Exploring Idiomaticity with Variant-based Distributional Measures and Shannon Entropy

Marco S. G. Senaldi[1]

Gianluca E. Lebani[2]

Alessandro Lenci[2]

[1] Scuola Normale Superiore, Pisa

[2] University of Pisa

DGfS 2017 – Saarbrücken | 9th March 2017

Summary

1. **Idiom type identification task** on **90 Italian V-N** combinations and **26 Italian Adj-N** combinations

    - **distributional indices of compositionality** that leverage the restricted lexical substitutability of idiom constituents

2. **Predicting human ratings on idiom syntactic flexibility** from the **indices in (1)** and **entropy-based indices of formal flexibility**

Summary

1. **Idiom type identification task** on **90 Italian V-N** combinations and **26 Italian Adj-N** combinations
   - **distributional indices of compositionality** that leverage the restricted lexical substitutability of idiom constituents

2. **Predicting human ratings on idiom syntactic flexibility** from the **indices in (1)** and **entropy-based indices of formal flexibility**

Idiomaticity and Compositionality

- **Idioms: non-compositional multiword expressions** (Nunberg et al. 1994; Sag et al. 2001; Cacciari 2014)
- **Lexical substitutability**
  - *to read a book* → *to read a novel*
  - *to spill the beans* → *to spill the peas* (just literal)
- **Systematicity** (Fodor & Lepore 2002)
  - If we can understand *drop the peas* and (literal) *spill the beans*, we can also understand *drop the beans* and *spill the peas*
  - This does not apply to idiomatic *spill the beans*

Idiom Type Identification: Previous Approaches

- **LIN 1999**; **FAZLY ET AL. 2009**
  - initial set of V-N pairs
  - generate lexical variants replacing the constituents with thesaurus synonyms
    - < *spill, bean* > → < *pour, bean* >, < *spill, corn* >, *etc.*
  - < *spill, bean* > labeled as non-compositional iff PMI(< *spill, bean* >) significantly different from PMI(< *pour, bean* >), PMI(< *spill, corn* >), etc.

Idiom Type Identification: Previous Approaches

- In **Distributional Semantic Models** (**DSMs**) target words and expressions are represented as **distributional vectors** in a **high-dimensionality space**
  - The vectors record the co-occurrence statistics of the targets with some contextual features
- Compositionality is assessed by measuring the **distributional similarity** between the **vector of a phrase** and the **vectors of its constituents** (BALDWIN ET AL. 2003; VENKATAPATHY & JOSHI 2005; FAZLY & STEVENSON 2008)

Our Proposal

for a target multi-token construction

1 FIND SYNONYMS

find the synonyms of the tokens that compose the construction

2 BUILD VARIANTS

build the lexical variants by combining the synonymic tokens

3 MEASURE SIMILARITY

measure the similarity between the lexical variants and the target construction

4 CLASSIFY

idioms are expected to be less similar to their variants

Our Proposal

tagliare la corda ('to flee', lit. 'to cut the rope')

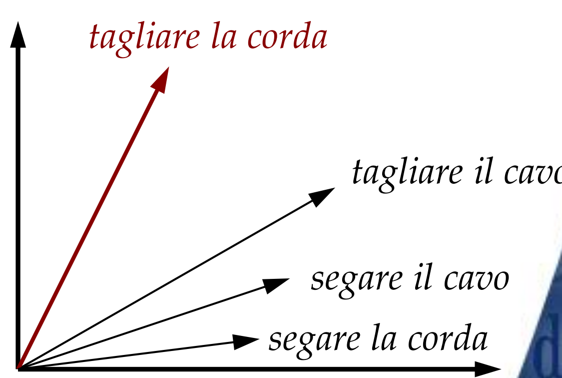
1 FIND SYNONYMS

tagliare → segare, recidere …
corda → cavo, fune …

BUILD VARIANTS 2

tagliare il cavo, segare il cavo, recidere il cavo, tagliare la fune, segare la fune, recidere la fune, segare la corda, recidere la corda …

3 MEASURE SIMILARITY



CLASSIFY 4

Our Proposal

tagliare la corda ('to flee', lit. 'to cut the rope')

