Measuring Distributional Semantic Effects in Syntactic Variation

Kristina Gulordava

University of Geneva
CLCL group

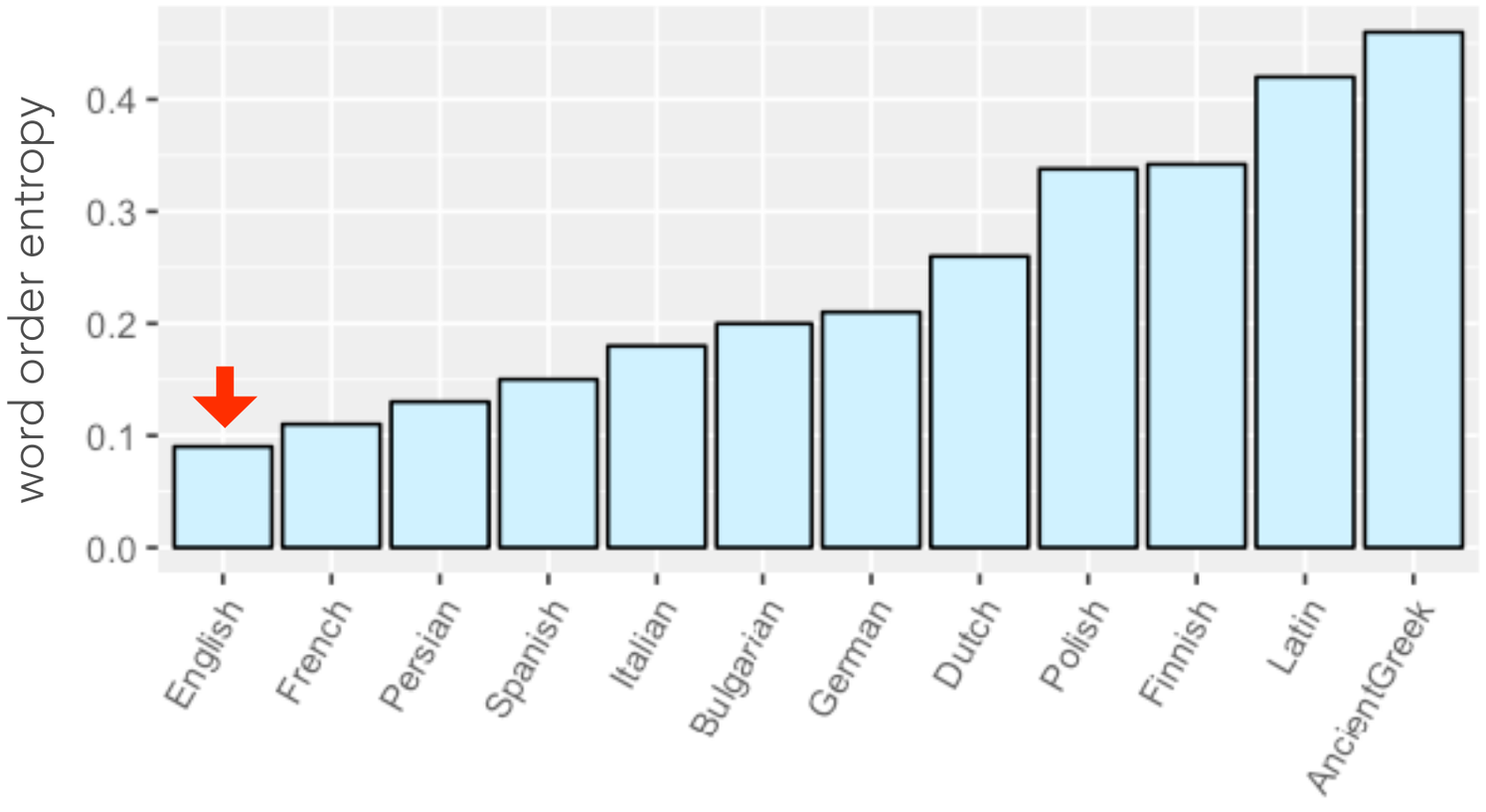
Why study syntactic variation?

