SAPIENZA
UNIVERSITÀ DI ROMA

THE UNIVERSITY
of EDINBURGH

Scan pattern similarity predicts the semantic similarity of sentences across languages above and beyond their syntactic structure

Moreno I. Coco, Eunice G. Fernandes,
Manabu Arai, and **Frank Keller**

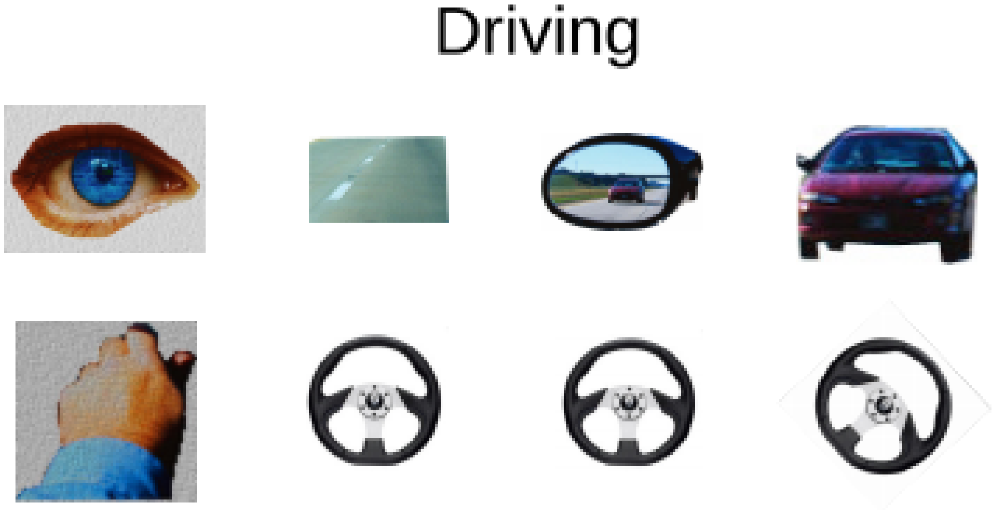
Overview

- Human **cognition** is a highly integrated system: **processes and representations** are multimodal.
- The words we utter to describe a visual scene are grounded in the objects we attend.
- **Visual attention and language production** correlate across modalities (Coco & Keller, 2012).
- We can predict what you will say based on where you look.

Research Questions
Does this effect generalize across different languages?
Is the correlation driven by semantics or syntax?
What does this tell us about visual grounding?

Background

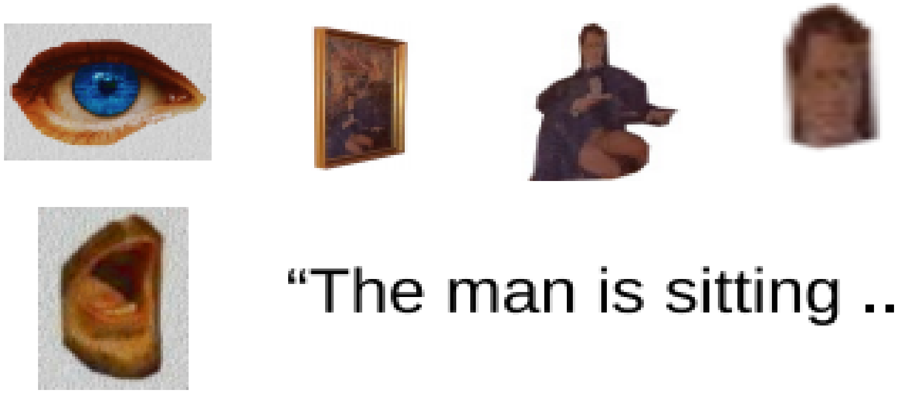
Many cognitive processes are multimodal:



Background

Many cognitive processes are multimodal:

Describing



“The man is sitting ...”

Background

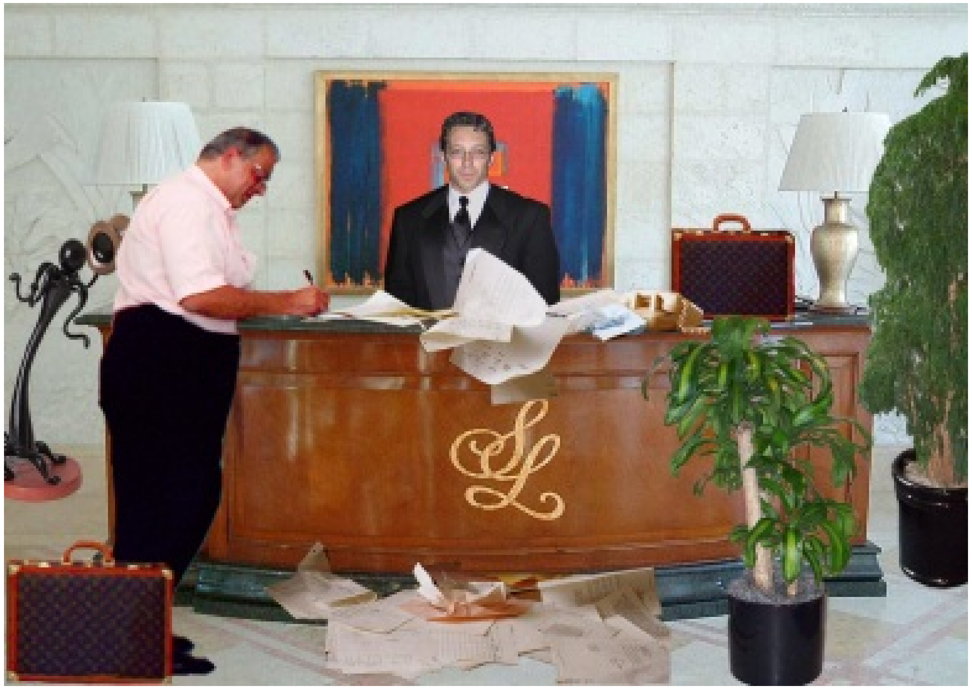
For many everyday tasks, processing streams in **multiple modalities** need to be coordinated:

- Motor tasks such as tea-making or driving (Land et al. 1999);
- Dialogue and collaborative problem solving (Coco et al. 2018);
- Language comprehension and production (Griffin & Bock 2000).

Here, we will focus on the coordination of **visual and linguistic processing.**

Scan patterns

In this work, we will use picture description to study multimodal processing:

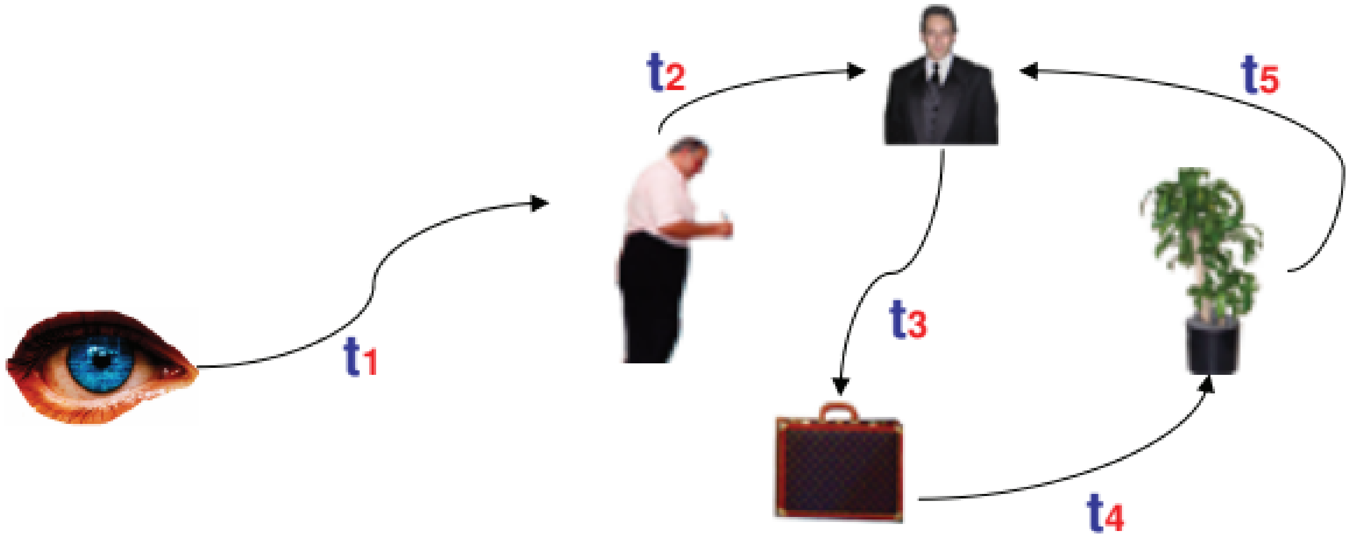


“The suitcase is on the counter next to the man.”

