

Implementing information structure in HPSG/MRS

DELPH-IN 2010

Sanghoun Song

UW Linguistics

Outline

- 1. Implementation: Translating Passives**
2. A Corpus Study: *The Little Prince*
3. Future Work

Motivation

- (1) a. Kim tore the book.
b. The book was torn by Kim.
- (2) a. Kim-ga sono hon-o yabut-ta.
 Kim-NOM DET book-ACC tear-PAST
 ‘Kim tore the book.’
b. ? sono **hon**-ga Kim-ni yabu-rare-ta.
 DET **book**-NOM Kim-DAT tear-PASS-PAST
 ‘The book was torn by Kim.’
- (3) a. Kim-ga/**wa** sono hon-o/**wa** yabut-ta.
 Kim-NOM/TOP DET book-ACC/TOP tear-
 PAST
b. sono hon-o/**wa** Kim-ga/**wa** yabut-ta.

Translating Passives

- All **passives** are not always translated into **passives**.
 - Passives are **not universal**.
 - With Passives : W/O Passives = **162 : 211** (WALS Info)
 - **Productivity** of passivization
 - It differs in different languages.
- **Information Structure** can be used to refine translations of passives.
 - Active/passive pairs are the typical cases of **allosentences**.

Information Structure

Q: By whom was the book torn?

A: **The book** was torn by KIM.

Topic: **the book** [B-accent]

Focus: KIM [A-accent]

The book was torn by **KIM**.

sono hon-wa Kim-ga yabut-ta
DET book-TOP Kim-NOM tear-PAST

- Using Information Structure can filter out **infelicitous translations**.

Assumptions

- 1) All sentences always have **at least one focus**, while all sentences do not always have a topic (Gundel, 1999).
- 2) **Contrast** exists as a category in information structure, with properties of both topicality and focality (Molnár, 2002).
- 3) Semantically empty categories are **informatively empty** as well.

It is **JOHN** ~~who~~ read this book. (cleft sentences)

This book was torn ~~by~~ Kim. ('by' in passives)

Implementation

- Type Hierarchies
 - *sform*: sentential forms
 - *mkd*: markedness
 - *info-str*: MRS
- Lexical Rules
 - Prosody in English
 - Topic-markers in Japanese, Korean
- Phrasal Rules
 - topic-comment
 - scrambling

Experiment

- Input: 24 English passive sentences
 - 3 verbal types (tear, chase, hit) × 8 allosentences types
- Output: Japanese / Korean sentences
 - Actives or Passives / scrambling
- Toy Grammars
 - The Grammar Matrix customization system
 - Passives / Information Structure

Languages					
	Passive	Animacy	Topic	Contrast	Focus
Eng	Productive	Less important	B-accent	A/B-accent	A-accent
Jpn	Unproductive	Important	wa	wa & scrambling	ga
Kor			(n)un	(n)un & scrambling	ka

Results

- The systems based on Information Structure
 - reduce the numbers of outputs
 - from 350 to 62 (**17.72%**) for Japanese
 - from 344 to 49 (**14.25%**) for Korean.

		E→J / E→K		
		‘tear’	‘chase’	‘hit’
5a	<i>topic-bg-focus</i>	4(2)/4(2)	8(4)/8(4)	8(4)/4(2)
5b	<i>topic-focus-bg</i>	2(0)/2(0)	4(0)/4(0)	4(0)/2(0)
5c	<i>topic-focus</i>	4(2)/4(2)	8(4)/8(4)	8(4)/4(2)
5d	<i>focus-bg</i>	1(0)/1(0)	2(0)/2(0)	2(0)/1(0)
5e	<i>all-focus</i>	2(0)/2(0)	4(0)/4(0)	4(0)/2(0)
5	baseline	16/16	24/32	24/16
6a	<i>topic-focus</i>	2(1)/2(1)	4(1)/4(1)	4(1)/2(1)
6b	<i>focus-bg</i>	1(0)/1(0)	2(0)/2(0)	2(0)/1(0)
6c	<i>all-focus</i>	1(0)/1(0)	2(0)/2(0)	2(0)/1(0)
6	baseline	2/2	4/4	4/2

Information Structure in MT

- Information Structure **improves MT**.
 - Using Information Structure in MT can
 - function as a filter to **exclude infelicitous translations**.
 - light the burden of the **transfer component** (Vauquois, 1968).
 - make the translations **more plausible**.
 - Remaining Problems
 - Resolving Information Structure within contexts

Outline

1. Implementation: Translating Passives

2. A Corpus Study: *The Little Prince*

3. Future Work

Subject vs. Topic

- Classification in Previous Studies
 - Li and Thompson (1976)

	subj-prominence	non-subj-prominence
topic-prominence	Japanese, Korean	Chinese
non-topic-prominence	English	Tagalog