- Understanding mechanisms of **language use**
  - by taking advantage of **corpus data** on many languages
- Flip side of syntactic processing
  - what are the common and distinct properties of mechanisms speakers use to process *and* produce sentences?
  - e.g. long syntactic dependencies are hard to process and are dispreferred in language use
- NLP perspective: language generation [White & Rajkumar 2012]

Word order variation

Across languages, word order is much more variable than in English

[Liu 2010; Futrell et al. 2015; Gulordava & Merlo 2016]



Cases of word order variation

- Dative alternation

I gave Mary a book

I gave a book to Mary

- Alternation of PPs

He left on Monday by car

He left by car on Monday

- Verb-particle split

She threw the trash out

She threw out the trash

Assumption: different word order - the same meaning

Lexico-semantic effects in Verb-Particle split order (Lohse et al. 2004)

She threw$_{\mathbf{V}}$ [the trash]$_{\mathbf{NP}}$ out$_{\mathbf{Prt}}$ **?** She threw$_{\mathbf{V}}$ out$_{\mathbf{Prt}}$ [the trash]$_{\mathbf{NP}}$

Entailment tests for Verb ↔ Particle semantic relation:

[X V NP Prt] entails [X V NP] → **V** interpretation is *independent*

[X V NP Prt] entails [NP PredV Prt] → **Prt** interpretation is *independent*

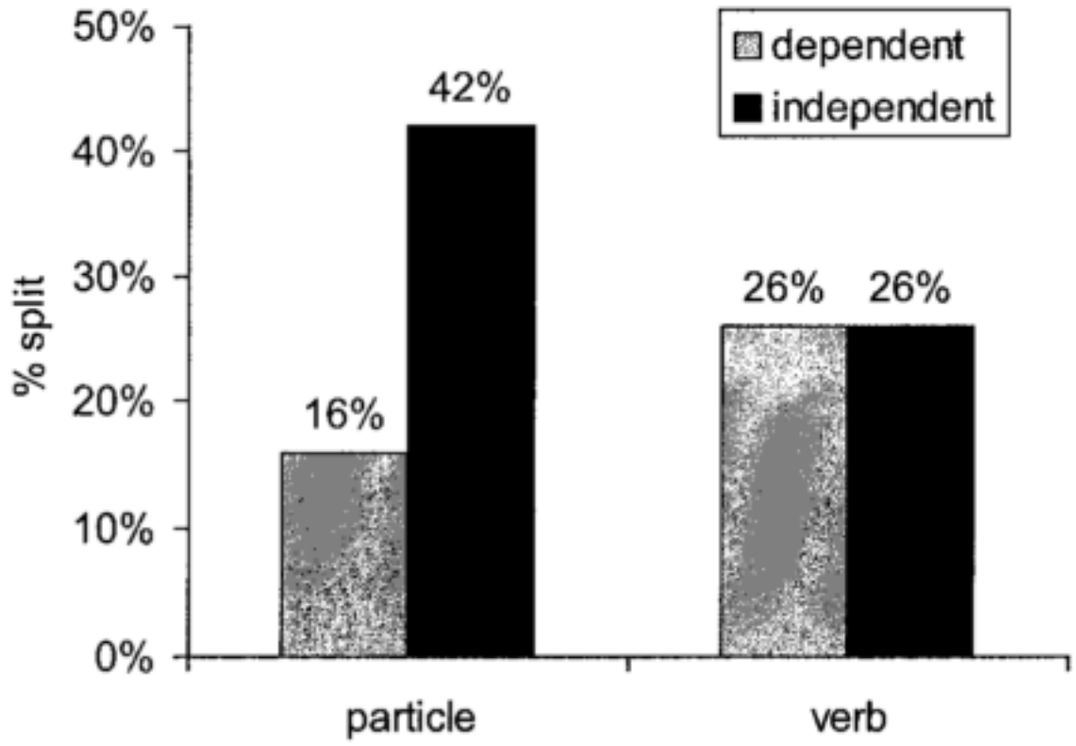
non-compositional

	V is dependent	**V is independent**
Prt dependent	*rip off* this guy	*call* parents *up*
Prt independent	*turn off* the lights	*bring in* her chair

fully compositional

Lexico-semantic effects in Verb-Particle split order (Lohse et al. 2004)

- When particle meaning is opaque the split order is much less frequent than when particle is semantically transparent
- No symmetric effect for verbs



- Compositionality of V-Prt constructions affects their word order
- Similar results for V PP1 PP2 alternation [Wiechmann 2011]

Compositionality of verb-particle phrases: DS perspective

- Many distributional accounts of non-compositionality, also for verb-particle constructions
  - *McCarthy et al. 2003; Bannard et al. 2003; Kim & Baldwin 2007*
- Prediction task: non-compositionality was annotated using entailment tests, binary or ranking judgments
  - Agreement ranges between 55%-80%* in different studies
- Categorical distinction or a continuum of non-compositionality?