When describing an image, a participant follows a scan pattern while a uttering a sentence.

Scan patterns

When viewing a scene, participants generate scan patterns – sequences of fixated objects:



Sentences are sequences of words:
the, suitcase, is, on, the, counter, next, to, the, man

How does multimodal coordination happen?

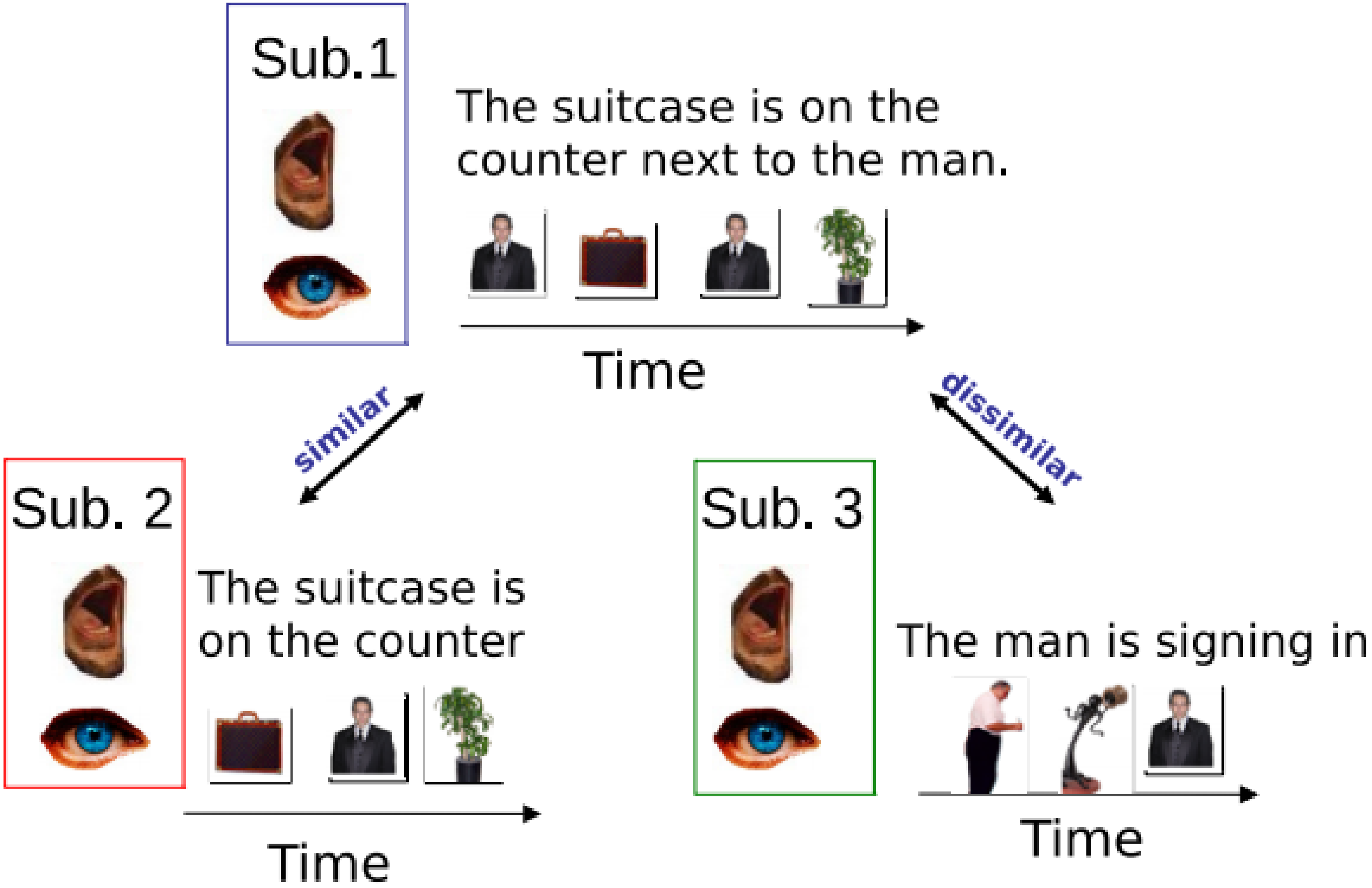
Hypothesis: Speakers use **referents** to coordinate across modalities: for example noun phrases are **grounded** to objects.

Based on this hypothesis, we predict:

- two participants that follow similar scan patterns on an image also produce similar sentences describing it;
- ultimately, we should be able to predict what someone will say if we know their scan pattern.

To test this, we need **similarity measures** for both scan patterns and sentences.

Comparing scan patterns


Sub.1
The suitcase is on the counter next to the man.
Time
similar
dissimilar
Sub. 2
The suitcase is on the counter
Time
Sub. 3
The man is signing in
Time

Testing the grounding hypothesis

Hypothesis: Speakers use **referents** to coordinate across modalities: for example noun phrases are **grounded** to objects.

