Quantitative Computational Syntax: dependencies, intervention effects and word embeddings

Paola Merlo

Computational Learning and Computational Linguistics group (CLCL)
University of Geneva

SyntaxFest, Paris, August 2019

Preamble

- I have been pursuing a research agenda that I call **quantitative computational syntax** (Merlo, 2016): quantitative *differentials are the expression of underlying grammatical properties.*
- We study the quantitative aspects of traditional syntactic phenomena, in a computational, corpus-driven framework.
- Word order in the noun phrase: universal 18, universal 20, Dependency Length Minimisation effects
- Causative alternations and typology
- **Long-distance dependencies**

Related to interests in human processing and language optimisation, evolution, efficiency

In this talk (Merlo and Ackermann, CoNLL 2018); Merlo (BBNL, 2019)

- Neural networks work in practice, but do they learn **in theory**? (Steedman, LTA 2018)

- **Long-distance dependencies** are the hallmark of human languages.

What do vectorial spaces really learn?

- Several pieces of work have recently studied core properties of language in syntax. Results are inconclusive.
  - Linzen et al 2016: RNNs could predict the right agreement word but with some mistakes
  - Gulordava et al 2018: RNNs can learn agreement patterns in four languages with almost human performance
  - Kunkoro et al 2018: Gulordava effect is artifact of learning first word in sentence.
- Studies of **long-distance dependencies** equally inconclusive
  - Wilcox et al 2019: RNNs learn basic properties of long-distance constructions
  - Merlo and Ackermann 2018: word embeddings do not correlate with experimental results in intervention effects

Long-distance dependencies and intervention

Not all long-distance dependencies are equally acceptable.

(1a) What do you think **John** bought <what> ?

(1b) * What do you wonder **who** bought <what>?

(2a) Show me the tiger that **the lion** is washing <the tiger>.

(2b) Show me the tiger that <the tiger> is washing the lion.

(3) ??/ok Jules sourit aux étudiant(s) que l'orateur <**étudiant(s)**> endort <étudiant(s)> sérieusement depuis le début.
'Jules smiles to the students who the speaker is putting seriously to sleep from the beginning.'

Intervention theory (Rizzi 1990, 2004)

- Core to the explanation of these facts is the notion of **intervener.**
- Intervener: an element that is **similar** to the two elements that are in a long-distance relation, and **structurally intervenes** between the two, blocking the relation (shown in bold).
- N.B. Intervention is defined structurally and not linearly.
  *When do you wonder who won?
  You wonder who won at five
  When did the uncertainty about who won dissolve?
  The uncertainty about who won dissolved at five

Gradation in intervention

Long-distance dependencies exhibit gradations of acceptability

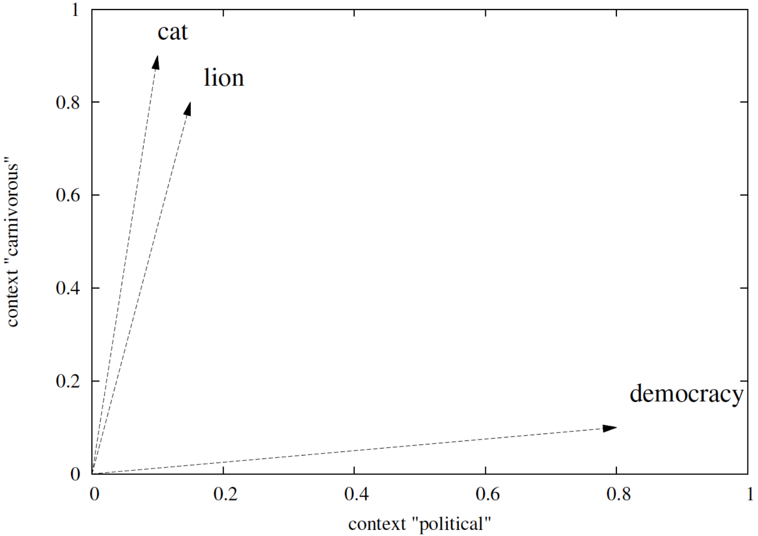
- a. *What do you wonder who bought?
- b. ??Which book do you wonder who bought?
- c. ?Which book do you wonder which linguist bought?

