Corpus Informed Enrichment of Lexical Resources (with Fixed Similes) via Facebook-enabled Crowdsourcing

Stella Markantonatou[1], Jelena Mitrovic[2]
[1] Institute for Language and Speach Processing/
Athena RC,Greece
[2] Universität Passau, Germany

The rhetorical figure *Simile*

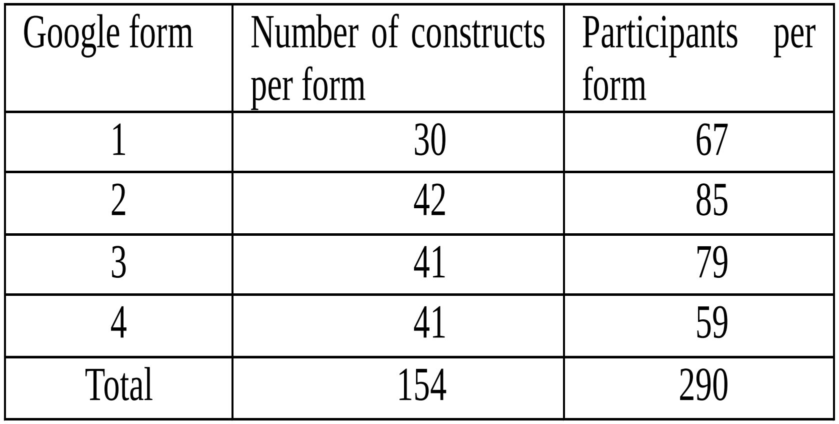
- Appears in a number of syntactic patterns
- Basic pattern: Adjective (‘property’) + comparator + (Det) + Noun (‘vehicle’): *white as snow*
- Similes categorise the predicated entity (‘tenor’) by reference to some, normally highly conventionalised, property of the vehicle (Hanks 2005)
- We studied similes that denote the tenor property explicitly: ‘closed’ similes because the implication made by the ‘vehicle’ (Noun) is clearly stated with the ‘property’ (Adjective) and is not ‘open’ to conjecture (Qadir et al. 2016).

The crowdsourcing experiment: the data

- ~2000 phrases with potential Greek similes from the Hellenic National Corpus (HNC) and from a corpus (~110M words) obtained with web crawling (Mastropavlos and Papavasiliou, 2011).
- corpora searched with the pattern “adjective σα(ν) (Det) noun” (*σα(ν)* = ‘as’, ‘like’).
- Low frequency:
  - (i) ~500 phrases in the second corpus, only 250 featured a fixed simile
  - (ii) most similes occurred only once in this corpus.
- In all, 154 candidate similes were used (as in the parallel Serbian experiment **Mladenović et al. (2016)**).

Crowdsourcing

- Google Forms were circulated via Facebook
- Native speakers of Greek were presented with a form that allowed them to click on a construct
- **“tick what you use in your everyday talk**” and NOT “tick the constructs that you know and/or use”



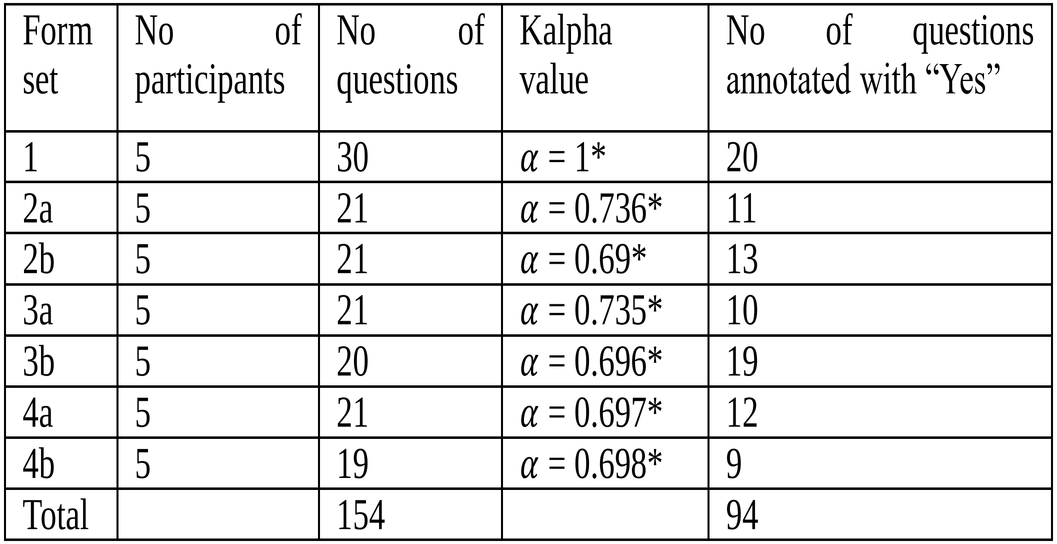
Google form	Number of constructs per form	Participants per form
1	30	67
2	42	85
3	41	79
4	41	59
Total	154	290

Crowdsourcing

•

Crowdsourcing followed Krippendorff’s recommendations for obtaining usable data (Artstein & Poesio, 2008: 21):

