Exploring Idiomaticity with Variant-based Distributional Measures and Shannon Entropy

Marco S. G. Senaldi
Gianluca E. Lebani
Alessandro Lenci

1 Scuola Normale Superiore, Pisa
2 University of Pisa
DGfS 2017 – Saarbrücken | 9th March 2017
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
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**
• **Idioms:** non-compositional multiword expressions (Nunberg et al. 1994; Sag et al. 2001; Cacciari 2014)

• **Lexical substitutability**
  - *to read a book* $\rightarrow$ *to read a novel*
  - *to spill the beans* $\rightarrow$ *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
    - \(<\text{spill, bean}> \rightarrow <\text{pour, bean}>, <\text{spill, corn}>, \text{etc.}\)
  - \(<\text{spill, bean}>> labeled as non-compositional iff $\text{PMI}(<\text{spill, bean}>)$ significantly different from $\text{PMI}(<\text{pour, bean}>)$, $\text{PMI}(<\text{spill, corn}>)$, 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

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

**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 ...

**Classify** 4

**Find Synonyms** 1

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

**Measure Similarity** 3

tagliare la corda
tagliare il cavo
segare il cavo
segare la corda

*tagliare la corda* (‘to flee’, lit. ‘to cut the rope’)*
Our Proposal

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

Find Synonyms

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

Build Variants

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

Measure Similarity

tagliare la corda
tagliare il cavo
segare il cavo
segare la corda

Very Idiomatic
Our Proposal

**1. Find Synonyms**

- `scrivere` → `comporre, realizzare` ...
- `libro` → `romanzo` ...

**2. Build Variants**

- `scrivere un libro, comporre un libro, scrivere un romanzo, comporre un romanzo` ...

**3. Measure Similarity**

- `scrivere un libro` → `scrivere un romanzo`
- `comporre un libro` → `comporre un romanzo`
Our Proposal

**Build Variants**

`scrivere un libro` ('to write a book')

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

**Find Synonyms**

`scrivere → comporre, realizzare ...`

`libro → romanzo ...`

**Measure Similarity**

`scrivere un libro`

`scrivere un romanzo`

`comporre un romanzo`

`comporre un libro`

**Very Literal**
• 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 ($t=11.99, p < .001$)
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 \([-2, 2]\) 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
Potential variants for our targets were generated by combining:

- noun synonyms with the original verb
  - e.g. *tagliare la corda* → *tagliare il cavo, tagliare la fune*, etc.

- verb synonyms with the original noun
  - e.g. *tagliare la corda* → *segare la corda, recidere la corda*, etc.

- verb synonyms with noun synonyms
  - e.g. *tagliare la corda* → *recidere il cavo, segare la fune*, 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
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
• 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

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variants source</td>
<td>DSM, iMWN</td>
</tr>
<tr>
<td>Variants filter</td>
<td>cosine (DSM, iMWN)</td>
</tr>
<tr>
<td></td>
<td>raw frequency (iMWN)</td>
</tr>
<tr>
<td>Variants per target</td>
<td>15, 24, 35, 48</td>
</tr>
<tr>
<td>Non-attested variants</td>
<td>not considered (no)</td>
</tr>
<tr>
<td></td>
<td>orthogonal vectors (orth)</td>
</tr>
<tr>
<td>Measures</td>
<td>Mean, Max, Min, Centroid</td>
</tr>
</tbody>
</table>

- **96 models** resulting from the combinations of all the possible values for all the parameters
## Top IAP, F and ρ models