- Lexical restriction improves acceptability. Acceptability judgements ($<$ = better): c $<$ b $<$ a.
- Agreement features: number creates intervention effects (so decreases acceptability) but person doesn't.
- Animacy: children don't seem to mind in relative clauses but intervention effects have been found in weak-islands (Franck et al., 2015).

Intervention theory notion of similarity: summary

- Long-distance dependencies are acceptable if there is no intervener.
- Establishing if an element is an intervener requires the calculation of similarity of feature vectors, where some features are morpho-syntactic and some are semantic.
- This is very reminescent of current notions of similarity over distributional semantic spaces.

Vectorial spaces



Vector spaces

- **Word embeddings:** definition of lexical proximity in feature spaces, vectorial representation of the meaning of a word, defined as the usage of a word in its context.
- Tasks that confirm this interpretation are association, analogy, lexical similarity, entailment.
- Does the similarity space defined by word embeddings capture the **grammatically-relevant notion of similarity** at work in long-distance dependencies?
- The work is done on French.

Weak island intervention and animacy

Data kindly provided to us by Sandra Villata and Julie Franck.

Weak islands, ANIMACY MISMATCH

Quel cours te demandes-tu **quel étudiant** a apprécié?
[+Q,+N,-A] [+Q,+N,+A]

Which class do you wonder which student appreciated?

Weak islands, ANIMACY MATCH

Quel professeur te demandes-tu **quel étudiant** a apprécié?
[+Q,+N,+A] [+Q,+N,+A]

Which professor do you wonder which student appreciated?

Weak island intervention and animacy

Quel cours te demandes-tu **quel étudiant** a apprécié?
[+Q], [+N], [-A] [+Q], [+N], [+A] ANIMACY MISMATCH
Which class do you wonder which student appreciated?

Quel professeur te demandes-tu **quel étudiant** a apprécié?
[+Q], [+N], [+A] [+Q], [+N], [+A] ANIMACY MATCH
Which professor do you wonder which student appreciated?

- Experiment 1 manipulated the lexical restriction of the *wh*-elements (both bare vs. both lexically restricted), and the match in animacy between the two *wh*-elements, as shown. All verbs required animate subjects.
- Data: acceptability judgments collected off-line on a seven-point Likert scale. No time constraints.
- Results: clear effect of animacy match for lexically restricted phrases and less so for bare *wh*-phrases.

Weak island intervention and animacy

Quel cours	te demandes-tu	**quel étudiant**	a apprécié?	
[+Q], [+N], [-A]		[+Q], [+N], [+A]		ANIMACY MISMATCH

Which class do you wonder which student appreciated?

Quel professeur	te demandes-tu	**quel étudiant**	a apprécié?	
[+Q], [+N], [+A]		[+Q], [+N], [+A]		ANIMACY MATCH

Which professor do you wonder which student appreciated?

- Both the pair *(class, student)* and the pair *(professor, student)* are close in a semantic space that measures semantic field and association-based similarity.
- Human speakers rate the first sentence as on average a little better as there is a mismatch in animacy, hence the effect of intervention is weaker.
- If word embeddings learn grammatically-relevant notions of similarity, then *(professor, student*) should be more similar, predicting lower acceptability, since they are both animate, compared to *(class, student)*, a pair with a mismatch in animacy.

Object relatives intervention and number

Object relatives, NUMBER MATCH

Jules sourit à l' étudiant que l' **orateur** $<$**étudiant**$>_2$ endort $<$étudiant$>_1$ sérieusement depuis le début.

Jules smiles to the student who the speaker is putting seriously to sleep from the beginning.

Object relatives, NUMBER MISMATCH

Jules sourit aux étudiants que l' **orateur** $<$**étudiants**$>_2$ endort $<$étudiants$>_1$ sérieusement depuis le début.

Jules smiles to the students who the speaker is putting seriously to sleep from the beginning.