More related work on MWEs

- **Syntactic "fixedness"** as a correlate of **semantic idiosyncrasy**
  - modification: "break the diplomatic ice"
  - passivisation: "it's a shame the beans were spilled"
  - *Fazly & Stevenson 2007; Bannard 2007; Baldwin & Kim 2010*
- Such syntactic properties were used to *predict the semantic class of MWEs*
  - syntactic behaviour → lexico-semantic properties
  - **?** lexico-semantic properties → *systematic* syntactic behaviour

Case study: Adjective-noun order in Italian

- There is substantial variation between prenominal and postnominal order **30% / 70%**
  - lexically conditioned - adjective lemma accounts for a lot of variance
- Extensively studied, but mostly in connection with:
  - Meaning change in basic cases of Adj-N vs N-Adj with bare nouns
  - Relative order of several adjectives
    [Cinque 1994; McNally & Boleda 2004; Vecchi 2013; …]
  - Adj+N+**PP** cases are not well studied but are very frequent

***~25%** of all adjective cases*
= every 5th sentence

ADJECTIVE ORDER IN COMPLEX NPS

Oggi devo finire [un **importante**$_A$ compito$_N$ di matematica$_{PP}$] e quindi non posso uscire. **Adj N PP**

Oggi devo finire [un compito$_N$ **importante**$_A$ di matematica$_{PP}$] e quindi non posso uscire. **N Adj PP**

default order?

Oggi devo finire [un compito$_N$ di matematica$_{PP}$ **importante**$_A$] e quindi non posso uscire. **N PP Adj**

'Today I have to finish an important math homework so I can't go out.'

ADJECTIVE ORDER IN COMPLEX NPS

PP phrase	#	Adj N PP	N Adj PP	N PP Adj
None	**3933**	**0.31**	**0.69**	**–**
A	137	0.46	0.54	0.01
Che	236	0.42	0.58	–
Da	41	0.31	0.69	–
Di	**1132**	**0.53**	**0.45**	**0.03**
In	148	0.43	0.55	0.01
Op	189	0.46	0.53	0.01

[extracted from UD 1.2 Italian treebank]

N-DI-N PHRASES AND ADJECTIVES: CLASSIFICATION CHALLENGES

- Lexicalized phrases (compositional / non-compositional?) are very fixed:
  - punto di vista ('point of view'), colpo di stato ('coup d'etat'), uovo di Pasqua ('Easter egg')
  - but: "… quel punto *particolare* di vista"
    "… la rottura dell'uovo *artigianale* di Pasqua più grande del mondo, alto 8,5 metri"
- Frequent (compositional) collocations can also be relatively fixed:
  - regalo di Natale, regalo di soldi, regalo di 18 anni
  - colpo di pistola "colpo mortale/preciso/improvviso di pistola"

Hypothesis

- Non-compositionality of N-di-N determines its syntactic fixedness = adjacency

- Adjacency of N-di-N affects adjective placement in *N Adj di N* position

- *Note:* adjective placement is determined by a number of factors with which N-di-N fixedness interacts

Data

- Cases of Adj + N-di-N phrases extracted from automatically parsed Italian Wikipedia (200.000.000 words) [Baroni et al. 2009]
  - based on PoS tags, dependency relations
  - only simple adjectives
- > 100.000 observations (~ 70.000 unique N-di-N phrases)
- **Kept only adjectives that appear both prenominally and postnominally**
- Kept 1000 most frequent N-di-N phrases
- Binary regression task: predict **Order {N Adj di-N, other}**

FACTORS

- **Compositionality**:
  - sim(N1-di-N2, N1)
  - sim(N1-di-N2, di-N2)
    - cosine similarity in a count and PPMI vector space extracted from the Wikipedia (window size = 5, 10)
- Frequency
- PMI of N-di-N
  - log P(N1-di-N2) / P(N1) * P(di-N2)

- Random factor: adjective lemma
  - to account for lexical variation

Preliminary Results

Factors	effect on N1 **Adj** di N2	*p*
Sim N1	n.s.	0.3
Sim N2	↑	<0.001
Frequency	↓	<0.001
PMI	↓	<0.001

N = 7008, # adjective lemmas = 261

- PMI is a strong, robust effect
- Semantic similarity measures are not robust on infrequent phrases
- The current model doesn't explain all the variance

Conclusions

- Instead of looking at syntactic properties as classification features for semantic classes of MWEs we can do the inverse: look at semantic properties of phrases as cause of syntactic variation

- New data on adjective variation in Italian
  - word order is conditioned by the N-PP relationship

- More questions than answers:
  - how do distributional properties of phrases affect their syntactic behaviour?
  - how the syntax-lexicon interface is organised?

Thank you!