Fixed Similes: Measuring Aspects of the Relation between MWE Idiomatic Semantics and Syntactic Flexibility

Stella Markantonatou
Institute for Language and Speech Processing, Athena RIC, Greece
stiliani.markantonatou@gmail.com

Panagiotis Kouris
School of Electrical and Computer Engineering, National Technical University of Athens, Greece
pkouris@islab.ntua.gr

Yanis Maistros
School of Electrical and Computer Engineering, National Technical University of Athens, Greece
maistros@cs.ntua.gr

Abstract

We shed light on aspects of the relation between the semantics and the syntactic flexibility of multiword expressions (MWE) by investigating fixed adjective similes (FS), a predicative. MWE class not studied in this respect before. We work on Modern Greek data[1] and find that only a subset of the observed syntactic structures is related with idiomaticity. We identify and measure two aspects of semantic idiomaticity, one of which seems to allow to predict FS syntactic flexibility. Our research draws on a resource developed with the semantic and detailed syntactic annotation of web-retrieved material, indicating frequency of use of the individual similes.

1 Introduction

The relation between the idiomatic semantics of multiword expressions (MWE) and their syntactic flexibility, namely the ability of a MWE to occur in different syntactic configurations without loss of the idiomatic meaning, has offered a fertile field of research because it challenges the notion of compositionality, namely the derivation of the meaning of an utterance from the meaning of its components and its syntactic structure. The (predicative) free subject verb MWEs of the type verb+noun (V+N) have been one of the privileged fields of these studies. We add to the understanding of this relation by investigating a class of predicative (non-verb) MWEs that, to the best of our knowledge, has not been studied from this point of view so far, namely the class of fixed similes (FS) of the type adjective+connector+(article)+noun such as *sweet like honey*. Modern Greek, a lesser studied language in this respect, is our object language. We ask whether the occurrence of the syntactic structures considered as manifestations of syntactic flexibility is related with (idiomatic) semantics and we find that this holds only for some structures –a subset of which could be characterised as syntactic alternatives. Our results corroborate the idea that syntactic structures demonstrate varying sensitivity to the semantics of their components and that the presence of syntactic variants may not be dependent only on idiomatic semantics. Next, we investigate the notion of idiomaticity. We understand idiomaticity as the degree of similarity between the FS semantics and the semantics of their free property (i.e., the free adjective, see Table 1) and, we identify and measure two types of similarity, one of which allows us to make predictions about FS syntactic flexibility. Our results were drawn from a resource developed as part of the research presented here by extensively annotating a large amount of web-retrieved unique FS usage examples. The special feature of this resource is that it offers an as good as possible approximation of the actual FS usage in texts, both in terms of frequency and in terms of structure selection.

Table 1 shows the terminology for the parts of an FS adopted from Hanks (2005):

[1]We acknowledge support of this work by the project "Computational Sciences and Technologies for Data, Content and Interaction" (MIS 5002437) which is implemented under the Action "Reinforcement of the Research and Innovation Infrastructure", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), pages 51–61
Santa Fe, New Mexico, USA, August 25-26, 2018.

entity modified by the FS	adjective	connector	(article) noun
tenor	property		vehicle

Table 1: Terminology for the parts of an FS.

In §2 we contextualise our work with respect to state-of-the-art research. In §3 we talk about data collection and annotation and the development of the resource. In §4 we describe the results we receive by quantitatively studying the resource. We present our conclusions and our plans for the future in §5.

2 Related work

Simile is a figure of speech, which, unlike metaphor, draws attention to the likeness between the tenor and the vehicle that are implied to share certain properties (Veale and Hao, 2007). Similes draw on conventional beliefs about likeness and effectively convey the speaker's superlative evaluation of the tenor (Israel et al., 2004). They follow the same structure as comparative, e.g. *Knowledge is **sweeter than honey***, equative, e.g. *Her argument was **as clear as glass*** and similative statements, e.g. *My hands were **cold like ice*** (Chila-Markopoulou, 1986; Israel et al., 2004; Niculae and Danescu-Niculescu-Mizil, 2014; Mpouli and Ganascia, 2015). Many conventional similes tend to be fixed and have idiom status (Hanks, 2005). For instance, FS6 κόκκινος σαν αστακός (kokinos san astakos) 'red as lobster' (see Table 2) applies mainly to persons or body (parts) (96,4% in our data) to denote blushing or sunburned people; this meaning can not be derived from the parts of the FS. Some English FS tend to select tenors from a small range of semantic fields (Mpouli and Ganascia, 2015); for Greek see Figure 1. FS have been found to share a good part of simile usage in English (Niculae and Danescu-Niculescu-Mizil, 2014).

Modern Greek FS assume the syntactic functions of the adjectives (Markantonatou et al., 2016). They appear as complements of the copula (1), as phrasal adjuncts controlled by the subject of the main verb (3), in verbless sentences (4), in the typical adjective position between the determiner and the noun (5) and, rarely, as subjects and objects of verbs or prepositions. In all these structures, FS can be replaced with the superlative of the adjective (exemplified only for the complement of copula (2)). They may occur with the adverb πολύ (poli) 'much' or, rarely, with some other intensifying adverb. In this case, the simile is not pronounced as a unit but as if a punctuation mark existed between the property and the σαν+vehicle part. We annotated these structures as *empp* (phrasal emphasis (Table 3)). We will treat FS as MWE adjectives.