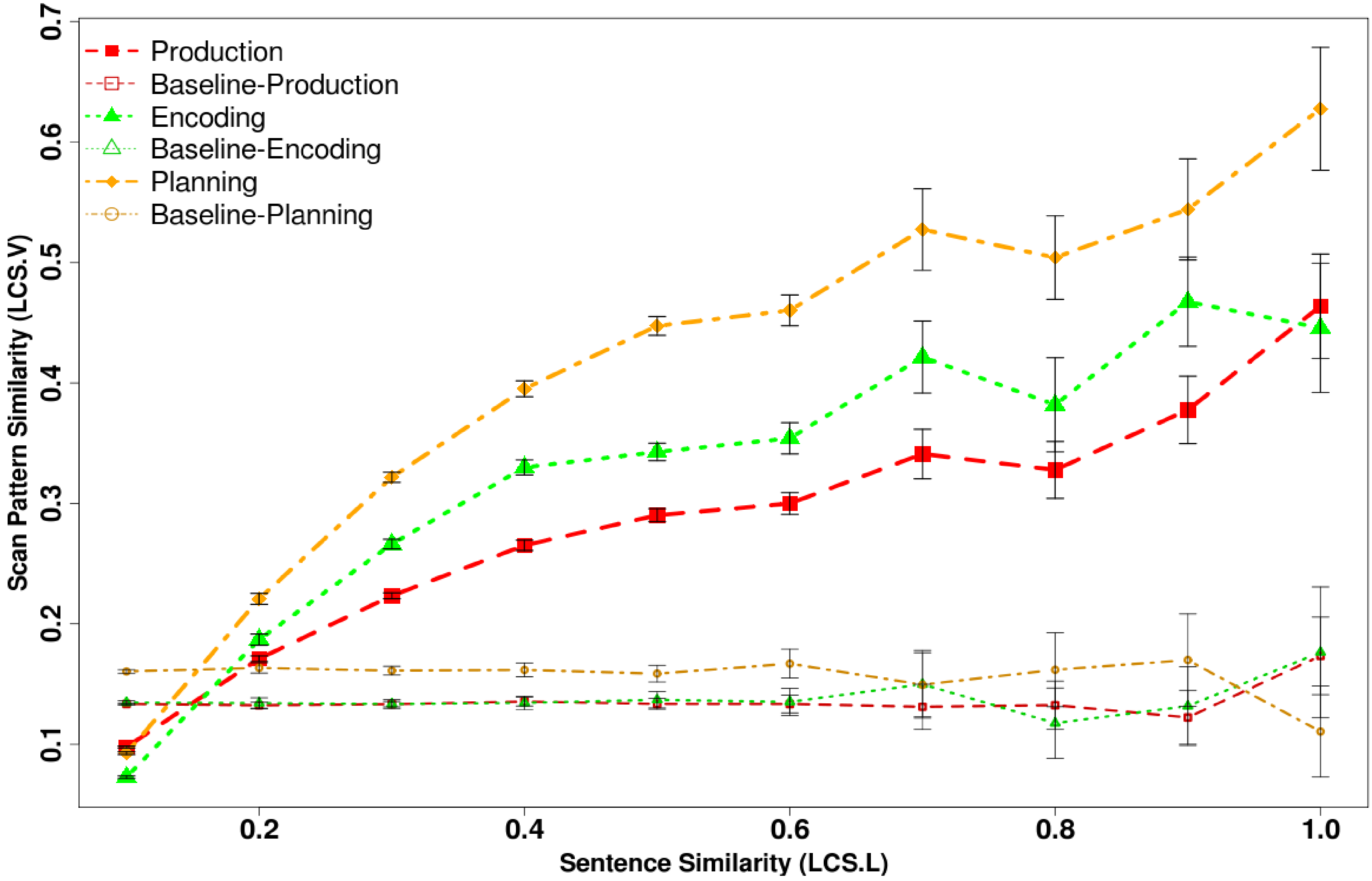
Now let's give participants a scene description task and measure:

- the similarity of the scan patterns they produce;
- the similarity of the sentences they generate.

If the grounding hypothesis correct, then the two similarity measures should correlate.

(We will provide details about how to measure similarity later.)

Testing the grounding hypothesis (Coco & Keller, 2012)



Similar sentences are associated with similar scan patterns

But what about other languages?

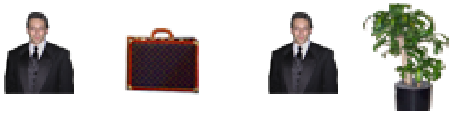
- Human **cognition** is a highly integrated system: **processes and representations** are multimodal.
- The words we utter to describe a visual scene are grounded in the objects we attend.
- **Visual attention and language production** correlate across modalities (Coco & Keller, 2012).

Research Questions
Does this effect generalize across different languages?
Is the correlation driven by semantics or syntax?
What does this tell us about visual grounding?

Grounding across scenes and languages

Let’s test the generality of the grounding hypothesis!

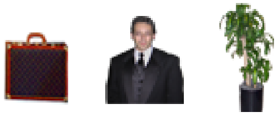
The suitcase is on the counter next to the man.



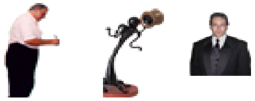
Similarity

Dissimilarity

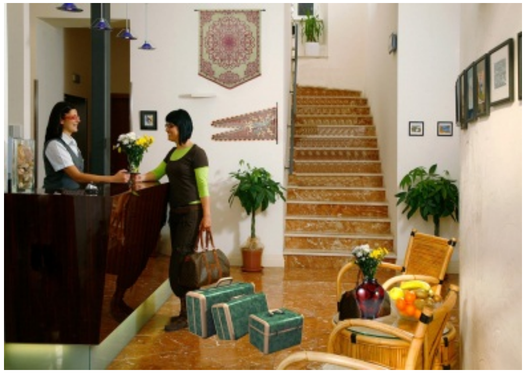
The suitcase is on the counter



The man is signing in



within the same scene



Does this hold between different scenes?

A mulher corre na estrada
(Portuguese, **SVO**)

女性は道路を走っている

woman on road running is
(Japanese, **SOV**)

Does this hold across different languages?

This study: data collection

74 participants (24 British English, 28 European Portuguese and 20 Japanese) were asked to **describe** scenes (24) after being **prompted** with an **ambiguous cue** word while being eye-tracked.

N = 1,776 sentences paired with scan patterns generated during sentence production.

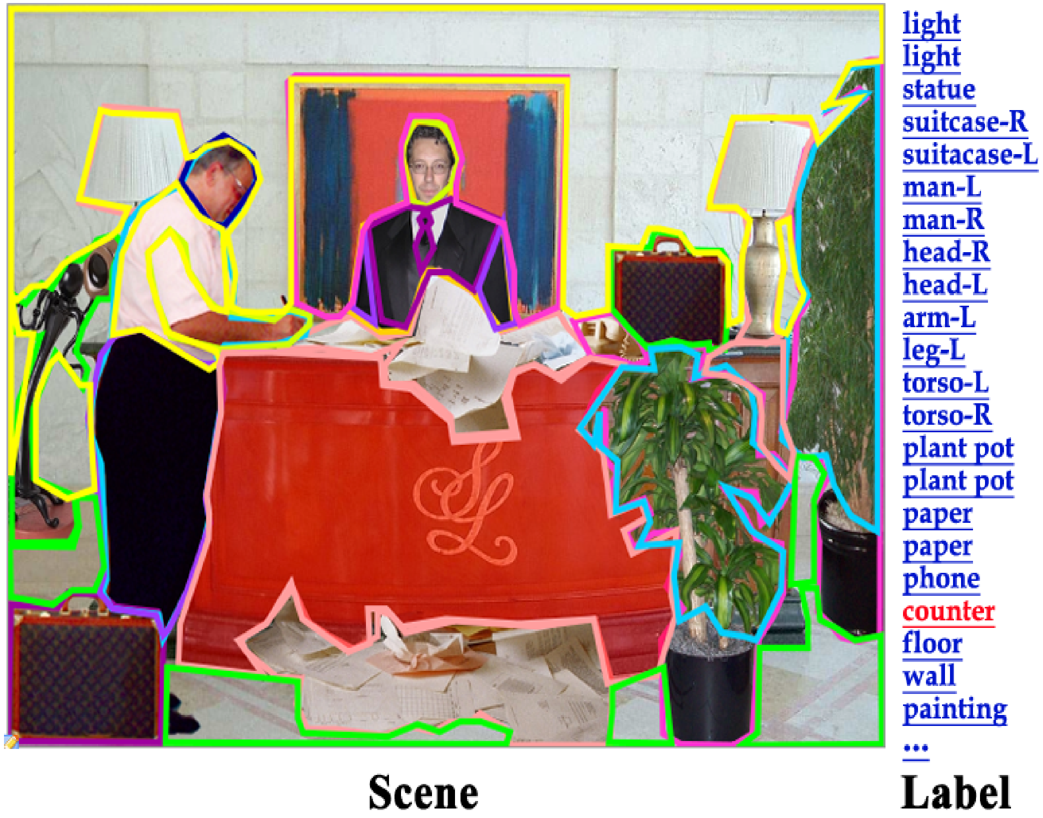
(No. of objects: μ = **28.65**; σ = **11.30**)

Sentence complexity:

- *One man waits for another man to fill out the registration form for a hotel*
- *o homem está a fazer check-in no hotel*
- ホテルのロビーでは男性が2人話をしている。