<table>
<thead>
<tr>
<th>Top IAP Models</th>
<th>IAP</th>
<th>F</th>
<th>ρ</th>
</tr>
</thead>
<tbody>
<tr>
<td>iMWN$<em>{\cos}$ 15$</em>{\text{var}}$ Centroid$_{\text{no}}$</td>
<td>.91</td>
<td>.80</td>
<td>-58***</td>
</tr>
<tr>
<td>iMWN$<em>{\cos}$ 24$</em>{\text{var}}$ Centroid$_{\text{no}}$</td>
<td>.91</td>
<td>.78</td>
<td>-62***</td>
</tr>
<tr>
<td>iMWN$<em>{\cos}$ 35$</em>{\text{var}}$ Centroid$_{\text{no}}$</td>
<td>.91</td>
<td>.82</td>
<td>-60***</td>
</tr>
<tr>
<td>DSM 48$<em>{\text{var}}$ Centroid$</em>{\text{no}}$</td>
<td>.89</td>
<td>.82</td>
<td>-64***</td>
</tr>
<tr>
<td>Top F-measure Models</td>
<td>IAP</td>
<td>F</td>
<td>ρ</td>
</tr>
<tr>
<td>----------------</td>
<td>-----</td>
<td>---</td>
<td>----</td>
</tr>
<tr>
<td>iMWN$<em>{\cos}$ 35$</em>{\text{var}}$ Centroid$_{\text{no}}$</td>
<td>.91</td>
<td>.82</td>
<td>-60***</td>
</tr>
<tr>
<td>DSM 48$<em>{\text{var}}$ Centroid$</em>{\text{no}}$</td>
<td>.89</td>
<td>.82</td>
<td>-64***</td>
</tr>
<tr>
<td>DSM 48$<em>{\text{var}}$ Centroid$</em>{\text{orth}}$</td>
<td>.89</td>
<td>.82</td>
<td>-60***</td>
</tr>
<tr>
<td>iMWN$<em>{\cos}$ 15$</em>{\text{var}}$ Centroid$_{\text{no}}$</td>
<td>.91</td>
<td>.80</td>
<td>-58***</td>
</tr>
<tr>
<td>DSM 24$<em>{\text{var}}$ Centroid$</em>{\text{no}}$</td>
<td>.89</td>
<td>.80</td>
<td>-60***</td>
</tr>
<tr>
<td>Top ρ Models</td>
<td>IAP</td>
<td>F</td>
<td>ρ</td>
</tr>
<tr>
<td>---------------</td>
<td>-----</td>
<td>---</td>
<td>----</td>
</tr>
<tr>
<td>iMWN$<em>{\cos}$ 48$</em>{\text{var}}$ Centroid$_{\text{orth}}$</td>
<td>.86</td>
<td>.80</td>
<td>-67***</td>
</tr>
<tr>
<td>iMWN$<em>{\cos}$ 35$</em>{\text{var}}$ Centroid$_{\text{orth}}$</td>
<td>.72</td>
<td>.44</td>
<td>-66***</td>
</tr>
<tr>
<td>iMWN$<em>{\cos}$ 24$</em>{\text{var}}$ Centroid$_{\text{orth}}$</td>
<td>.85</td>
<td>.78</td>
<td>-66***</td>
</tr>
<tr>
<td>iMWN$<em>{\cos}$ 15$</em>{\text{var}}$ Centroid$_{\text{orth}}$</td>
<td>.88</td>
<td>.80</td>
<td>-65***</td>
</tr>
<tr>
<td>iMWN$<em>{\text{freq}}$ 15$</em>{\text{var}}$ Centroid$_{\text{orth}}$</td>
<td>.66</td>
<td>.51</td>
<td>-65***</td>
</tr>
<tr>
<td>Random</td>
<td>.55</td>
<td>.51</td>
<td>.05</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Model</th>
<th>Adjusted $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>IAP</td>
<td>0.90</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.52</td>
</tr>
<tr>
<td>ρ</td>
<td>0.94</td>
</tr>
</tbody>
</table>
Parameters and Feature Ablation