- data for each construct were received from more than three people (in the reported experiment, five speakers at least)
- the criterion presented to the speakers was only one and it was clear
- the speakers worked independently



Form set	No of participants	No of questions	Kalpha value	No of questions annotated with “Yes”
1	5	30	$\alpha = 1$*	20
2a	5	21	$\alpha = 0.736$*	11
2b	5	21	$\alpha = 0.69$*	13
3a	5	21	$\alpha = 0.735$*	10
3b	5	20	$\alpha = 0.696$*	19
4a	5	21	$\alpha = 0.697$*	12
4b	5	19	$\alpha = 0.698$*	9
Total		154		94

Observed WEB populations

- A thorough check of simile properties, for instance what type of ‘tenors’ they select and whether the similes are flexible or not, has to draw on Google search results.
- We searched the Web with “construct” and checked all the retrieved examples. We used the full morphological paradigms of the similes in normal and in inverted word order. Only non-identical examples were stored.
- “Yes”-constructs return results ranging from 5 to hundreds of single occurrences.
- Certain “No” constructs could be considered “Yes”-constructs if their frequency of use on the Web was used as a criterion of “simile-hood”.

	YES		G	NO		G
1	Αδύνατος σαν οδοντογλυφίδα	Skinny as toothpick	26	Βαρύς σαν πέτρα	Heavy as stone	112
2	Αδύνατος σαν σκελετός	Skinny as skeleton	33	Γαλανός σαν (τη) θάλασσα	Blue as the sea	4
3	Αδύνατος σαν στέκα	Skinny as cue	27	Γερός σαν ταύρος	Healthy as bull	111
4	Αθώος σαν άγγελος	Innocent as angel	22	Γυμνός σαν άγαλμα	Naked as statue	5
5	Αθώος σαν παιδί	Innocent as child	54	Καθιστός σαν το Βούδα	Seated as the Buddha	0
5	Ακλόνητος σαν βράχος	Immobile asrock	96	Κίτρινος σαν φλουρί	Yellow as the lire	150
7	Αλαφρύς σαν πούπουλο	Light as dawn	200	Κολλημένος σαν πεταλίδα	Stuck as limpet	0
8	Ανάλαφρος σαν αεράκι	Light as breeze	30	Σιωπηλός σαν σφίγγα	Silent as sphinx	8
9	Απαλός σα μετάξι	Soft as silk	364	Στρογγυλός σαν τόπι	Round as ball	55
10	Απαλός σαν χάδι	Soft as stroke	282	Φουσκωμένος σαν παγόνι	Bloated as peakock	25
11	Αργός σαν χελώνα	Slow as turtle	21	Φωτεινός σαν φεγγάρι	Bright as moon	10
12	Άσπρος σαν το γάλα	White as the milk	249			
13	Άσπρος σαν το πανί	White as the cloth	505			
14	Άσπρος σαν το χιόνι	White as the snow	231			
15	Αστραφτερός σαν το διαμάντι	Shiny as the diamont	19			
16	Βαρύς σαν μολύβι	Heavy as lead	143			
17	Βρεγμένος σαν πάπια/το παπί	Wet as duck/duckling	16			
18	Βρώμικος σαν γουρούνι	Dirty as pig	5			
19	Γλυκός σαν μέλι	Sweet as honey	307			
20	Γρήγορος σαν αστραπή	Fast as lighting	234			
21	Γρήγορος σαν λαγός	Fast as hare	19			
22	Δειλός σαν κότα	Coward as hen	7			
23	Δυνατός σαν ταύρος	Strong as bull	120			
24	Στολισμένος σαν φρεγάτα	Adorned as frigate	15			
25	Φωτεινός σαν ήλιος	Bright as sun	75			

We asked ‘tick what you use’ and..

- The “No”-construct *κίτρινος σαν το φλουρί* (yellow as lire) seems to be old-fashioned, therefore the speakers correctly did not give it a YES vote
- The “No”-construct *βαρύς σαν πέτρα* (heavy as stone) often occurs in fixed phrases and in genres like poetry, artistic prose and songs. However, free, everyday usages do exist
- *γερός σαν ταύρος* (healthy as bull) is a problem because many of the examples retrieved with Google search seem to originate in social media texts.
- Several “Yes”-constructs return few Google search results.
  - *βρώμικος σαν γουρούνι* (dirty as pig) competes with the morphologically and semantically related fixed verb simile *βρωμάω σα γουρούνι* (I stink like pig)
  - Others are highly marked: *στολισμένος σαν φρεγάτα* (adorned like a frigate) is a bit old-fashioned and normally applies only to women.

Bottom line

- Only “Yes” cases with significant populations (say > 100) and “No” cases with zero populations could be encoded reliably as fixed similes of the “live” Greek language– or be rejected respectively.

- It is on the linguist/lexicographer to decide which of the remaining structures she will document as fixed similes --also what type of features such as ‘colloquial’, ‘old-fashioned’ etc she will employ in order to better depict the combination of Web data with native speaker intuitions (as they were encoded with the crowdsourcing experiment).

THANK YOU!

ΕΥΧΑΡΙΣΤΟΥΜΕ