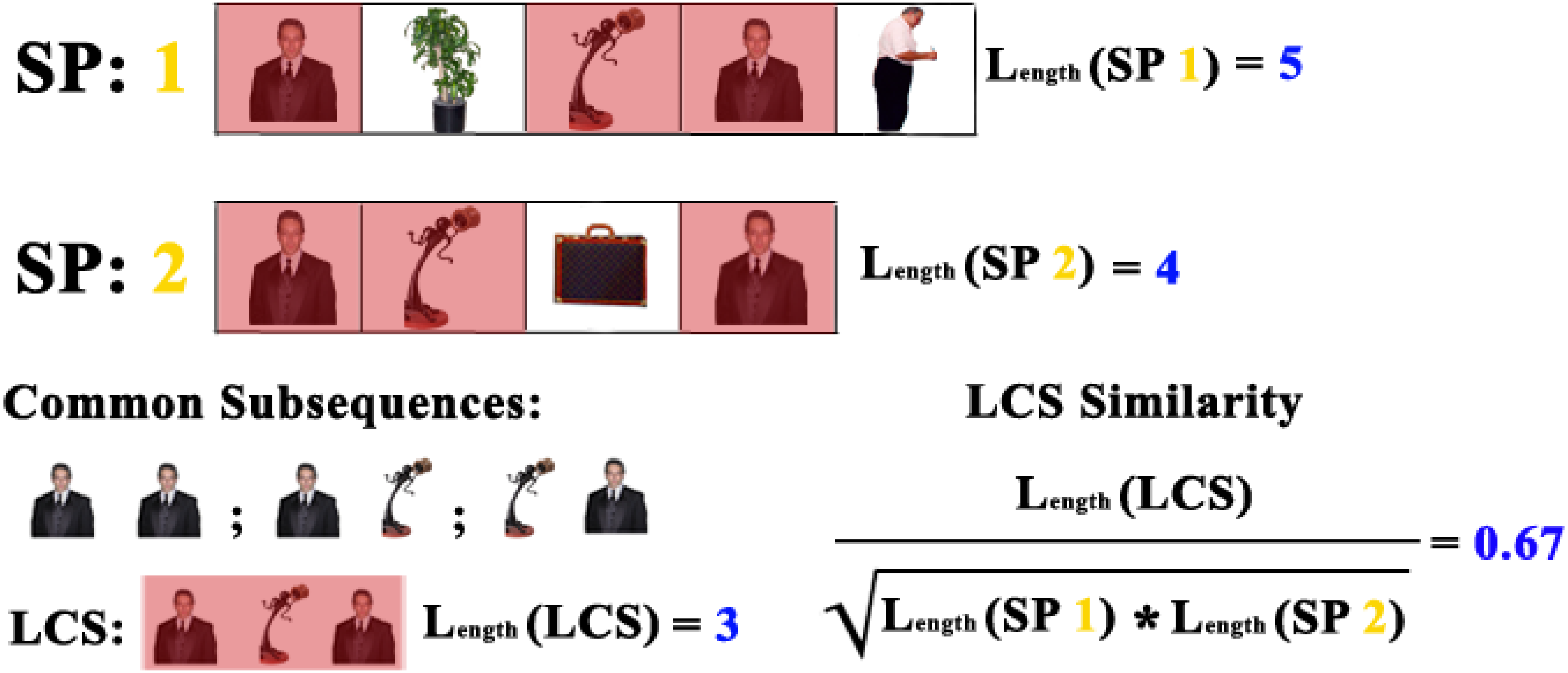
Scan pattern length:

- Preparation for production:
- **min = 800 ms**; **max = 10205 ms**
- During production:
- **min = 2052 ms**; **max = 18361 ms**



Scan pattern similarity

Longest **C**ommon **S**ubsequence (Gusfield, 1997)



SP: 1 $\quad \text{Length}(\text{SP } 1) = 5$

SP: 2 $\quad \text{Length}(\text{SP } 2) = 4$

Common Subsequences:

LCS: $\quad \text{Length}(\text{LCS}) = 3$

LCS Similarity

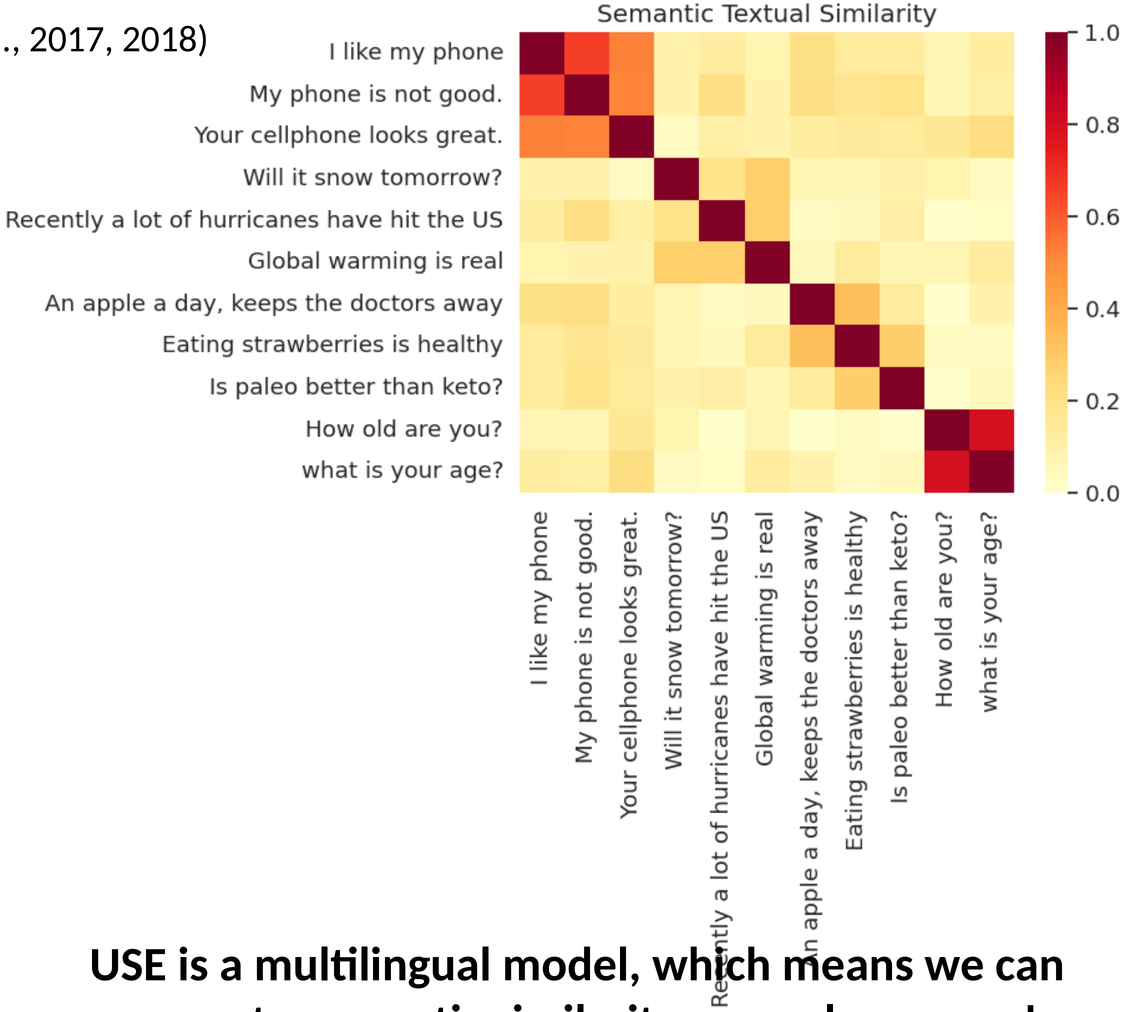
$$\frac{\text{Length}(\text{LCS})}{\sqrt{\text{Length}(\text{SP } 1) * \text{Length}(\text{SP } 2)}} = 0.67$$