1 FIND SYNONYMS

tagliare* → *segare, recidere …
corda* → *cavo, fune …

BUILD VARIANTS **2**

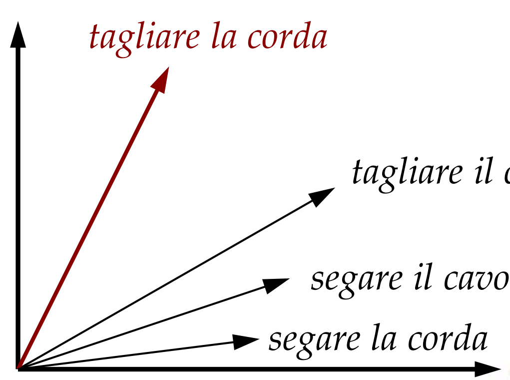
tagliare il cavo, segare il cavo, recidere il cavo, tagliare la fune, segare la fune, recidere la fune, segare la corda, recidere la

CLAS

VERY IDIOMATIC

1
2
3
4

3 MEASURE SIMILARITY



Our Proposal

scrivere un libro
('to write a book')

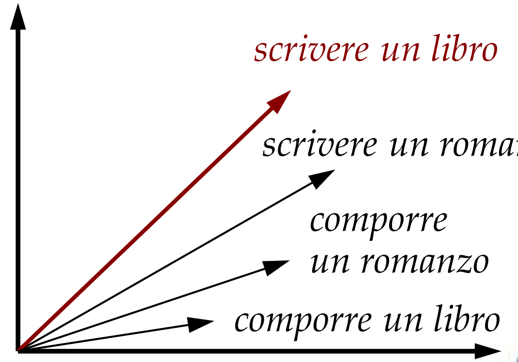
1 FIND SYNONYMS

scrivere → comporre, realizzare …
libro → romanzo …

BUILD VARIANTS 2

scrivere un libro, comporre un libro, scrivere un romanzo, comporre un romanzo ...

3 MEASURE SIMILARITY



CLASSIFY 4

Our Proposal

scrivere un libro
('to write a book')

1 FIND SYNONYMS

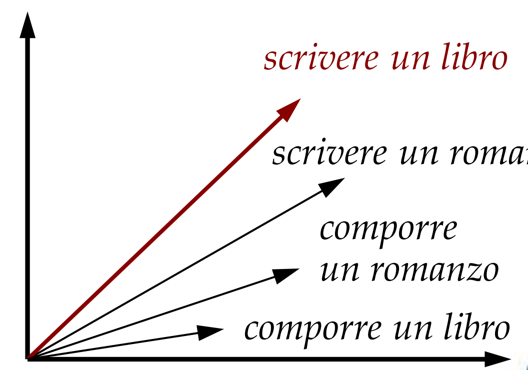
scrivere → comporre, realizzare …
libro → romanzo …

BUILD VARIANTS 2

scrivere un libro, comporre un libro, scrivere un romanzo, comporre un romanzo …

VERY LITERAL

3 MEASURE SIMILARITY



Our Targets

- **90 V-NP and V-PP constructions**
  - **45 idiomatic** constructions
    - » frequencies range from 364 (*ingannare il tempo* 'to while away the time') to 8294 (*andare in giro* 'to get about')
  - **45 compositional** constructions
    - » frequency-matched (e.g. *scrivere un libro* 'to write a book')
- **1-7 idiomaticity judgments** from 9 Linguistics students:
  - Krippendorf's $\alpha = 0.77$
  - Idioms obtained significantly higher ratings ($t=11.99$, $p < .001$)

Variant Extraction

- For both the verb and the noun of each target, **3, 4, 5 and 6 synonyms** were extracted from:
  - a Distributional Semantic Model (**DSM**):
    - » top cosine neighbors in a DSM built by looking at the [±2] content words linear context in the La Repubblica corpus (BARONI ET AL., 2004: 331M tokens)
  - Italian MultiWordNet lexicon (PIANTA ET AL., 2002: **iMWN**):
    - » candidates were **lemmas occurring in the same (manually selected) synsets** and **co-hyponyms**
    - » top 3, 4, 5 and 6 candidates filtered

1
2
3
4

Build Variants & Measure Similarity

- Potential variants for our targets were generated by combining:
  - noun synonyms with the original verb
    - » e.g. *tagliare la corda* → *tagliare il <u>cavo</u>, tagliare la <u>fune</u>*, etc.
  - verb synonyms with the original noun
    - » e.g. *tagliare la corda* → *<u>segare</u> la corda, <u>recidere</u> la corda*, etc.
  - verb synonyms with noun synonyms
    - » e.g. *tagliare la corda* → *<u>recidere</u> il <u>cavo</u>, <u>segare</u> la <u>fune</u>*, etc.
- A linear DSM from itWaC (BARONI ET AL. 2009; about 1,909M tokens) was built to represent both the targets and the variants that were found in the corpus as vectors
  - co-occurrences recorded how often each construction occurred in the same sentence with each of the 30,000 top content words