(1) Ήταν ο ύπνος γλυκός σαν το μέλι της κερήθρας.
Itan o ipnos glikos san to meli tis kierithras.
was the sleep.nom sweetnom like the honey the honeycomb.gen
'Sleep was as sweet as the honey of the honeycomb.'

(2) Ήταν ο ύπνος πάρα πολύ γλυκός.
Itan o ipnos para poli glikos.
was the sleep.nom much very sweet.nom
'Sleep was extremely sweet.'

(3) Ελαφρύς σαν πούπουλο, πήδηξε στο μαξιλάρι.
Elafris san pupulo, pidikse sto maksilari.
light.nom like feather, jumped on.the pillow
'As light as the feather, he jumped on the pillow.'

(4) Ύπνος γλυκός σαν μέλι.
Ipnos glikos san meli.
sleep.nom sweet.nom like honey
'A sleep as sweet as honey.'

(5) Ένας γλυκός σαν μέλι ύπνος τον τύλιξε.
Enas glikos san meli ipnos ton tilikse.
a sweet like honey sleep him wrapped
'He had a very sweet sleep.'

Computational methods distinguishing between similes and other comparisons take advantage of the attested semantic distance between the similes' tenor and vehicle and of definiteness (Niculae and Danescu-Niculescu-Mizil, 2014). However, we show that in Modern Greek, determiner distribution is independent of FS idiomaticity (§4.1).

The study of large amounts of data of V+N MWEs has shown that MWE syntactic flexibility is related to MWE idiomaticity modeled as the semantic distance between the environments where the MWE and its parts occur (Fazly et al., 2009). Our research on FS sheds more light on this attested relation.

A much debated proposal claims that MWE syntactic flexibility is related with MWE semantic analysability, namely the degree to which the idiomatic semantics of the MWE can be derived synthetically from the (idiomatic) meanings of its parts (indicatively, Nunberg et al. (1994), Kay and Sag (2012) argue in favor of the syntactic analysability approach and Laporte (2018) argues against it). Interestingly, (Kay and Sag, 2012, 4) maintain that while "meaningful idiom words can be modified and can appear in syntactic contexts that meaningless ones cannot", only those language structures that "do not entail interpretive consequences (as does, for example, English topicalization) are precisely the syntactic contexts that permit meaningless idiom words to be "displaced."" We demonstrate that the syntactic environments in which FS occur can be split into FS semantics sensitive and insensitive ones.

3 The resource

N	Greek FS	English translation	Instances	Percentage
1	άσπρος (λευκός) σαν το πανί	as white as a sheet	462	9.5%
2	στολισμένος σαν φρεγάτα	Lit. adorned like a frigate	16	0.3%
3	απαλός σαν πούπουλο	Lit. soft like feather	36	0.7%
4	απαλός σαν χάδι	Lit. soft like cuddle	174	3.6%
5	ελαφρύς (αλαφρύς) σαν πούπουλο	as light as a feather	201	4.1%
6	κόκκινος σαν αστακός	Lit.red like lobster	104	2.1%
7	οπλισμένος (αρματωμένος) σαν αστακός	Lit. armed like a lobster	388	8.0%
8	μαλακός σαν βούτυρο	as soft as butter	162	3.3%
9	γερός σαν ταύρος	as strong as a bull	102	2.1%
10	πιστός σαν σκύλος (σαν σκυλί)	as faithful as a dog	69	1.4%
11	κόκκινος σαν παπαρούνα	as red as a poppy	173	3.6%
12	ντυμένος σαν αστακός	Lit. dressed like lobster	24	0.5%
13	κόκκινος σαν παντζάρι	as red as a beetroot	122	2.5%
14	γλυκός σαν μέλι	as sweet as honey	501	10.3%
15	άσπρος (λευκός) σαν το γάλα	as white as milk	517	10.6%
16	κρύος σαν τον πάγο	as cold as ice	272	5.6%
17	γρήγορος (γοργός) σαν την αστραπή	as quick as a flash	275	5.7%
18	μαύρος σαν το σκοτάδι	as black as pitch	81	1.7%
19	μπερδεμένος σαν κουβάρι	Lit. knotted like ball	84	1.7%
20	άσπρος (λευκός) σαν το χιόνι	as white as snow	1099	22.6%

Table 2: FS distribution in the resource.

3.1 Data collection

With a tailor-made Facebook application (Mitrović and Markantonatou, 2018) we asked 260 native speakers of Modern Greek to specify which of 152 similes they would use in their everyday exc hange.[2] Krippendorff's alpha coefficient (Artstein and Poesio, 2008) was used to evaluate inter-speaker agreement. 85 similes were found to be used by a critical number of speakers. According to Baldwin and Kim

[2]The similes were collected from the Hellenic National Corpus (HNC), http://hnc.ilsp.gr/, and a corpus of 100 million words that was collected with crawlers. The corpora were searched for the pattern adjective+σαν+noun and only structures that appeared more than once were retained.