Object relatives intervention and number

Object relatives, NUMBER MATCH
Jules sourit à l' étudiant que l' **orateur** <**étudiant**>$_2$ endort <étudiant>$_1$ sérieusement depuis le début.
Jules smiles to the student who the speaker is putting seriously to sleep from the beginning.

Object relatives, NUMBER MISMATCH
Jules sourit aux étudiants que l' **orateur** <**étudiants**>$_2$ endort <étudiants>$_1$ sérieusement depuis le début.
Jules smiles to the students who the speaker is putting seriously to sleep from the beginning.

- Experiment: items crossing structure (object relative clauses vs. complement clauses) and the number of the object (singular vs. plural).
- Data: On-line reading times (milliseconds). Interference examined on the agreement of the verb in the subordinate clause.
- Results: Speed-up effect in number mismatch configurations.

Object relatives intervention and number

Object relatives, NUMBER MATCH
Jules sourit à l' étudiant que l' **orateur** $<$**étudiant**$>_2$ endort $<$étudiant$>_1$ sérieusement depuis le début.
Jules smiles to the student who the speaker is putting seriously to sleep from the beginning.

Object relatives, NUMBER MISMATCH
Jules sourit aux étudiants que l' **orateur** $<$**étudiants**$>_2$ endort $<$étudiants$>_1$ sérieusement depuis le début.
Jules smiles to the students who the speaker is putting seriously to sleep from the beginning.

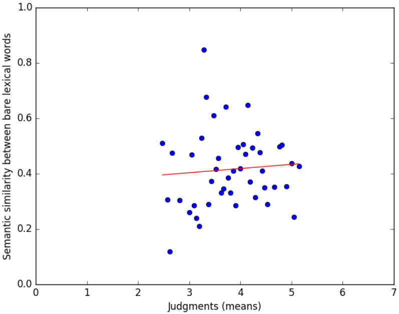
- In the NUMBER MATCH cases, the intermediate trace causes intervention effects (the presence of a trace is supported by other experiments on agreement errors).
- Human speakers read the verb *endort* in the second sentence on average faster than in the first, as there is a mismatch in number, hence the effect of intervention is weaker.
- If word embeddings learn grammatically-relevant notions of similarity, then *(student, speaker)* should be more similar, predicting slower reading times, since they are both singular, compared to *(students, speaker)*, a pair with a mismatch in number.

Calculating the word and phrase vectors

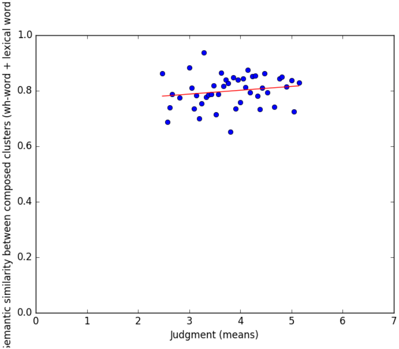
- The pairs of words or phrases (indicated in bold in the examples) were used to construct the vector-based similarity space.
- For each of these words, French FastText word embeddings (Bojanowski et al., 2016). 5-word window on Wikipedia data using the skip-gram model resulting in 300-dimension vector Every word is represented as an *n*-gram of characters.
- Quality of resulting similarity spaces was inspected.
- The cosine is a well-known and efficient measure of vector similarity. It is a symmetric measure. It has been shown to capture analogical semantic similarity in vector space.

Results with the cosine operator: weak islands

Bare nouns

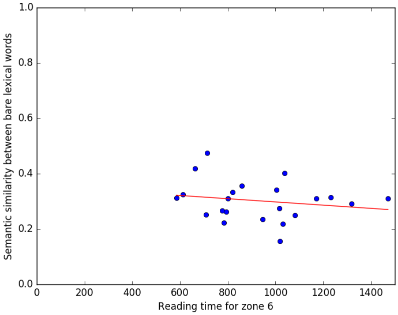


Composed phrases

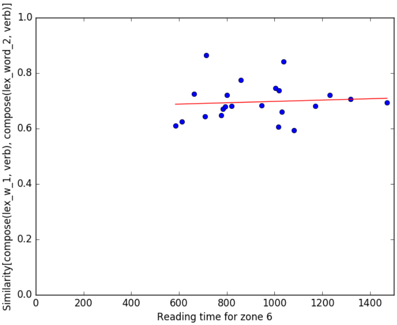


Results with the cosine operator: object relatives

Bare nouns



Composed phrases



Analysis of the results: do we capture a binary distinction?