1
2
3
4

Compositionality Indices

- Compositionality indices were built in four different ways:
  - **Mean** - mean cosine similarity between the target and its variants
  - **Max** - maximum cosine between the target and its variants
  - **Min** - minimum cosine between the target and its variants
  - **Centroid** – cosine between the target and the centroid of its variants
- We tried keeping **15, 24, 35** and **48 variants per target**
- Variants missing from itWaC were treated in two ways:
  - **no** models - they are ignored
  - **orth** models - encoded as vectors orthogonal to the targets

Evaluation

- Our targets were sorted in ascending order according to each of the four indices
- Idioms (our positives) expected to occur at the top of the ranking
  - **Spearman's ρ correlation** with our idiomaticity judgements
  - Interpolated Average Precision (**IAP**): the average Interpolated Precision at recall levels of 20%, 50% and 80% (following FAZLY ET AL., 2009)
  - **F-measure** at the median

Parameters

Parameter	Values
Variants source	DSM, iMWN
Variants filter	cosine (DSM, iMWN) raw frequency (iMWN)
Variants per target	15, 24, 35, 48
Non-attested variants	not considered (no) orthogonal vectors (orth)
Measures	Mean, Max, Min, Centroid

- **96 models** resulting from the combinations of all the possibile values for all the parameters

Top IAP, F and ρ models

Top IAP Models	**IAP**	**F**	**ρ**
$iMWN_{cos}$ 15_{var} $Centroid_{no}$	**.91**	.80	-.58***
$iMWN_{cos}$ 24_{var} $Centroid_{no}$	**.91**	.78	-.62***
$iMWN_{cos}$ 35_{var} $Centroid_{no}$	**.91**	.82	-.60***
DSM 48_{var} $Centroid_{no}$	**.89**	.82	-.64***
DSM 48_{var} $Centroid_{orth}$	**.89**	.82	-.60***
Top F-measure Models	**IAP**	**F**	**ρ**
$iMWN_{cos}$ 35_{var} $Centroid_{no}$	.91	**.82**	-.60***
DSM 48_{var} $Centroid_{no}$	.89	**.82**	-.64***
DSM 48_{var} $Centroid_{orth}$	.89	**.82**	-.60***
$iMWN_{cos}$ 15_{var} $Centroid_{no}$	.91	**.80**	-.58***
DSM 24_{var} $Centroid_{no}$	.89	**.80**	-.60***
Top ρ Models	**IAP**	**F**	**ρ**
$iMWN_{cos}$ 48_{var} $Centroid_{orth}$	.86	.80	**-.67*** **
$iMWN_{cos}$ 35_{var} $Centroid_{orth}$	.72	.44	**-.66*** **
$iMWN_{cos}$ 24_{var} $Centroid_{orth}$	.85	.78	**-.66*** **
$iMWN_{cos}$ 15_{var} $Centroid_{orth}$	.88	.80	**-.65*** **
$iMWN_{freq}$ 15_{var} $Centroid_{orth}$	.66	.51	**-.65*** **
Random	.55	.51	.05

