

# Building an English-Chinese advertisement corpus

Scarlet W. Y. Li, Shan Wang, Grégoire Winterstein  
The Hong Kong Institute of Education

# BACKGROUND, MOTIVATION

# Background

- The language of advertisement has been studied rather extensively (since Leech, 1966)
- However:
  - Most studies are **qualitative**
  - Most studies focus on **one language** (some exceptions: Tanaka, 1994)
  - Beyond a discourse analysis approach, the study of advertisement also offers interesting insight for **semantics and pragmatics**

# Goals

- Construction of a **bilingual advertisement** corpus:
  - Chinese (Mandarin/Cantonese) and English
- **Annotation** of the corpus
  - Argumentative relations
  - Alignment of discourse markers
- Open **access** of the data

# Argumentation theory

- Linguistic Argumentation Theory (Anscombe & Ducrot, 1983) postulates that every utterance **targets an argumentative goal**
- At its core, LAT studies **argumentative markers** and how they affect the argumentative potential of an utterance
  - John was **barely** late.  $\Rightarrow$  John is reliable/serious.
  - John was **almost** late.  $\Rightarrow$  John is not reliable/serious
- Markers have received detailed formal descriptions (Anscombe & Ducrot, 1983; Winterstein, 2010), but with **little empirical backing**

# Argumentation in Advertisement

- A recurring problem when studying argumentation is the **abduction problem**:
  - Given an utterance, how is it possible to reconstruct the goal targeted by the utterance?
- Generally, the question cannot be answered from linguistic material alone, which makes massive quantitative approaches impractical
- Advertisements have the advantage of having a relatively **clear/obvious goal**: promotion of a service/sell a product etc.

# Argumentative Markers

	Valence 1	Valence 2	Valence 1 or 2
Marker(s)	<i>Almost, (but) also, exactly, indeed, just, merely, moreover, nearly, (not) only, probably, quite, really, totally, very, even if</i>	<i>But, yet, because (of), since, though, unless, however, despite, In addition</i>	<i>Even (though)</i>

Table 1. Types of marker in the corpus

- Examples in the corpus:
  - Return Fare from **just** HK\$4,850
  - Our schools' international curriculum uses English as the language of instruction. **However**, Chinese also plays an important part in the curriculum, as all students are required to learn Putonghua, the official language in China.

# METHODOLOGY



# Methodology

- **Manual collection** of material taken from:
  - Internet
  - TV advertisements
- All material is **bilingual** (either Written Chinese / English or Cantonese/English)
  - The same content exists in both languages
  - Most of the material was prepared for the HK market
- **Manual annotation** of
  - Argumentative information
  - Alignment information between languages

# Metadata

- Advertisement and promotional material **in both English and Chinese** used by Hong Kong based companies.
- Two main sources of material:
  - Texts from the official promotional **websites** of various companies (1255 texts)
  - Transcripts of **TV advertisements** (150 ads)

# Metadata

- Metadata descriptors for the Advertisements:
  - The name of the advertising company
  - The nature of its services
  - A link to the website/ TV ad (if available online)
  - The type of advertised product
  - A screen capture in the case of a website (not used at the moment)

# Metadata

```
<ad id="ad87">
  <organization type="Tourism/ Travel Agency">Ngong Ping 360 Limited</organization>
  <product>Attraction</product>
  <source type="web">http://www.np360.com.hk/en/visitors-information/index.asp?id=143</source>

  <content lang="en" file="87-en.png">
    <title>Motion 360</title>
    <p>
      Stimulate all 5 senses at once
      All aboard the magical spaceship for an exciting ride around breathtaking Lantau Island.
      Fly over the Big Buddha, before diving underwater with Hong Kong's famous white dolphins.
      Indulge all your senses in the magnificence of Lantau, sky, land and sea, from the highest mountain peak, to the deepest gorge.
      only$245up
      Online offer 10% off
      Ngong Ping Walk n Motion Pass
    </p>
  </content>

  <content lang="zh" file="87-zh.png">
    <title>360動感影院</title>
    <p>
      五維動感影院 激新感官體驗
      踏上神秘飛船，與船長展開首次刺激歷險的飛行任務！
      飛越天壇大佛、潛入中華白海豚水域、穿梭心經簡林，帶你從多角度遨遊大嶼山，感受不一樣的視覺、聽覺、嗅覺、觸覺及動感全方位震撼。
      只需$245起
      網上尊享9折優惠
      昂坪大動感同行套票
    </p>
  </content>
</ad>
```


