Parallel Text Annotation for Information Structure:

using *The Little Prince*

El Principito
Маленький Принц
어린 왕자
小王子

Sanghoun Song and Varya Gracheva

DELPH-IN Summit

June 26, 2011

Introduction

- Deep analysis of Information Structure
  - Various linguistic features
- Previous Corpus studies for Information structure
  - Bilingual Texts
    - Japanese to English (Komagata, 1999)
    - Swedish to English (Johansson, 2001)
    - Norwegian to English (Bouma et al., 2010)
  - Multilingual Texts
    - Calhoun et al. (2005), Dipper et al. (2007)
  - **Multilingual Parallel Texts**?

Why Multilingual Parallel?

- A comparative study about **distributional differences of Information Structure** in different languages
  - Languages use different phonological, morphological, and syntactic means of marking Information Structure.
  - For many languages, the full range of Information Structure marking possibilities remains unknown.
- Fully parallel text for several languages
  - How **Information Structure strategies** in different languages are related to each other.
  - How to find systematic methods to **identify topics and foci**.

Goal

- Providing a fully annotated multilingual data
  - covering Information Structure itself
  - covering relevant linguistic domains
- This data can be used
  - to support previous theoretical work

    (Engdahl and Vallduví, 1996; Lambrecht, 1996; Gundel, 1999)
  - to aid in the development of computational models involving Information Structure

Implications

- Multilingual Anaphora Resolution
  - distributional differences of dropped elements
- Grammar Libraries
  - building up a grammar library
    - for Information Structure itself
  - improving the previous libraries
    - argument optionality, cognitive status
    - tense/aspect, negation, etc.
- Transfer-based MT
  - An essential part of translation is reshaping the means of conveying information (i.e. involving IS) instead of simply changing the words or reordering phrases.

of Sentences

# of sets of sentences	**1,755**
# of sentences (eng)	1,754
# of sentences (spa)	1,742
# of sentences (rus)	1,740 (?)
# of sentences (kor)	1,753
# of sentences (cmn)	1,741

Progress

- Writing up an Annotation Guideline
  - More precise criteria for
    - multilingual annotation
    - dropped elements
- Progress
  - Target languages so far
    - English, Spanish, Russian, and Korean
    - About half of all chapters
  - Other languages
    - Chinese, Japanese, German

Future Plan

- We're planning to
  - finish the **multilingual annotation** for four languages (eng, spa, rus, and kor) this summer
  - expand annotation to **other languages**
    - Chinese, Japanese, German
  - construct an **HPSG/MRS-based treebank**
    - Redwoods Treebank (Oepen et al., 2004)
    - Hinoki Treebank (Bond et al., 2004)
  - explore the differences in the variations across languages

Preliminary Findings

- (Multilingual) Anaphora Resolution
  - Spanish: the dropped subject can be mostly resolved **within the current sentence** (i.e. the verbal inflection).
  - Korean: we have to look at **a wider scope**, because less salient elements tend to be freely dropped in Korean.
- Grammar Libraries
  - Argument Optionality
    - Any languages should have the **[focus –]** feature for dropped elements.
  - Cognitive Status
    - Pronouns need to have a close relation with **salience**.
  - Information Structure
    - It has to run parallel to **optionality and salience**.

Preliminary Findings (cont'd)

- Machine Translation
  - If the source is an English-like language and the target belongs to pro-drop languages,
    - the translation hinges on **information structure.**
  - If the source is a pro-drop language, and the target is an English-like language,
    - the dropped subject is translated into a **pronoun.**

Issues in Annotation

- Relative clauses
  a. flower$_i$ that cannot be seen.
  b. poi-ci anh-nun kkoch$_i$
     seen-COMP not-REL flower (Korean)
- Possessives
  a. [You]$_i$ have a good house.
  b. [Your house]$_j$ is very good. (Korean)
  c. A good house belongs to you. (Finnish)
  d. At you good house. (Russian)

An Example of Sentences

Q: What are you doing there?