- *Animacy* in *wh*-islands: expected inverse correlation between mean similarity and mean acceptability.

  (Match: mean sim=0.394, mean acc=3.65; mismatch: mean sim=0.293, mean acc=4.00.)

- *Number* in relative clauses: no expected direct correlation between mean similarity and mean reading time.

  (Match condition: mean sim=0.678, mean RT=962.96; mismatch: mean sim=0.705, mean RT=896.03).

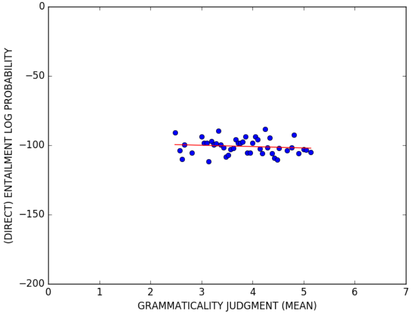
- Also notice that the average similarity score for the number match condition is lower than for the number mismatch condition.

Asymmetric operator

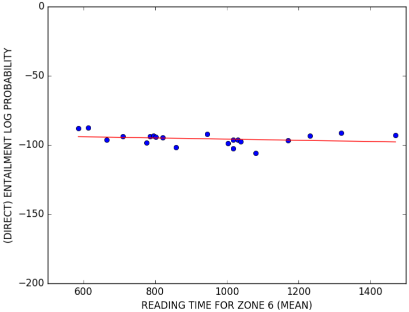
- Human grammaticality judgments differ depending on whether the feature set of the long-distance element is properly included or properly includes the feature set of the intervener. If the features of the long-distance dependency are a superset of the features of the intervener, sentences are judged more acceptable (Rizzi, 2004).
- These fine-grained differences in grammaticality judgments suggest that it might be more appropriate to calculate similarity with an asymmetric operator.
- The asymmetric measure we use here has been developed to capture the notion of entailment. This operator has been shown to learn the notion of hyponymy with good results (Henderson and Popa, 2016).

Results with asymmetric operator

Weak islands, bare nouns.



Object relatives, bare nouns.



Discussion

- These results also confirm a **lack of correlation.**
- The convergence of these results is important as null effects are always hard to confirm and explain.
- All experiments,
  - across constructions (weak island and object relatives),
  - across type of noun phrase (bare or composed),
  - across measurement method of the experimental dependent variable (off-line grammaticality judgments and online reaction times),
  - and across operators (symmetric and asymmetric)

  show a consistent lack of correlation between experimental results, and the notion of similarity encoded in word embeddings.

Extension to sentence embeddings and prediction task

- Prediction task: can we identify right sentence type?
- Translate items also into a new language: English.
- Sentence embeddings: additive bag of vectors model (same word embeddings as previously).
- Classifier: Multi-layer perceptron (4 outputs, 2 hidden layers, 50 and 30 dims). n-fold cross-validation (each quadruple of stimuli is used for testing).
- Dependent variable: Accuracy, as a measure of how much the information in the input embeddings supports the discrimination of the four sentence types in a categorical classifier.

Long-distance dependencies stimuli

Weak islands

LexI **Which class** do you wonder **which student** liked?

LexA **Which professor** do you wonder **which student** liked?

BareI **What** do you wonder **who** liked?

BareA **Who** do you wonder **who** liked?

Object Relatives

ORCsg Julie smiles to the **student** that the **speaker** is putting to sleep seriously from the beginning.

ORCpl Julie smiles to the **students** that the **speaker**is putting to sleep seriously from the beginning.

CMPsg Julia points out to the **student** that the **speaker** has been yawning frequently from the beginning.

CMPpl Julia points out to the **students** that the **speaker** has been yawning frequently from the beginning.