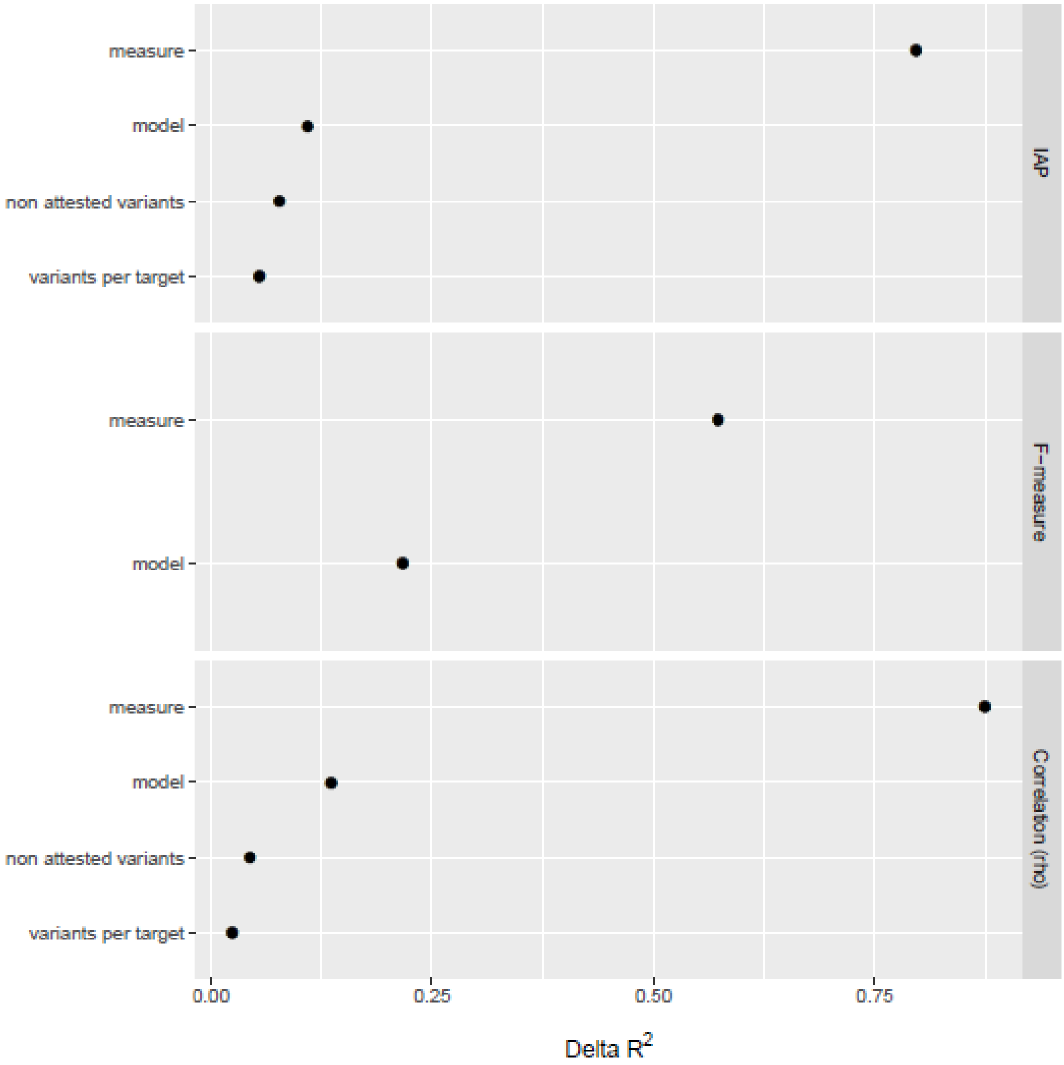
Influence of Parameters on Performance

- Linear regressions to assess the influence of the parameter settings on the performances of our models (cf. LAPESA & EVERT 2014)
- **Predictors**: parameter settings
- **Dependent variables**: IAP, F-measure and ρ of our models

Model	Adjusted R^2
IAP	0.90
F-measure	0.52
ρ	0.94

Parameters and Feature Ablation

(***model*** = variants source + variants filter)



Extending our Approach to Adj-N Combinations

- **13 idiomatic** (*alte sfere* 'high places') + **13 frequency-matched literal** targets (*nuova legge* 'new law')
- Variants also from a **Structured DSM** (co-occurrences like $<w_1, r, w_2>$)
- **Mean**, **Max**, **Min** and **Centroid** compared to **reference indices**:
  - **Additive** model: the similarity between the target and the sum of the vectors of its components (see KRČMÁŘ ET AL., 2013)
  - **Multiplicative** model: the similarity between the target and the product of the vectors of its components (see KRČMÁŘ ET AL., 2013)

Adjective-Noun Pairs: Best Models

Top IAP Models	IAP	F	ϱ
Additive	.85	.77	-.62***
Structured DSM Mean_{orth}	.84	.85	-.68***
iMWN_{syn} Centroid_{orth}	.83	.85	-.57**
iMWN_{ant} Centroid_{orth}	.83	.77	-.52**
iMWN_{ant} Mean_{orth}	.83	.69	-.64***
Top F-measure Models	**IAP**	**F**	**ϱ**
Structured DSM Mean_{orth}	.84	.85	-.68***
iMWN_{syn} Centroid_{orth}	.83	.85	-.57**
Additive	.85	.77	-.62***
iMWN_{ant} Centroid_{orth}	.83	.77	-.52**
iMWN_{syn} Centroid_{no}	.82	.77	-.57**
Top ϱ Models	**IAP**	**F**	**ϱ**
Structured DSM Mean_{orth}	.84	.85	-.68***
Linear DSM Mean_{orth}	.75	.69	-.66***
iMWN_{syn} Mean_{orth}	.77	.77	-.65***
iMWN_{syn} Mean_{no}	.70	.69	-.65***
iMWN_{ant} Mean_{orth}	.83	.69	-.64***
Multiplicative	**.58**	**.46**	**.03**
Random	**.55**	**.51**	**.05**