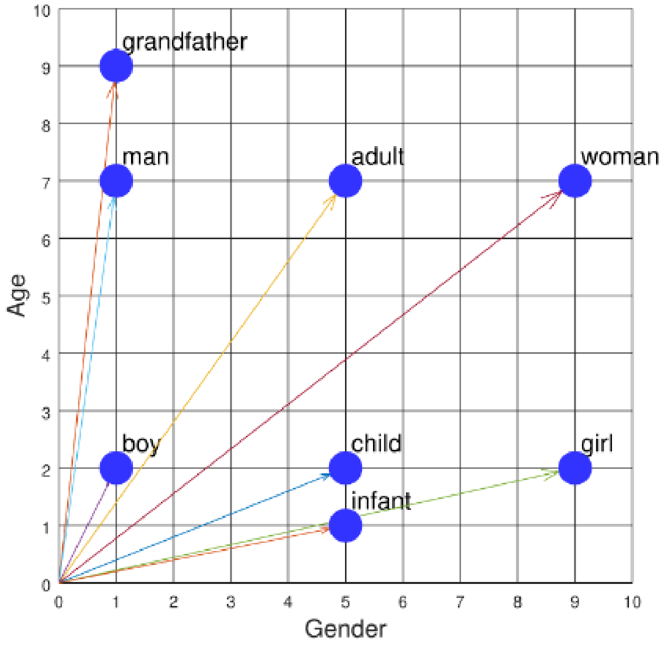
Longest subsequence common to two sequences among all possible sub-sequences.

Semantic similarity

Universal sentence encoder (Cer, et al., 2017, 2018)

A deep neural network trained to extract **sentence embeddings** which are vectors encoding meaning: sentences closer in vector space are semantically more similar.





Word embeddings

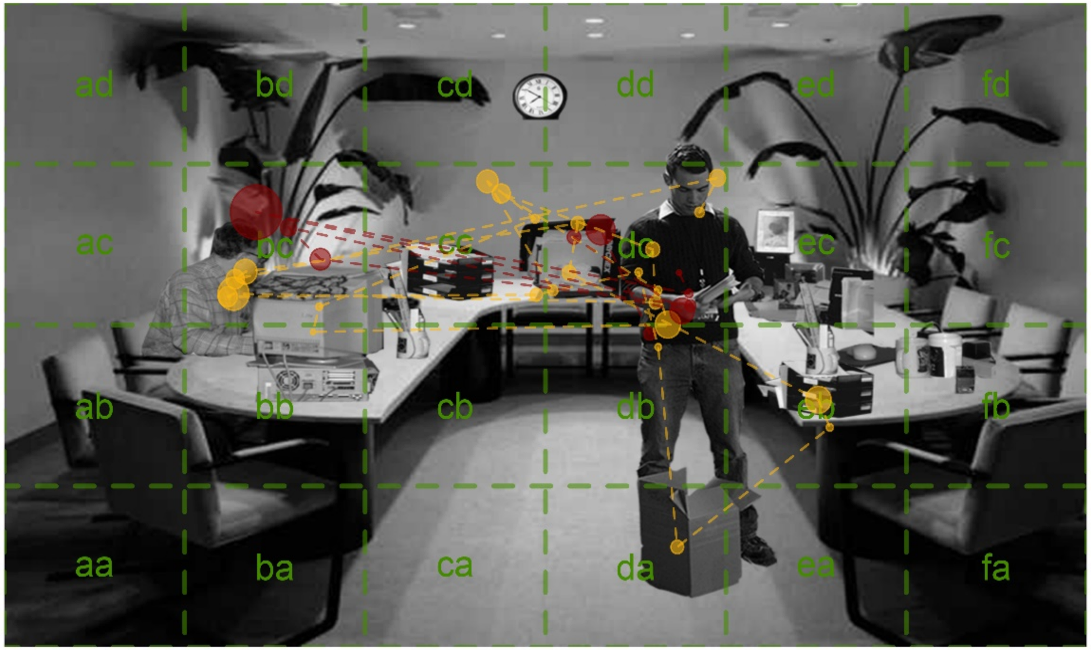
(taken from https://www.cs.cmu.edu/)

USE is a multilingual model, which means we can compute semantic similarity across languages!

We use the dot product to measure vector distance.

Syntactic similarity

Scan patterns of two Japanese speakers (yellow and red):



Cue word: Man (男性)

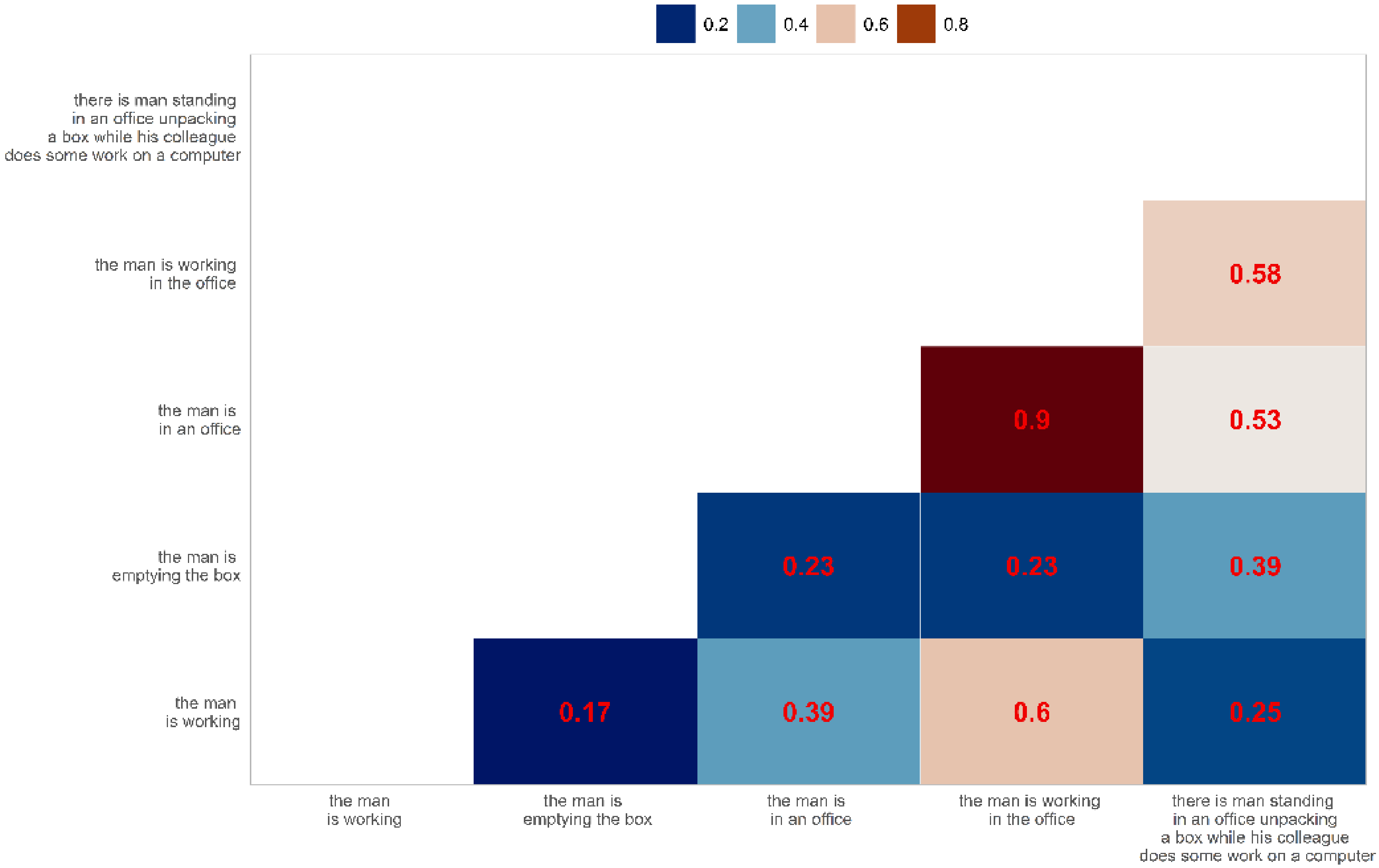
Participant 1:
事務所に2人の男性が働いている
NOUN NOUN ADP NOUN ADP NOUN ADP VERB SCONJ VERB
(two men working in the office)

Participant 2:
男性が2人オフィスにいます
NOUN ADP NOUN NOUN ADP VERB AUX
(two men are in the office)

We part of speech tag descriptions using SpaCy, employing a set of language-independent PoS labels.