Weak Islands Expectations and Results

Expectations		
Acc(LexA)	$<$	Acc(LexI)
Acc(BareA)	$<$	Acc(BareI)
Acc(LexA)	$>$	Acc(BareA)
Acc(LexI)	$>$	Acc(BareI)

	French	English
BareA	0.909	0.272
BareI	0.788	0.485
LexA	0.151	0.091
LexI	0.303	0.151

- For French, the prediction on the effect of animacy in the lexically specified case is confirmed, but the others are not.
- For English, the prediction for the effect of animacy is confirmed both in bare *wh*-phrases and in lexicalised *wh*-phrases, but the others are not.

Relative Clause Expectations and Results

Expectations
Acc(ORCsg) $<$ Acc(ORCpl)
Acc(CMPsg) $=$ Acc(CMPpl)
Acc(ORCsg) $<$ Acc(CMPsg)
Acc(ORCpl) $=$ Acc(CMPpl)

	French	English
ORCsg	0.250	0.417
ORCpl	0.125	0.375
CMPsg	0.291	0.292
CMPpl	0.500	0.292

- For French, none of the predictions is confirmed.
- For English, the only confirmed prediction says that *number*, whether singular or plural should be roughly similar in completives, the control case.

Discussion

- Current word embeddings, i.e. dictionaries in a multi-dimentional vectorial space, clearly encode a notion of similarity, as shown by many experiments on analogical tasks and textual and lexical similarity.
- They do not however encode the notion of similarity that has been shown in many human experiments to be at work and to be definitional in long-distance dependencies.
- They do not encode therefore a core linguistic notion.

Discussion – Finer-grained distinctions among intervention theories

- Narrow intervention (grammar-based, explains ungrammaticality, weak islands): only morpho-syntactic features are relevant to define intervention, so the fact that word embeddings — meant to capture semantic notion of similarity — do not correlate with grammar-based notion of similarity is to be expected. ☺
- Cue-based memory based models (processing-based, explain difficulty, object relatives): similarity can take any feature type into account (as demonstrated in experiment on weak islands above, which also manipulate semantic reversibility) and intervention is a kind of interference at retrieval in memory. Correlation is expected. ☹

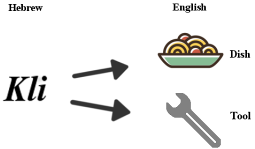
Cross-lingual word embeddings and the bilingual lexicon (Merlo and Rodriguez, CoNLL 2019)

Do cross-lingual word embeddings have the same structure as the bilingual lexicon?

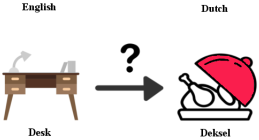
The bilingual lexicon is a space of distributed word representations where word forms from different languages map onto a common abstract conceptual code (Van Hell and de Groot, 1998).

Shared translation, false and true friends effects

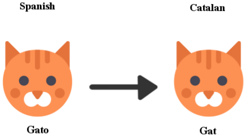
- **Shared translations effect** Task: similarity rating



- **False friends effect** Task: cross-modal picture decision.



- **True friends effect** Task: production, picture-naming.



Word pairs types

False friends words with same form, but semantically different.

Real translations of the false friends: the real L2 translations of the L1 word that also has a false friend.

True friends words sharing form and meaning.

Normal translations words semantically equivalent, but with a different form.

Uncorrelated words words lexically and semantically uncorrelated.

Word pairs types

FALSE FRIENDS		REAL TRANSLATIONS		TRUE FRIENDS		NORMAL TRANSLATIONS	
arrange	arrangiare	arrange	disporre	family	famiglia	jam	marmellata
		arrange	sistemare	fantastic	fantastico	overview	panoramica
		arrange	organizzare	future	futuro	journey	viaggio
attend	attendere	attend	frequentare	general	generale	keep	tenere
		attend	assistere	generation	generazione	kind	tipo
bald	baldo	bald	calvo	guide	guida	leave	partire
		bald	pelato	historial	storica	light	luce
brave	bravo	brave	coraggioso	industry	industria	mean	significare
		brave	valoroso	local	locale	mood	umore