Interim conclusions

- variant-based distributional indices are effective for idiom type identification
- **Centroid** and **Mean** perform the best
- **DSM variants comparable to iMWN** but less time-consuming!
- most best models for **Adj-N idioms** are ***orth*** ≠ **V-N idioms**
- **additive** model performs comparably
- **product** comparable to random baseline

Summary

1. **Idiom type identification task** on **90 Italian V-N** combinations and **26 Italian Adj-N** combinations

   - **distributional indices of compositionality** that leverage the restricted lexical substitutability of idiom constituents

2. **Predicting human ratings on idiom syntactic flexibility** from the **indices in (1)** and **entropy-based indices of formal flexibility**

Our Dataset

- **54 Italian V-NP** and **V-PP idioms**
  - e.g. *tagliare la corda* (‘to flee’, lit. ‘to cut the rope’)

    cadere dal cielo (‘to be heaven-sent’, lit. ‘to fall from the sky’)
  - **frequency > 75 tokens** in ‘La Repubblica’
- **54 Italian V-NP** and **V-PP literals**
  - e.g. *leggere un libro* (‘to read a book’)

Syntactic Flexibility Judgments on CrowdFlower

- For each idiom and literal, **5 sentences** were created

1) **base form**

   Pietro alza il gomito quando va a cena da Teresa.

   «Pietro raises the elbow when he has dinner at Teresa's»

2) **adverb insertion**

   Pietro alza sempre il gomito quando va a cena da Teresa.

   «Pietro always raises the elbow when he has dinner at Teresa's»

3) **adjective insertion**

   Pietro alzò il solito gomito quando andò a cena da Teresa.

   «Pietro raised the usual elbow when he had dinner at Teresa's.»

4) **left dislocation**

   Il gomito Pietro lo alza quando esce con Giovanni

   «The elbow Pietro raises it when he goes out with Giovanni.»

5) **wh-movement**

   Che gomito ha alzato Pietro quando è andato alla festa di Teresa?

   «Which elbow did Pietro raise when he went to Teresa's party?»

Syntactic Flexibility Judgments on CrowdFlower

- **1-7 acceptability judgments**
  - Each sentence rated by 20 contributors

	Idioms Avg.	Literals Avg.	t-test
Base form	6.31	6.40	$p = 0.32$
Adverb	6.22	6.21	$p = 0.68$
Adjective	5.00	6.02	$p < 0.05$
Left Dislocation	4.09	4.71	$p < 0.001$
Wh-movement	3.11	4.31	$p < 0.001$

- Overarching **SYNTACTIC FLEXIBILITY** index
  - average of the differences between the mean acceptability of each variant and the mean acceptability of the base form

Measuring Formal Flexibility with Shannon Entropy

- **SHANNON** (1948) **Entropy** measures the average degree of uncertainty in a random variable ***X***

$$H(X) = \sum_{x \in X} p(x) log \frac{1}{p(x)}$$

  - Each $x \in X$ represents a state of the system
  - **The higher the entropy, the more unpredictable the outcome of the random system**

Measuring Formal Flexibility with Shannon Entropy

1. **LEXICAL VARIABILITY** of the free slot (e.g. *to cast a shadow on the problem*, *to cast a shadow on the institution*, etc.)
2. **MORPHOLOGY** of the arguments and the verb (e.g. *to cast a shadow-S*, *to cast many shadows-P*, etc.)
3. **ARTICLES** variability (e.g. *to cast a shadow*, *to cast Ø shadows*, etc.)
4. **LINEAR ORDER** of the constituents (e.g. *to bring a project to light*, *to bring to light a project*, etc.)
5. **TOKEN DISTANCE** of the arguments from the verb (e.g. *to cast a shadow* (1), *to cast a big shadow* (2), etc.)
6. Presence of **INTERVENING ADJECTIVES**, **PPS** and **ADVERBS** (e.g. *to cast a big shadow*, *to cast a huge shadow*, etc.)
7. The **SYNTACTIC FRAME** it occurs in (e.g. *to open the floodgates to*, *to open the floodgates for*, etc.)

Measuring Formal Flexibility with Shannon Entropy

- **Lexical entropy** (e.g. *to cast a shadow on X*)

$$H(X) = \sum_{x \in X} p(x) log \frac{1}{p(x)}$$

  - each *x* represents a possible lemma
  - e.g. *to cast a shadow on* X $\rightarrow$ x_1 = institution, x_2 = project, x_3 = problem, etc.
  - the higher the entropic value, the more lexically variable the free slot is and vice versa

Measuring Formal Flexibility with Shannon Entropy

- MORPHOLOGICAL ENTROPY of the arguments
  - x_1 = *to cast a <u>shadow (SING.)</u> on*

    x_2 = *to cast <u>shadows (PLUR.)</u> on,* etc.
- ARTICLES ENTROPY
  - x_1 = *to cast <u>a (IND)</u> shadow on*

    x_2 = *to cast <u>the (DEF)</u> shadow on*

    x_3 = *to cast <u>(∅)</u> shadows on*
- Etc.