# Argumentative annotation

- Annotation done in two steps:
  - **Automatic** annotation of argumentative markers
  - **Manual** annotation of scope and bilingual relations
- Two phases
  - English / Chinese (done)
  - Chinese / English (underway)
- Annotation tool: **Webanno** (Yimam et al., 2013)

# WebAnno

The screenshot displays the WebAnno web interface. At the top, there is a navigation bar with a logo, the title "Annotation", and a "WebAnno | Home |" link. On the right, it shows the user "scarlet" and a "Log out" button. Below this is a toolbar with several sections: "Document" (Open, Prev, Next, Export, Settings), "Page" (First, Prev, Go to 1, Next, Last), "Help" (Guidelines), and "Workflow" (Done). The main content area shows a list of sentences from an "Argumentation Annotation/AnnotationRaw1-500.tsv" file, with "showing 1-200 of 3035 sentences". Sentence 3 is selected and expanded to show its annotations. The text of sentence 3 is: "Hong Kong Brand New Company only HK \$ 4,600 , Hong Kong Shelf Ccompany only HK \$ 3,688 , Company Secretary only HK \$ 400 , Registered Office only HK \$ 700 , BVI Company , Seychelles Company and Samoa Company only HK \$ 5,500 .". Annotations include "Marker" and "Argument" boxes connected by "InScope" arrows. There are also "PragLink | Bilingual" and "Synonym | Bilingual" links between different parts of the sentence. The interface also shows other sentences like "Read more about our Latest Promotions" and "最新優惠". At the bottom, there is a footer with the text "Technische Universität Darmstadt – Computer Science Department – WebAnno – 2.1.0 (2015-03-04 19:33:16, build 2467)".

# Manual annotation

- For all the markers automatically pre-annotated:
  - Annotation of the **scope** of the marker
  - **Link** between scope and marker  
*John almost hit the wall.*  

  - **Alignment** with a marker in the other language

# Bilingual relations

- Bilingual relations were annotated between:
  - Argumentative Markers
  - Scope of the markers
- Use of the scheme of Bond & Wang (2014):
  - Synonym (=): 因為/*because*
  - Pragmatic Link ( $\approx$ ): 咁/*but*
  - Lexical Link ( $\sim$ ): 更可/*also*
  - Partial translation (#):  
*China 's taxation can be categorized/稅收劃分為*
  - Hypernym (>)
  - Hyponym (<)
  - Antonym (!)



# More examples

Relations	Examples
<b>Synonym: =</b>	<p>We &lt;&lt;<b>also</b>&gt;&gt; sell examination publications on behalf of our partnering examination bodies .</p> <p>我們&lt;&lt;<b>亦</b>&gt;&gt;有為合作機構代售考試刊物。</p>
<b>Partial Translation: #</b>	<p>It is &lt;&lt;<b>also</b>&gt;&gt; used in some Light Buses , Vans and Passenger Cars .</p> <p>電裝在香港擁有領導地位，其中超過百分之九十五的雙層及單層巴士都裝用電裝空調系統，另多款私家車、輕型客貨車及小巴&lt;&lt;<b>亦有</b>&gt;&gt;採用。</p>
<b>Pragmatic Link: ≈</b>	<p>This may be due to large amounts of cash being excessively invested in fixed assets , or &lt;&lt;<b>because</b>&gt;&gt; the inventory turnover ratio is low , or credit policy is too loose , etc .</p> <p>其中的原因可能是大量現金被過多地投放於固定資產，也&lt;&lt;<b>可能</b>&gt;&gt;是存貨周轉率低，或信用政策過於寬鬆等。</p>

# An example

- [...] which **not only (=) resolve the problem (<)** to facilitate the business development for enterprise, **but also (=) effectively use various finance tools (=)** to raise capital for the enterprise to facilitate their business development.
- **不僅為企業解決了資金鏈的問題，更有效地利**用**各種融資工具**為企業籌集發展業務的資金。

# Contents of the corpus

- **1405** documents in total
  - **1255** texts from internet
  - **150** TV ads transcripts
- **150** different companies
- Varied services:
  - Banking, finance
  - Entertainment
  - Retail
  - Food industry
  - ...