The six experimental predictions

HYP. 1	Cross-lingual word embeddings pairs are more similar than their aligned monolingual counterparts
HYP. 2	For two L2 words sharing a translation in L1, cross-lingual word embeddings are more similar than monolingual word embeddings
HYP. 3	Real translations are more similar than their corresponding false friends
HYP. 4	False friends are more similar than uncorrelated pairs
HYP. 5	True friends are more similar than normal translation pairs
HYP. 6	Normal translation pairs are more similar than real translations of false friends

Cross-lingual word embeddings models

VECMAP, cross-lingual word embedding, the state-of-the-art for bilingual lexicon induction (Artetxe et al., 2018)

M2VEC, a weakly-supervised, concept-based adversarial model (Wang, Henderson and Merlo, 2019). This method is based on the idea that languages use similar words to express similar concepts. It uses concepts, drawn from Wikipedia, rather than words to learn competitive cross-lingual word embeddings.

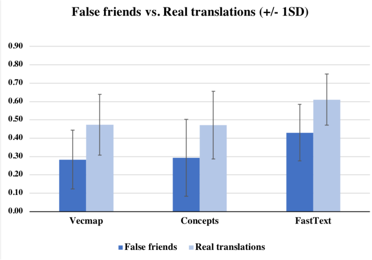
FastText,subword sequences, is important for the false and true friends experiments. Then trained with VecMap.

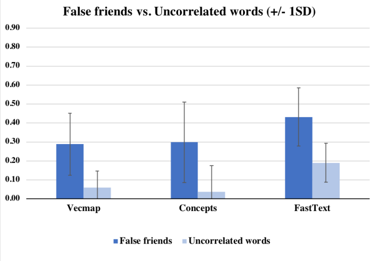
Shared translation effect results

translation pairs	shared translation pairs
wood-legno	legno bosco
wood-bosco	
block-blocco	blocco ceppo
block-ceppo	blocco bloccare
block-bloccare	blocco ostacoalre
block-ostacolare	ceppo bloccare
	ceppo ostacolare

- Both cross-lingual models show higher mean similarity scores for L2-words that share a common L1 source than the monolingual model ($p < 0.021$).

False and true friends effect results

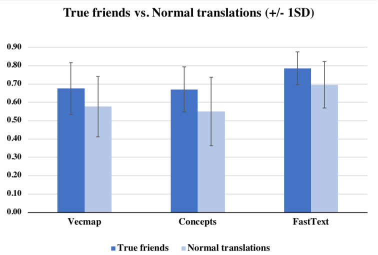


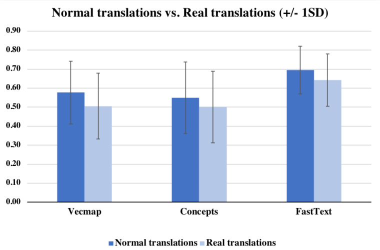


HYPOTHESIS 3 Confirmed: real translations have a better similarity score than their corresponding false friends.

HYPOTHESIS 4 Confirmed: False friends are significantly more similar than uncorrelated words.

False and true friends effect results





HYPOTHESIS 5 Confirmed: true friends have a better similarity score than normal translation pairs

HYPOTHESIS 6 Confirmed: normal pairs of words have a higher similarity score than real translations of false friends.

Discussion

- **Current word embeddings have the same structure as the bilingual lexicon.**
- **Total order of similarit**y: true friends > normal translations > real translations > false friends > uncorrelated pairs.
- True friends match both in form and meaning, normal and real translations match only in meaning, and false friends match only in form.
  This order indicates that **similarity based on meaning is more important that similarity based on form.**

Conclusions

- Human languages exhibit the ability to interpret elements distant from each other in the string as if they were adjacent.
- Results show that word embeddings and the similarity spaces they define do not encode this notion of intervention similarity in long-distance dependencies, and that therefore they fail to represent this core linguistic notion of similarity.
- Current word embeddings have the same structure as the bilingual lexicon.

Future work

- We will, grudgingly, try context-aware word embeddings (ELMO, BERT and other muppets).

The end

- Thank you.