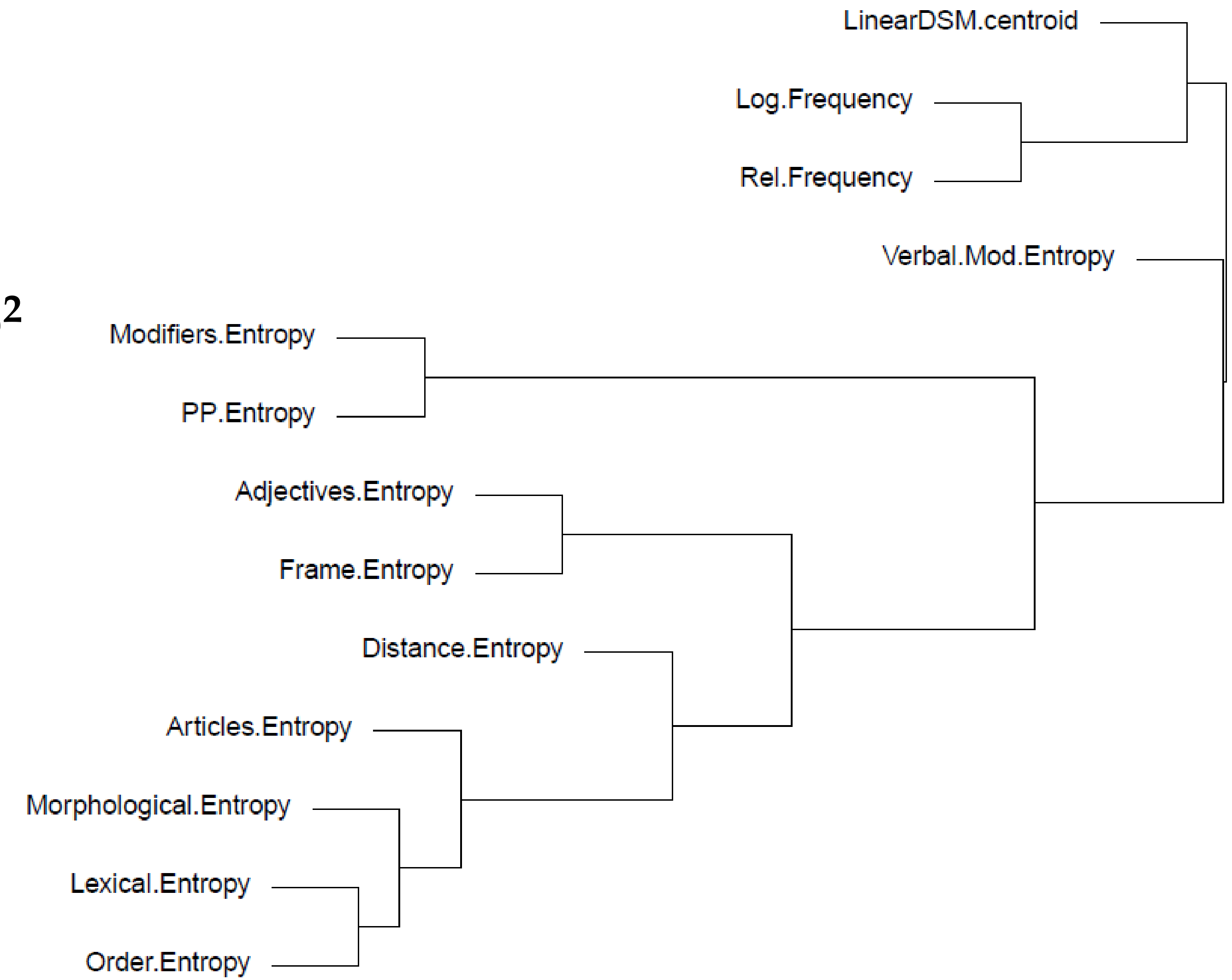
Regression analysis on the acceptability ratings

- **PREDICTORS**
  1. **Entropies** (lexical, morphological, order, token distance, articles, adjectives and PPs, frame)
  2. **DSM Centroid** (the best performing one)
  3. **Log frequency** and **relative frequency**
- **DEPENDENT VARIABLE**
  1. **Syntactic flexibility** judgments

