Quantifying intra-linguistic diversity: Case study of multiword expressions

Agata Savary, Yağmur Öztürk, Adam Lion-Bouton

1. Objectives

Quantifying intra-linguistic **diversity** of ***multiword expression*** (**MWE**) in annotated text.

↳ How can diversity be measured (**2.**)

↳ validating (**3.**) diversity measures

2. Diversity

In a population of MWE occurrences I and MWE types T

$$I \xrightarrow{\tau} T$$

Never gonna **give** you **up**	↦	{give, up}
They **gave up** long ago	↦	{give, up}
Who r u, who r so **wise in the ways of** science	↦	{in, of, the, way, wise}
Another one **bites the dust**	↦	{bite, dust, the}

Diversity → decomposed into:

- ***Variety*** : How many types

  ↳ *Richness*: $|\{\tau(i)|i \in I\}| = 3$

  ↳ *Normalized Richness*: $\frac{|\{\tau(i)|i \in I\}|}{|I|} = \frac{3}{4}$

- ***Balance*** : How balanced are types frequencies (f)

  With $E_{a,b} = \frac{N_a}{N_b}$, and $N_a = (\sum_{t \in T} f(t)^a)^{\frac{1}{(1-a)}}$

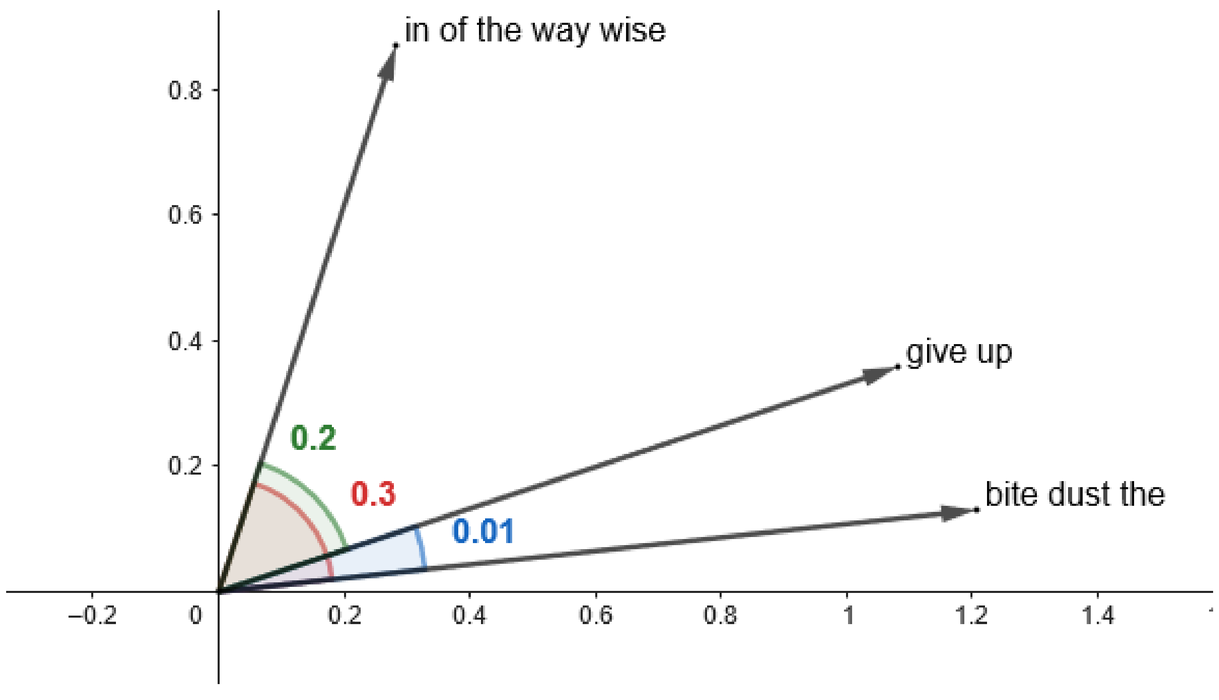
  ↳ $E_{1,0} \approx \frac{3}{2.58} \approx 0.86$