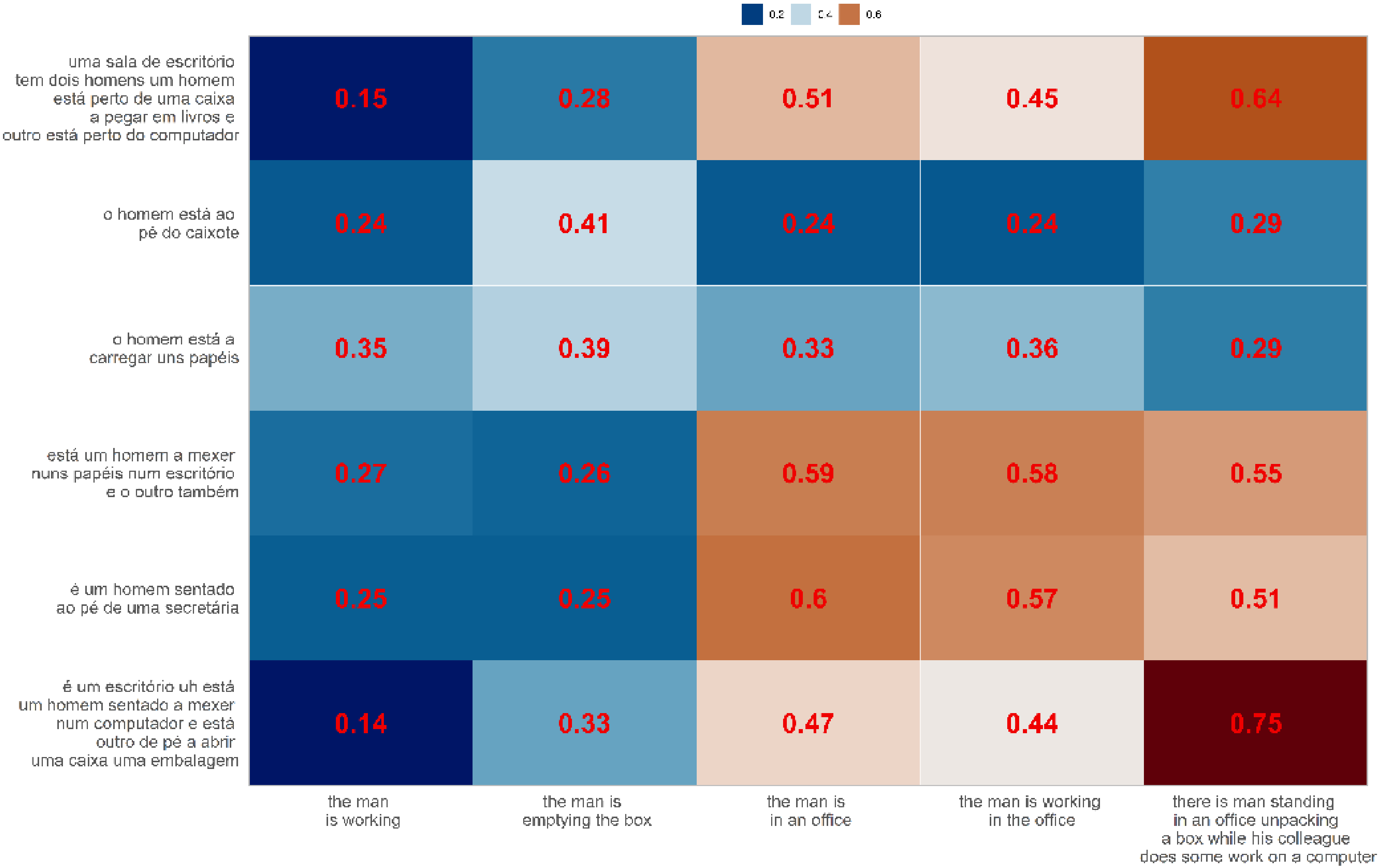
Pairwise similarity metrics for this trial:
LCS (Scan patterns): 0.57
Semantic similarity (dot product of USE vectors): 0.89
LCS (Parts of Speech): 0.71

Putting it all together



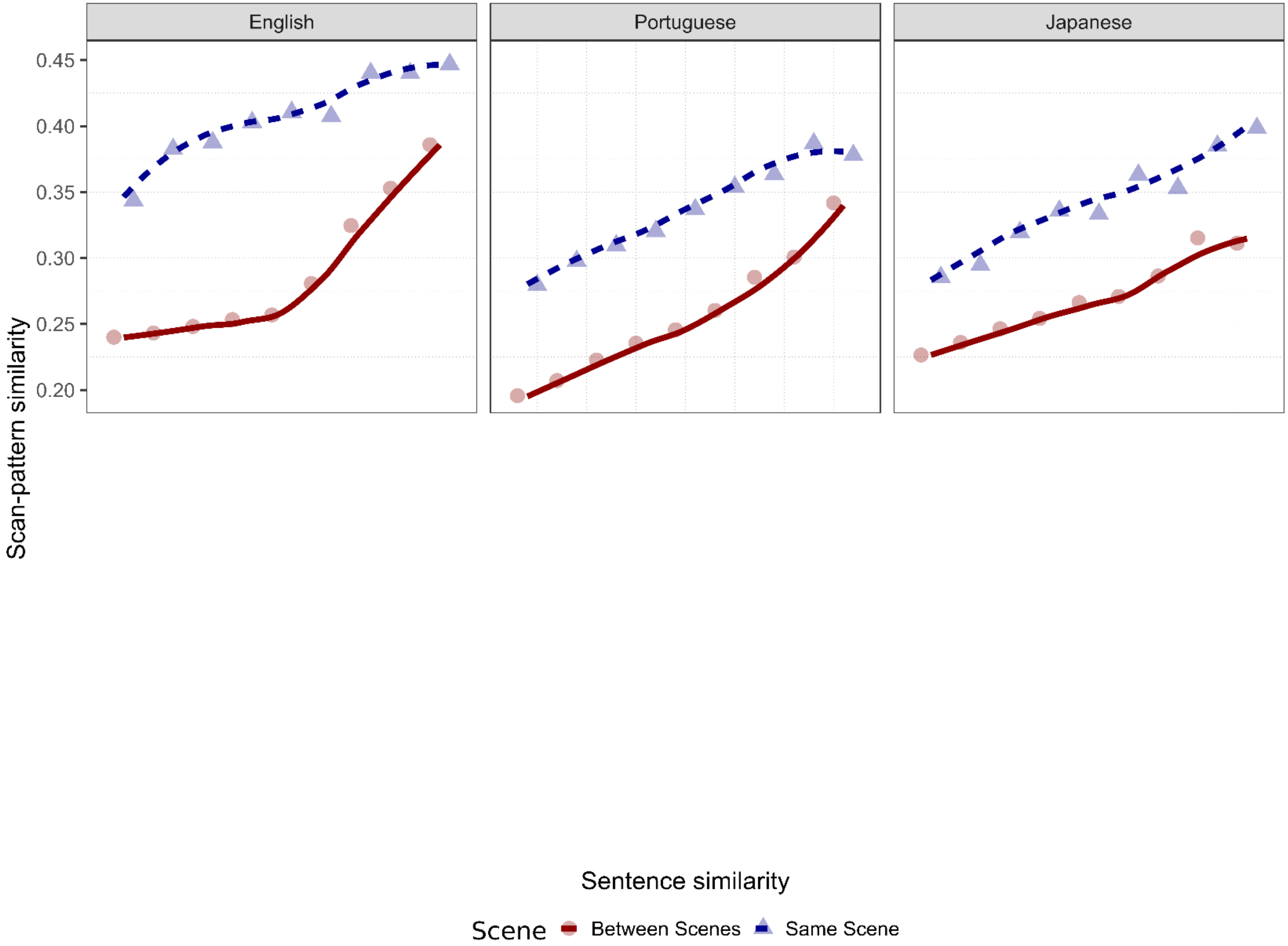
	the man is working	the man is emptying the box	the man is in an office	the man is working in the office	there is man standing in an office unpacking a box while his colleague does some work on a computer
there is man standing in an office unpacking a box while his colleague does some work on a computer					
the man is working in the office					0.58
the man is in an office				0.9	0.53
the man is emptying the box			0.23	0.23	0.39
the man is working		0.17	0.39	0.6	0.25