A:

a. I am drinking. (English)

b. Bebo.
drink-PRES (Spanish)

c. Пью.
Pjuu
drink-PRES (Russian)

d. 술 마신다
swul masi-n-ta
alcohol drink-PRES-DC (Korean)

e. 我 喝 酒
wo he jiu
I drink alcohol (Chinese)

Annotation

- Preliminary steps
  - obtaining raw texts (websites or books)
  - sentence-aligning (Python script)
- Annotation
  - Software: **EXMARaLDA** (http://www.exmaralda.org)
    - used in the PROJECT SFB632 (http://www.sfb632.uni-potsdam.de)
    - XML
  - Coverage: annotation of linguistic features at various layers, multiple tiers consisting of cell(s) for each word or phrase.
  - Schema: Dipper et al. (2007) → adapted to our IS research

Annotation (cont'd)

- Syntactic and semantic layers are mostly removed.
- Parsing our data with **DELPH-IN grammars**
  - ERG for English (Copestake and Flickinger, 2000)
  - SRG for Spanish (Marimon et al., 2007)
  - KRG for Korean (Song et al., 2010)
  - RRG for Russian (Avgustinova and Zhang, 2010)
- Resolving syntactic and semantic constructions in a semi-automatic way in treebanking

Annotation (cont'd)

- Morphological layer has not been significantly modified from Dipper et al. (2007).
  - MORPH
  - GLOSS

Annotation (cont'd)

EXAMPLES OF MORPH and GLOSS

1. Russian

TXT	#	- Пью	-мрачно	ответил	пьяница
MORPH		Пь-ю	мрачно	ответи-л	пьяниц-а
GLOSS		drink-IPFV.PRS.1SG	gloomily	answer-PFV.PST.3SG.M	tippler-SG.M.NOM

2. Spanish

TXT	#	-Bebo	-respondió.	el	bebedor,	con	aire	lúgubre
MORPH	#	Beb-o	Respond-ió.	el	bebedor	con	aire	lúgubre
GLOSS		drink-PRS.1SG	respond-PST.3SG	the:M	tippler:M	with	air	lugubrious

Annotation (NP_TYPE)

tag	determiner	example
all	universal quantifiers	‘all men’
any	NPI	‘any sound
bare	bare NPs	‘grown-ups’
def	definiteness	‘the little prince’
dem	demonstratives	‘this flower’
each	distributives	‘each day’
ind	indefiniteness	‘a sheep’
kind	kindness	‘such power’
neg	negative deteminers	‘no reply’
num	numeral expressions	‘six years’
poss	possessives	‘my cold’
prop	proper names	‘France’
wh	wh-words	‘what’

Annotation (cont'd)

- EXAMPLE OF NP_TYPE

TXT	“I	am	drinking”	- replied	the	tippler.
GLOSS						
NP_TYPE					def	

Annotation (DROPPED)

- DROPPED_WORD
  - missing expression
- DROPPED_FEAT
  - properties of the dropped element
- DROPPED_IDX
  - index of antecedent

Annotation (IF/OF)

IF/OF LAYERS

- IF (**Inner Frame**) and OF (**Outer Frame**) layers: differentiate two types of discourses
  - IF: dialogues between characters within the story
  - OF: author's narration
  - – the same set of fields with information structure layers

Annotation (IS)

INFORMATION-STRUCTURE (IS)

IS layer has not been significantly modified:

- **INFOSTAT** (information status: *given (giv-active or giv-inactive)*, *new, accessible*)
- **TOPIC** (*aboutness* or *frame-setting*)
- **FOCUS** (*new (un)solicited focus*)
- **CONTRAST** (*contrastive topic* or *contrastive focus*)

Annotation (cont'd)

IF/OF EXAMPLE

	“I	**am**	**drinking.”**	**said**	**the**	**tippler.**
OF-INFOSTAT					giv-inactive	
OF-TOPIC						
OF-FOCUS	nf-sol					
OF-CONTRAST						
IF-INFOSTAT	active					
IF-TOPIC	ab					
IF-FOCUS		nf-sol				
IF-CONTRAST						

Annotation (INDEX)

- Word Alignment
  - In a semi-automatic way
  - GIZA++

	12.6.1	12.6.2	12.6.3	sentential form
English	I(1)	am(2)	drinking(3)	topic-focus
Spanish	Ø	Bebo(3)		all-focus
Russian	Ø	P'ju(3)		all-focus
Korean	Ø	swul(4)	masinta(3)	all-focus

Any Questions or Comments?