- Huang (1984)

	zero-topic	non-zero-topic
pro-drop	Chinese, Japanese, Korean	Spanish, Italian
non-pro-drop	German	English, French

- Charles (2004)
 - English type / Italian type / Chinese type

The Little Prince



- Multilingual Texts
 - French, **English**, Spanish, German, **Korean**, Japanese, **Chinese**, ...
 - The whole text is fairly short.
 - Sentence pairs: 1,755
 - Subject-mappings: 2,542
- Goal
 - Corresponding patterns
 - Generalization of topic-drop
- Exploiting texts
 - Semi-auto alignment
 - Annotation
 - Tag Set
 - The Mapping Table

Nouns vs. Pronouns

- The **usage of pronoun** in Korean and Chinese is not so common as that in English.
- **Pronouns** tend to be dropped relatively freely in topic-prominent languages.

English	Korean	Chinese	Number	%
noun	noun	noun	489	19.24%
pronoun	pronoun	pronoun	477	18.76%
pronoun	unrealized	pronoun	356	14.00%
pronoun	unrealized	unrealized	261	10.27%
pronoun	noun	pronoun	152	5.98%

Topic-drop

- **Dropped topics** in Korean and Chinese correspond to **pronouns** in English in most cases (almost 90%).
 - with an antecedent

Korean			Chinese		
category	number	%	category	number	%
pronoun	384	87.87%	pronoun	122	88.40%
unrealized	42	9.61%	unrealized	14	10.14%
noun	11	2.52%	noun	2	2.46%

- without any antecedent

Korean			Chinese		
category	number	%	category	number	%
pronoun	109	87.20%	pronoun	43	91.49%
unrealized	15	12.00%	unrealized	3	6.38%
noun	1	0.85%	noun	1	2.13%

Other Findings

- Topic-drop
 - Topics in Korean are omitted more frequently than those in Chinese.
 - Topic-drop in Chinese and Korean **tends to occur at the same time.**
- Definiteness
 - Definiteness tends **not to be overtly** marked in Korean and Chinese
 - There is no significant correlation **between definiteness and the topic marker in Korean.**

Implications

- Machine Translation
 - Pronouns in English aren't always translated into pronouns in Korean and Chinese.
 - Dropped subjects in Korean and Chinese need to be translated to pronouns in English with a topic relation.
 - Definiteness does not have a direct relationship with topicality in terms of translations.
- Related Libraries
 - category (noun vs. pronouns)
 - *cog-st*
 - argument optionality

Outline

1. Implementation: Translating Passives
2. A Corpus Study: *The Little Prince*
- 3. Future Work**

Future Works

- Implementation
 - Implementing in **larger grammars**
 - Applying to **various phenomena**
 - I'm Mary : watashi-ga/wa Mery-desu.
- The Corpus Study
 - Other **languages**
 - Spanish: **pro-drop** (subject-drop with rich morphology)
 - German: both pro-drop and topic-drop
 - Japanese: comparing to Korean
 - Multilingual Treebanks with *the Little Prince*.

References

- Charles D. Yang. 2002. *Knowledge and Learning in Natural Language*. Oxford: Oxford University Press.
- Choi, Hye-Won. 1999. *Optimizing Structure in Context: Scrambling and Information Structure*. Stanford, CA: CSLI Publications.
- Gundel, Jeanette K. 1999. On Different Kinds of Focus. In Peter Bosch and Rob van der Sandt (eds.), *Focus: Linguistic, Cognitive, and Computational Perspectives*, 293–305, Cambridge: Cambridge University Press.
- Huang, James C. T. 1984. On the Distribution and Reference of Empty Pronouns. *Linguistic Inquiry*. 15:531–574.
- Ishihara, Shinichiro. 2001. Stress, Focus, and Scrambling in Japanese. *MIT Working Papers in Linguistics* 39: 142–175.
- Kuno, Susumu. 1973. *The Structure of the Japanese Language*. Cambridge, MA.: MIT press.
- Lambrecht, Knud. 1996. *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents*. Cambridge: Cambridge University Press.
- Li, Charles N. and Sandra A. Thompson. 1976. Subject and Topic: A New Typology of Language. In Li, Charles N (ed.), *Subject and Topic*. New York: Academic Press. 457–489.
- Molnár, Valéria. 2002. Contrast – from a Contrastive Perspective. In H. Hasselgrd, S. Johansson, B. Behrens, & C. Fabricius-Hansen (Eds.), *Information Structure in a Cross-linguistic Perspective* (pp. 147–162). Amsterdam: Rodopi.
- Vauquois, Bernard. 1968. A Survey of Formal Grammars and Algorithms for Recognition and Transformation in Mechanical Translation. IFIP Congress (2). 1114-1122.