\[(\text{model} = \text{variants source} + \text{variants filter})\]
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)
<table>
<thead>
<tr>
<th>Top IAP Models</th>
<th>IAP</th>
<th>F</th>
<th>q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additive</td>
<td>.85</td>
<td>.77</td>
<td>-.62***</td>
</tr>
<tr>
<td>iMWN&lt;sub&gt;syn&lt;/sub&gt; Centroid&lt;sub&gt;orth&lt;/sub&gt;</td>
<td>.83</td>
<td>.85</td>
<td>-.57**</td>
</tr>
<tr>
<td>iMWN&lt;sub&gt;ant&lt;/sub&gt; Centroid&lt;sub&gt;orth&lt;/sub&gt;</td>
<td>.83</td>
<td>.77</td>
<td>-.52**</td>
</tr>
<tr>
<td>iMWN&lt;sub&gt;ant&lt;/sub&gt; Mean&lt;sub&gt;orth&lt;/sub&gt;</td>
<td>.83</td>
<td>.69</td>
<td>-.64***</td>
</tr>
<tr>
<td>Top F-measure Models</td>
<td>IAP</td>
<td>F</td>
<td>q</td>
</tr>
<tr>
<td>Structured DSM Mean&lt;sub&gt;orth&lt;/sub&gt;</td>
<td>.84</td>
<td>.85</td>
<td>-.68***</td>
</tr>
<tr>
<td>iMWN&lt;sub&gt;syn&lt;/sub&gt; Centroid&lt;sub&gt;orth&lt;/sub&gt;</td>
<td>.83</td>
<td>.85</td>
<td>-.57**</td>
</tr>
<tr>
<td>Additive</td>
<td>.85</td>
<td>.77</td>
<td>-.62***</td>
</tr>
<tr>
<td>iMWN&lt;sub&gt;ant&lt;/sub&gt; Centroid&lt;sub&gt;orth&lt;/sub&gt;</td>
<td>.83</td>
<td>.77</td>
<td>-.52**</td>
</tr>
<tr>
<td>iMWN&lt;sub&gt;syn&lt;/sub&gt; Centroid&lt;sub&gt;no&lt;/sub&gt;</td>
<td>.82</td>
<td>.77</td>
<td>-.57**</td>
</tr>
<tr>
<td>Top q Models</td>
<td>IAP</td>
<td>F</td>
<td>q</td>
</tr>
<tr>
<td>Structured DSM Mean&lt;sub&gt;orth&lt;/sub&gt;</td>
<td>.84</td>
<td>.85</td>
<td>-.68***</td>
</tr>
<tr>
<td>Linear DSM Mean&lt;sub&gt;orth&lt;/sub&gt;</td>
<td>.75</td>
<td>.69</td>
<td>-.66***</td>
</tr>
<tr>
<td>iMWN&lt;sub&gt;syn&lt;/sub&gt; Mean&lt;sub&gt;orth&lt;/sub&gt;</td>
<td>.77</td>
<td>.77</td>
<td>-.65***</td>
</tr>
<tr>
<td>iMWN&lt;sub&gt;syn&lt;/sub&gt; Mean&lt;sub&gt;no&lt;/sub&gt;</td>
<td>.70</td>
<td>.69</td>
<td>-.65***</td>
</tr>
<tr>
<td>iMWN&lt;sub&gt;ant&lt;/sub&gt; Mean&lt;sub&gt;orth&lt;/sub&gt;</td>
<td>.83</td>
<td>.69</td>
<td>-.64***</td>
</tr>
<tr>
<td>Multiplicative</td>
<td>.58</td>
<td>.46</td>
<td>.03</td>
</tr>
<tr>
<td>Random</td>
<td>.55</td>
<td>.51</td>
<td>.05</td>
</tr>
</tbody>
</table>
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
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
• 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’)

Our Dataset
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?»
1-7 acceptability judgments

- Each sentence rated by 20 contributors

<table>
<thead>
<tr>
<th></th>
<th>Idioms Avg.</th>
<th>Literals Avg.</th>
<th>t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base form</td>
<td>6.31</td>
<td>6.40</td>
<td>p = 0.32</td>
</tr>
<tr>
<td>Adverb</td>
<td>6.22</td>
<td>6.21</td>
<td>p = 0.68</td>
</tr>
<tr>
<td>Adjective</td>
<td>5.00</td>
<td>6.02</td>
<td>p &lt; 0.05</td>
</tr>
<tr>
<td>Left Dislocation</td>
<td>4.09</td>
<td>4.71</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>Wh-movement</td>
<td>3.11</td>
<td>4.31</td>
<td>p &lt; 0.001</td>
</tr>
</tbody>
</table>

Overarching **Syntactic Flexibility** index

- average of the differences between the mean acceptability of each variant and the mean acceptability of the base form
• **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.)
• **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* → \( x_1 = \) institution, \( x_2 = \) project, \( x_3 = \) problem, etc.
- the higher the entropic value, the more lexically variable the free slot is and vice versa
• **Morphological entropy** of the arguments
  
  \[ x_1 = \text{to cast a shadow (SING.) on} \]
  
  \[ x_2 = \text{to cast shadows (PLUR.) on, etc.} \]

• **Articles entropy**
  
  \[ x_1 = \text{to cast a (IND) shadow on} \]
  
  \[ x_2 = \text{to cast the (DEF) shadow on} \]
  
  \[ x_3 = \text{to cast (Ø) shadows on} \]

• 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 $q^2$
Principal Component Analysis (PCA) on our predictors

- Condition number \((k) = 49.11\) (high collinearity)
Regression on the syntactic flexibility judgments

<table>
<thead>
<tr>
<th>Predictors</th>
<th>β</th>
<th>S.E.</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.81</td>
<td>0.11</td>
<td>-16.69</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td><strong>Centroid</strong></td>
<td>1.83</td>
<td>0.58</td>
<td>3.14</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Entropy PC1</td>
<td>-0.01</td>
<td>0.02</td>
<td>-0.94</td>
<td>n.s.</td>
</tr>
<tr>
<td><strong>Entropy PC2</strong></td>
<td>0.30</td>
<td>0.04</td>
<td>7.27</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Frequency PC1</td>
<td>-0.10</td>
<td>0.03</td>
<td>-2.30</td>
<td>&lt; 0.01</td>
</tr>
</tbody>
</table>

Best fitting model: \( \text{adjusted } R^2 = 0.67, F (4, 36) = 21.17, p < 0.001 \)
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!