  ↳ $E_{2,1} \approx \frac{2.58}{2.27} \approx 0.88$

- **Disparity** : How different are types from each other

  Disparity are based on distance-like measure

  Here a semantic space with: $\mathrm{d}(u,v) = (1 - \cos(u,v)) \div 2$

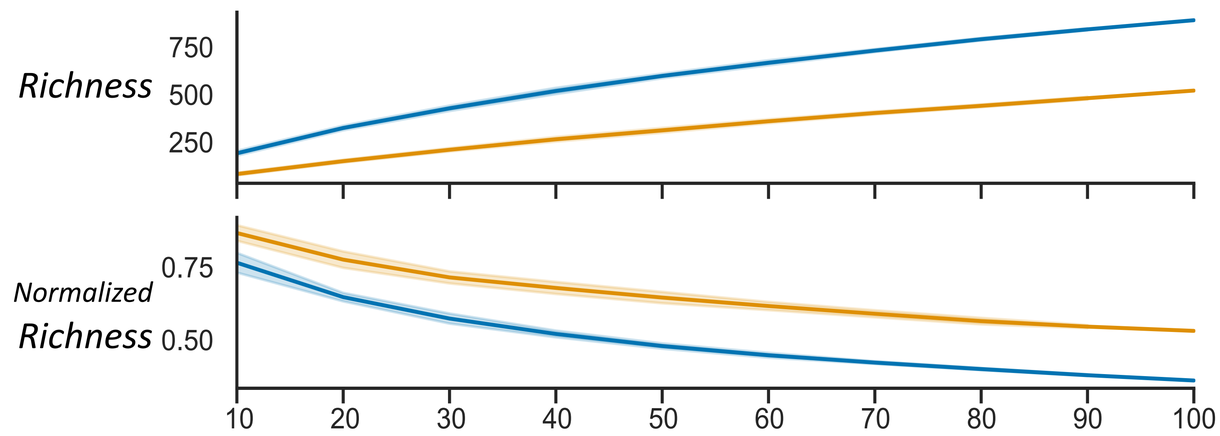


Pretend semantic space of T

↳ $\mathrm{D} = \frac{\sum_{(u\,v) \in T^2, u \neq v} d(u,v)}{|T|^2 - |T|} = \frac{2 \cdot 0.2 \cdot 0.3 \cdot 0.01}{6} = 0.17$

3. Impact of corpus size

Variety



Variety measures of **verbal** and **non-verbal** MWE in terms of sample size (% of sentences) of *Sequoia* Corpus

Both *variety* measures are affected by corpus size ⇒ **corpus of different size cannot be compared.**