Putting it all together



	the man is working	the man is emptying the box	the man is in an office	the man is working in the office	there is man standing in an office unpacking a box while his colleague does some work on a computer
uma sala de escritório tem dois homens um homem está perto de uma caixa a pegar em livros e outro está perto do computador	0.15	0.28	0.51	0.45	0.64
o homem está ao pé do caixote	0.24	0.41	0.24	0.24	0.29
o homem está a carregar uns papéis	0.35	0.39	0.33	0.36	0.29
está um homem a mexer nuns papéis num escritório e o outro também	0.27	0.26	0.59	0.58	0.55
é um homem sentado ao pé de uma secretária	0.25	0.25	0.6	0.57	0.51
é um escritório uh está um homem sentado a mexer num computador e está outro de pé a abrir uma caixa uma embalagem	0.14	0.33	0.47	0.44	0.75

Results: Semantics within/between languages

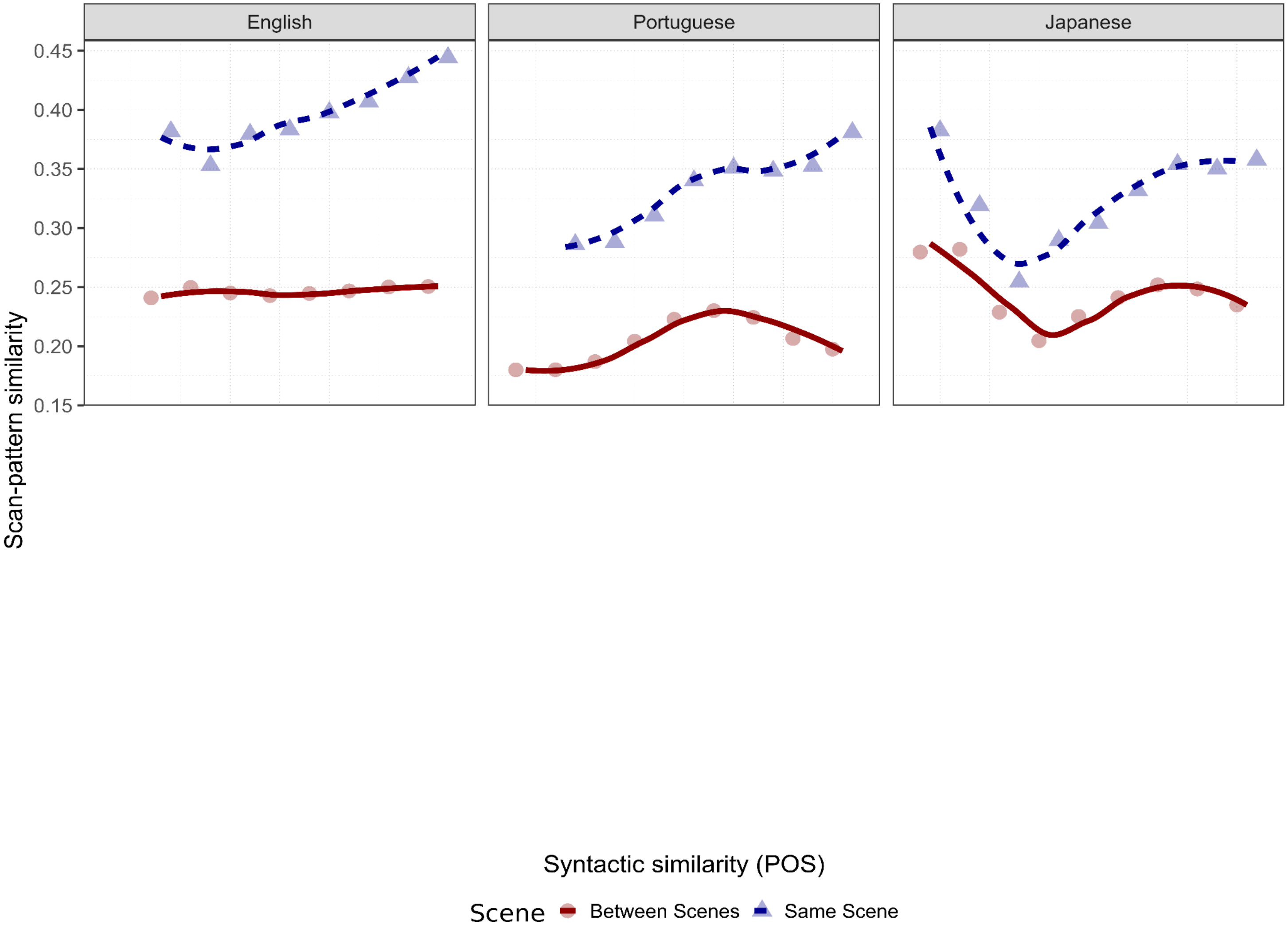


Within language

Between languages

Semantically similar sentences are associated with similar scan patterns – **within and between languages.**