# Contents: sizes

	English		Chinese	
	# Tokens	Avg. Tok. / ad.	# Char	Avg. Char. / ad.
Web material	152090	121.2	301254	240.0
TV advertisement	13026	86.8	21680	144.5
Total	165116	117.5	322934	229.8

# EXAMPLES OF USE

# Use of the corpus

- The corpus has different practical uses:
  - Study of the **advertisement discourse** in a comparative perspective
  - Study of argumentative markers and their **crosslinguistic differences**
  - Use for tasks related to **opinion mining** (Pang & Lee, 2008), and more generally machine-learning related tasks

# Proportion of translations

- Amsili et al. (2012) investigate the pressure to use **additive markers**
  - Usually those are markers supposed to be obligatory
    - *Jo had fish, and Mo did # (too).*
  - There are “fringe cases” where the use of an additive appears optional:

*Hartmann's joy was apparent in his beautifully cut hair, his expensive suit, his manicured hands, the faint aura of cologne that heralded his approach; in his mild and habitually smiling face, **too**, his expressive walk, in which the body, leaning slightly forward, seemed to indicate amiability*
- Does the pressure to use an additive varies cross-linguistically? Or is it a universal pragmatic constraint?

# Preliminary results

- Comparison of the rate of non-translation of *also* and *only*

	Translation	Non-translation
<i>Also</i>	157	65
<i>Only</i>	86	37

- The differences are not significant (Fisher's test,  $p = 0.9$ )
- This (weakly) argues for a general account of the usage of discourse markers, consistent with some of the literature (Zeevat, 2014)
- Current work:
  - Distinguish between types of translations
  - Look at the CN/ No EN translations (annotation underway)



# Argumentation algebra

- Argumentative operators are compositional (Winterstein, 2010)
  - Some combinations of markers are predicted to be more frequent than others:  
*X but only X*  
*almost X but did not X*
- Paired with a sentiment lexicon, the corpus can be used as a test bed for an argumentative algebra (e.g. Poria et al., 2014), under the assumption that utterances all argue for a similar goal

# Further annotation

- Beyond the identification of the scope of a marker, the identification of full argumentative schemes is planned:
  - Argumentative **premise**
  - Argumentative **conclusion**
  - **Type** of argumentation (against/for, opposition, addition, parallelism etc.)
- This should help improve systems to automatically detect argumentation schemes

**THANK YOU FOR YOUR ATTENTION**

# References

- Amsili P., E. Ellsiepen and G. Winterstein (2012) Parameters on the obligatoriness of *too*, in *Proceedings of LENLS 2012*, Miyazaki, Japan
- Anscombre, J.-C. and O. Ducrot (1983). *L'argumentation dans la langue*. Liège, Bruxelles: Pierre Mardaga.
- Bond F., Wang S. (2014). Issues in building English-Chinese parallel corpora with Wordnets. In *Proceedings of the 7th Global WordNet Conference (GWC 2014)* Tartu. pp 391–399
- Leech, G. (1966). *English in Advertising*. London: Longman.
- Pang, B. and L. Lee (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2(1–2), 1–135.
- Poria, S., E. Cambria, G. Winterstein, and G.-B. Huang (2014). Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems* 69, 45–63.
- Tanaka, K. (1994). *Advertising Language, a pragmatic approach to advertisements in Britain and Japan*. London: Routledge.
- Winterstein, G. (2010). *La dimension probabiliste des marqueurs de discours. Nouvelles perspectives sur l'argumentation dans la langue*. Ph. D. thesis, Université Paris Diderot.
- Yimam, S.M., Gurevych, I., Eckart de Castilho, R., and Biemann C. (2013): WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. In *Proceedings of ACL-2013*, demo session, Sofia, Bulgaria.
- Zeevat H. (2014). *Language production and interpretation: linguistics meets cognition*. Leiden: Brill.