Richness more interpretable ⇒ **Richness preferred**

Balance

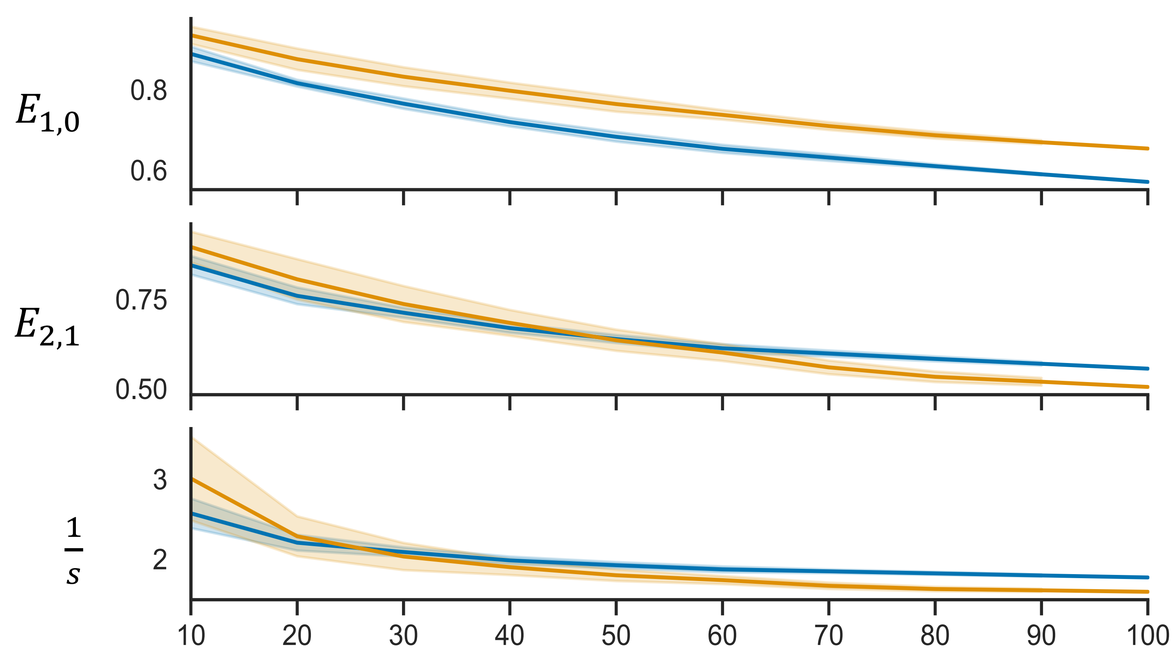
Hypothesis: MWEs follow a Zipfian distribution (of pmf $Z_{s,N}(x)$)

$$Z_{N,s}(x) = \frac{1}{x^s \sum_{n=1}^{N} n^{-s}}$$

N = number of types ↲ ↳ s → distribution's curvature

⇒ $\frac{1}{s}$ **acts as a balance measure** ⇒ Balance measures should act as $\frac{1}{s}$

We fit s on our samples using LSE and compare $E_{1,0}$ and $E_{2,1}$ to $\frac{1}{s}$



Balance measures of **verbal** and **non-verbal** MWE in terms of sample size (% of sentences) of Sequoia Corpus

$E_{1,0}$ → **verbal** MWEs **more** balanced than **non-verbal** MWE.

$E_{2,1}$ and $\frac{1}{s}$ → **verbal** MWEs **more** balanced than **non-verbal** MWEs on **small samples**, and **less** balanced on **large sample**.

$E_{2,1}$ acts more like $\frac{1}{s}$ than $E_{1,0}$ ⇒ $E_{2,1}$ **preferred**

Disparity

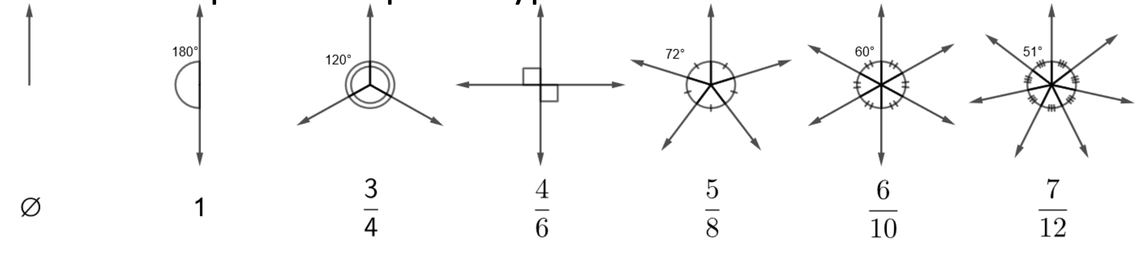
Early experiments → D **decreases** with corpus size

In a closed space $\left(\text{e.g. } \frac{1-\cos(i,j)}{2}\right)$:

↳ Large $|T|$ ⇒ Dense concentration of types

↳ High density ⇒ $\max D$ decreases ⇒ D tends to decease

$\max D$ in a 2D space for up to 7 types



université PARIS-SACLAY
UGA Université Grenoble Alpes
université de TOURS