Results: Syntax within/between languages

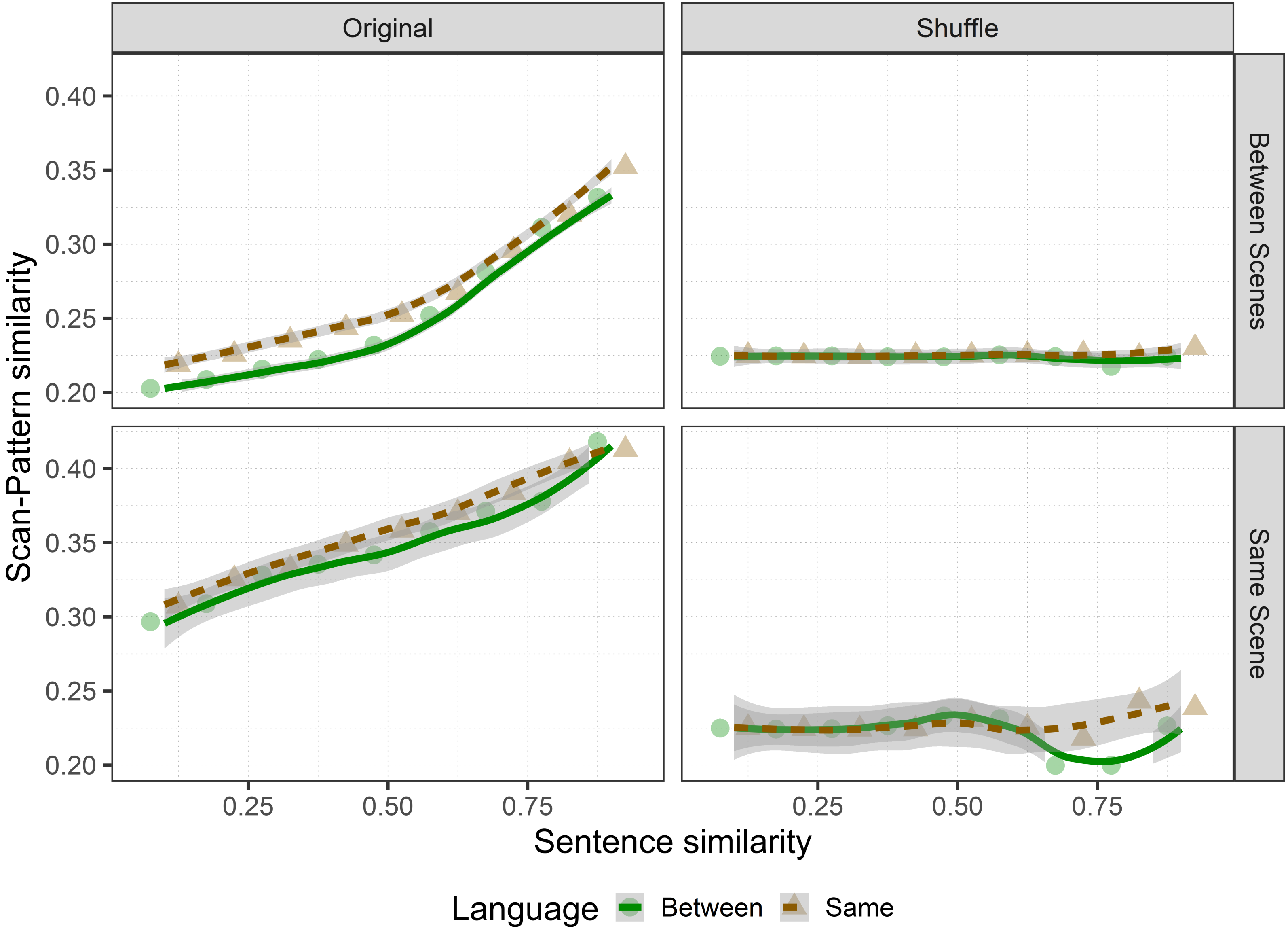


Within language

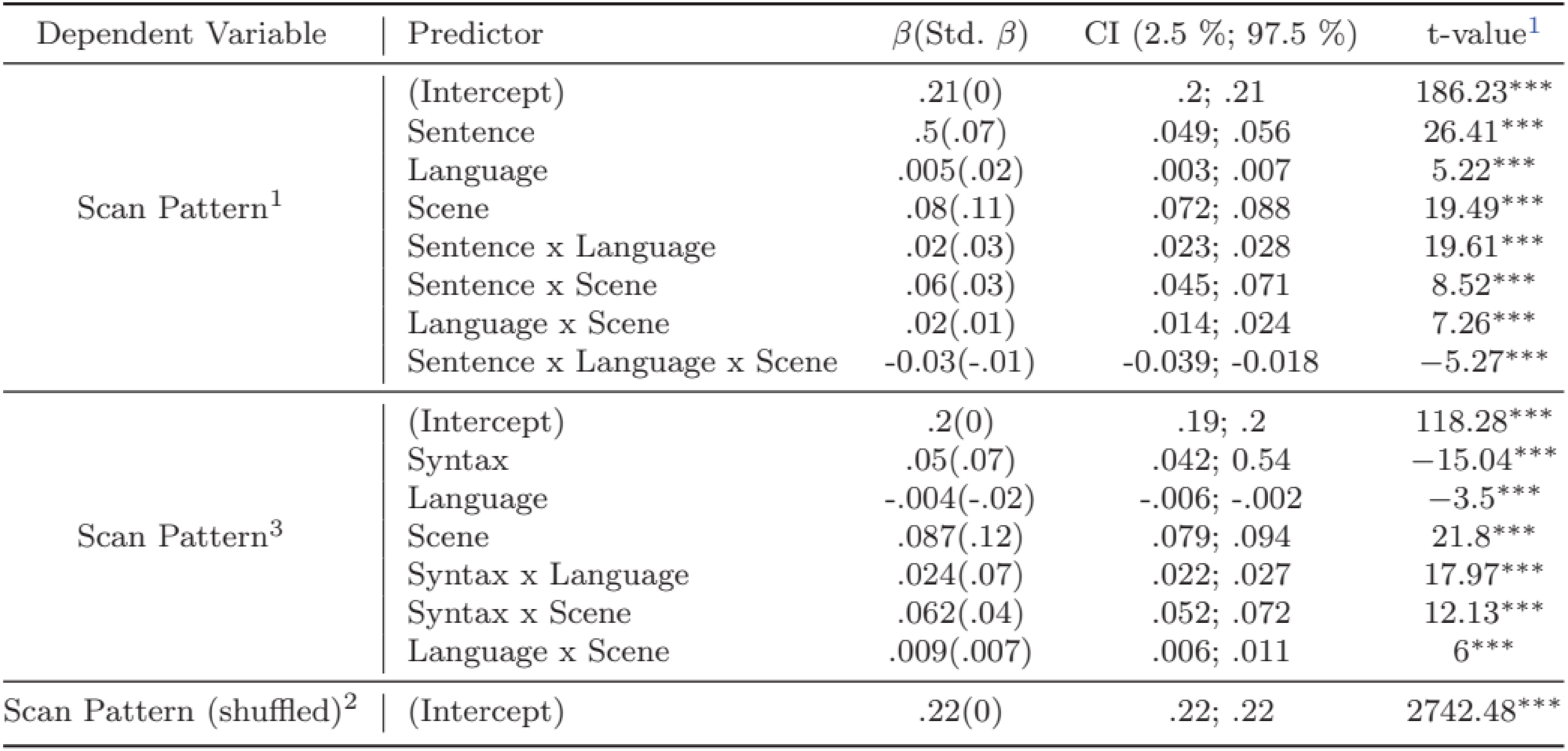
Between languages

Syntactically similar sentences are **not** associated to similar scan patterns – **within or between languages.**

Sanity check: Shuffling sentences and scan patterns



Results: Mixed effects models



Dependent Variable	Predictor	β(Std. β)	CI (2.5 %; 97.5 %)	t-value[1]
Scan Pattern[1]	(Intercept)	.21(0)	.2; .21	186.23***
	Sentence	.5(.07)	.049; .056	26.41***
	Language	.005(.02)	.003; .007	5.22***
	Scene	.08(.11)	.072; .088	19.49***
	Sentence x Language	.02(.03)	.023; .028	19.61***
	Sentence x Scene	.06(.03)	.045; .071	8.52***
	Language x Scene	.02(.01)	.014; .024	7.26***
	Sentence x Language x Scene	-0.03(-.01)	-0.039; -0.018	−5.27***
Scan Pattern[3]	(Intercept)	.2(0)	.19; .2	118.28***
	Syntax	.05(.07)	.042; 0.54	−15.04***
	Language	-.004(-.02)	-.006; -.002	−3.5***
	Scene	.087(.12)	.079; .094	21.8***
	Syntax x Language	.024(.07)	.022; .027	17.97***
	Syntax x Scene	.062(.04)	.052; .072	12.13***
	Language x Scene	.009(.007)	.006; .011	6***
Scan Pattern (shuffled)[2]	(Intercept)	.22(0)	.22; .22	2742.48***

Conclusions

- Semantically similar sentences are associated with similar scan patterns.
- This relationship holds **across scenes and across languages,** and even for languages with different surface realizations (e.g., English and Japanese).
- In contrast, syntactic sentence similarity is predictive of scan-pattern similarity **only within the same language and scene.**
- Suggests that **visual grounding** is mostly driven by semantics, unaffected by the syntax of individual languages.
- What about task generality? Our results might only hold for scene description.