Correlational structure of the predictors

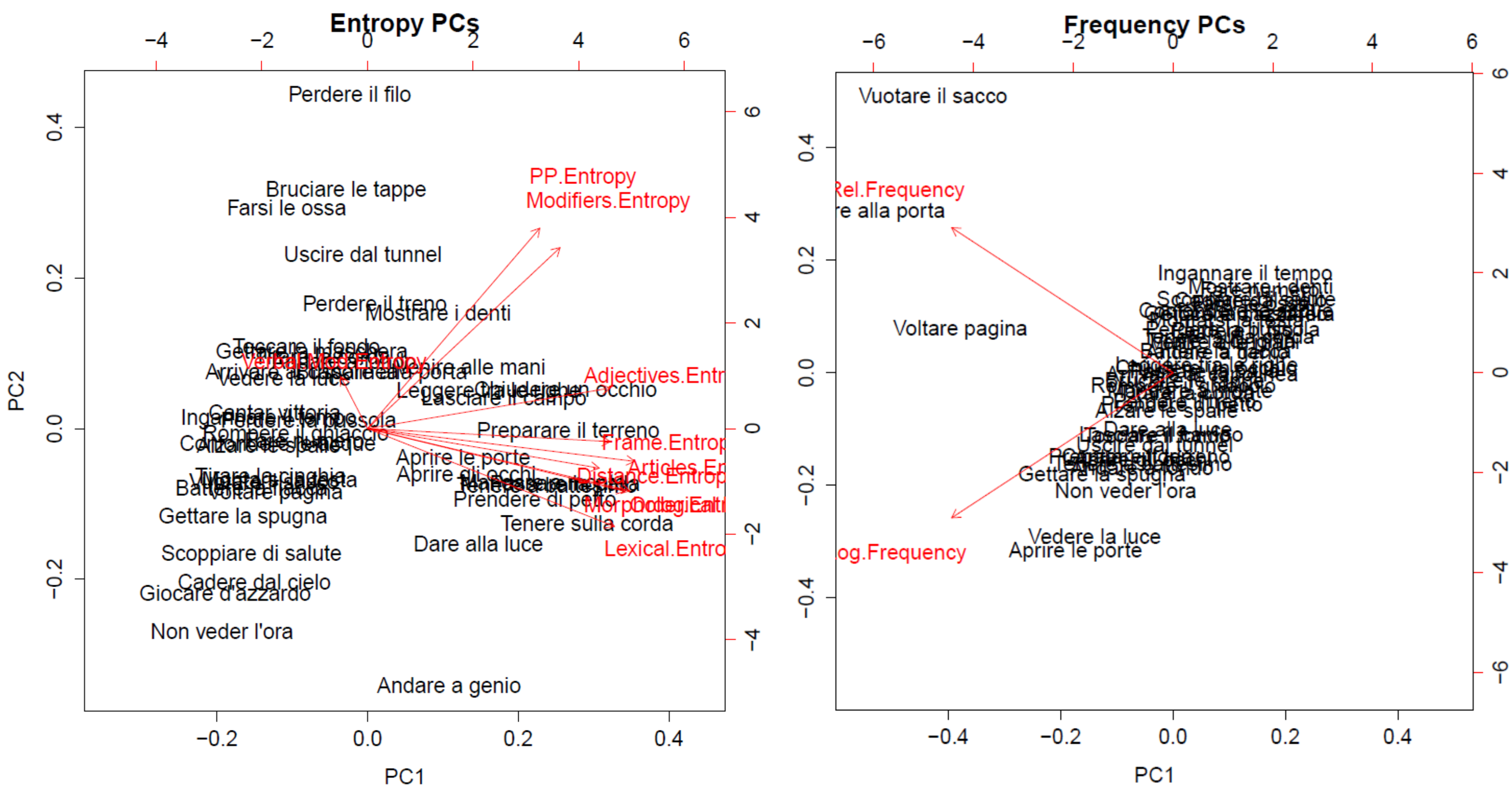
Metric:

Spearman's ρ^2



Principal Component Analysis (PCA) on our predictors

- Condition number (**k**) = **49.11** (**high collinearity**)

