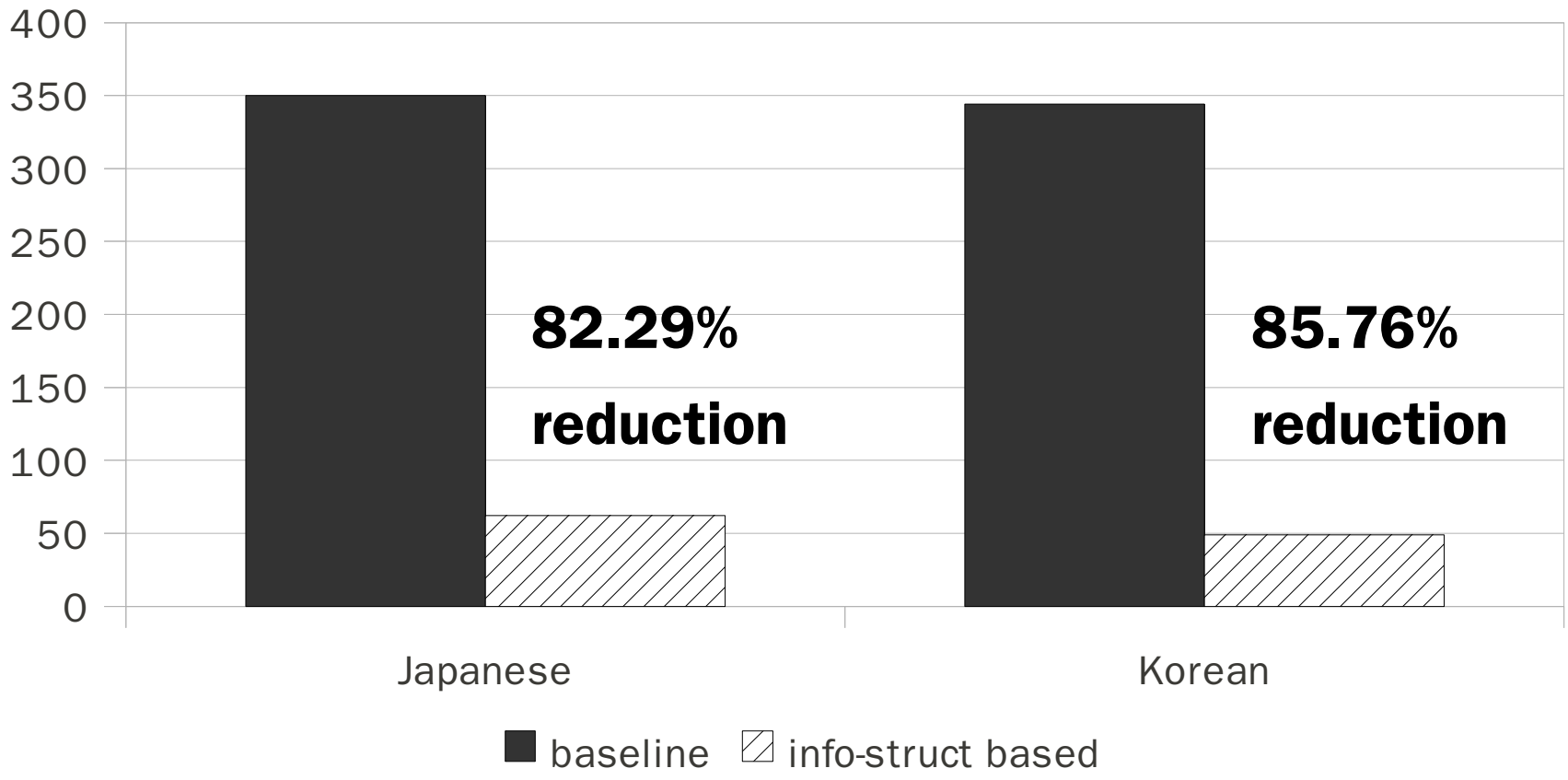


Figure: Evaluation

Contribution

- latency
  - The processing burden of MT component which ranks the translations and select only suitable results can be greatly lightened
- accuracy
  - Though it is still necessary to harness a re-ranking model for choosing translations, we can start from once-refined sets of translations.

Outline

Conclusion

- So far, we have covered
  - how Information Structure can be represented in HPSG/MRS
  - how the representation can be used to refine translations
- Implications
  - Type hierarchies in this proposal are constructed almost language-independently.
  - We effectively move further up the MT pyramid (Vauquois, 1968).

Future Work

- evaluating with various sentences
  - dropped elements
  - clefting
- corpus studies: using multilingual texts
- other languages
  - Japanese/Korean to English
  - Chinese, Spanish, Russian
- a Grammar Matrix library for information structure