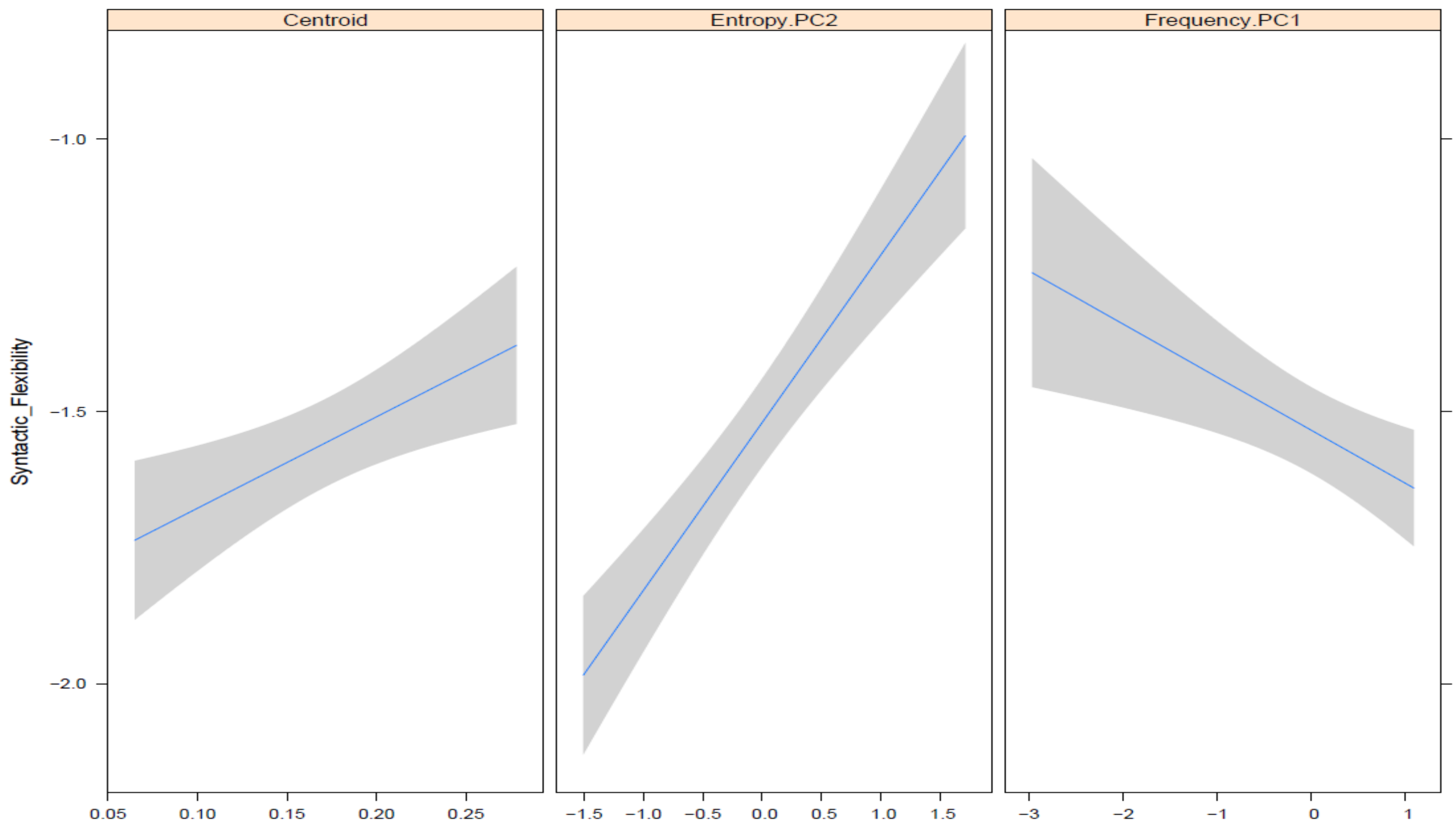


Regression on the syntactic flexibility judgments

Predictors	β	S.E.	t	p
Intercept	-1.81	0.11	-16.69	< 0.001
Centroid	1.83	0.58	3.14	< 0.01
Entropy PC1	-0.01	0.02	-0.94	n.s.
Entropy PC2	0.30	0.04	7.27	< 0.001
Frequency PC1	-0.10	0.03	-2.30	< 0.01

Best fitting model: **adjusted $R^2 = 0.67$, $F(4, 36) = 21.17$, $p < 0.001$**

Partial Effects
(*Centroid, Entropy PC2, Frequency PC1*)



Conclusions

- The best model consisted in a linear combination of **all our predictors**
  - **Entropy**: the more an expression formally varied in the corpus, the more the subjects perceived it to be flexible
  - **Distributional Centroid**: cfr. GIBBS & NAYAK (1989)
  - **Frequency**: more frequent expressions are perceived as less flexible
- Future directions of research
  - model other kinds of psycholinguistic data on idiom variation processing (e.g. **eye-tracking data**)

Thank you
for your attention!

University of Pisa

Computational Linguistics Lab

http://colinglab.humnet.unipi.it