(2010), the selected similes can be considered MWEs because (i) they fulfill the statistical idiomaticity criterion; as explained in §3.2, the similes are subject to minimal lexical variation (ii) most of them fulfill the semantic idiomaticity criterion (§2). Of them, we selected 20 FS (Table 2) that represent the classes defined in two different simile classifications by the semantics of the vehicle and of the property, one of Modern Greek FS (Mpolla-Mavridou, 1996) and one of English similes (Hanks, 2005). A corpus consisting of 4900 unique usage examples (instances from now on) was collected from the web with exhaustive searches, which allowed our resource to represent instance frequency.[3] Regular expressions were used to capture the rich morphology of Modern Greek, its flexible word order as well as lexical and syntactic variation. The retrieved material was cleaned from multiple occurrences of the same instance and from machine translation outputs. Table 2 shows the distribution of instances per FS in the corpus.

3.2 Lexical and syntactic variation in the case of Modern Greek FS

Lexical variation is not attested for most FS in our data (Table 2). The adjectives άσπρος (aspros) 'white', οπλισμένος (oplismenos) 'armed' and γρήγορος (grigoros) 'fast' variate between synonyms that differ in terms of style and the variants occur in considerable frequencies. FS1 άσπρος σαν το πανί (aspros san to pani) 'as white as sheet' is exceptional because there is a small number of adjectives with the sense *pale* that variate with άσπρος: κίτρινος (kitrinos) 'yellow', ωχρός (ochros) 'ashy, ghostly', χλωμός (chlomos) 'pale'. FS10 πιστός σαν **σκύλος** (pistos san skilos.masc) 'as faithful as a dog', variates with πιστός σαν **σκυλί** (pistos san skili.neut) because the vehicle occurs in both genders.

Syntactic variation is more frequent in our data. The FS alternative with connector σαν (san) 'like, as' (6) is the most frequently used form, which we will call the normative form. The same property and vehicle may appear in a comparative structure; (7) features a morphological comparative form and (8) a phrasal one (Chila-Markopoulou, 1986). They can also appear in an equative form (9). Similar observations have been made for English (Niculae and Danescu-Niculescu-Mizil (2014); Mpouli and Ganascia (2015)). We will use the term FS alternative to refer to syntactic variations (6)-(10): we have also treated as FS alternatives the structures that contain a punctuation mark (comma, full stop, dots) between the property and the connector+vehicle part of the FS (10) provided that the parts on either side of the punctuation mark are predicated of the same entity.

(6) Σε παίρνει ο γνωστός μεσημεριανός ύπνος, ο **γλυκός σαν μέλι**.
Se perni o gnostos mesimerianos ipnos, o glikos san meli.
you.acc take the familiar noon sleep.nom, the sweet like honey
'The familiar nap, that is sweet like honey, comes all over you.'

(7) Είναι **γλυκύτερος από μέλι** και δυστυχώς δεν δείχνει σκυλί επιβίωσης.
Ine glikiteros apo meli kie distichos den dichni skili epiviosis.
is sweeter than honey and unfortunately not shows dog of.survival
'It is sweeter than honey but, unfortunately, it does not seem to be a survivor dog.'

(8) Έλεγε τραγούδια γλυκά, **πιο γλυκά κι από το μέλι** της μέλισσας.
Elegie tragudia glika, pio glika ki apo to meli tis melisas.
said songs sweet more sweet and from the honey of.the bee
'He sang sweet songs, sweeter than the honey of the bee.'

(9) Αλείφουν τις βρύσες με μέλι, ώστε ο καινούριος χρόνος να είναι **τόσο γλυκός σαν το μέλι**.
Alifun tis vrises me meli, oste o kienurgios chrons na ine toso glikos san to meli.
rub the taps with honey so that the new year to be as sweet like the honey
'They rub the taps with honey to ensure that the new year will be as sweet as honey.'

(10) Και είναι **τόσο απαλό**. **Σαν το στερνό σου χάδι**.
Kie ine toso apalo. San to sterno su chadi.
and is so soft like the last your cuddle
'And it is so soft. Like your last cuddle.'

[3] The available Modern Greek corpora offered less than a hundred of instances.

3.3 Syntactic and semantic annotation

Syntactic annotation modeled FS syntactic flexibility; it captured the normative form and the deviations from it observed in the data, including FS alternatives. As Greek FS function as adjective MWEs (§2), we assume that their semantics is represented by the semantics of the NPs they select as tenors (Table 1). We used WordNet supersenses to annotate the tenors; this detailed semantic annotation allowed us to distinguish between types of idiomaticity. In the remainder, we will use the term FS semantics to denote the set of semantic annotations assigned to the tenors of a particular FS in the resource. High degrees of inter-annotator agreement were obtained as confirmed by Krippendorff's alpha coefficient.

3.3.1 Syntactic annotation

Table 3 summarises the labels used to annotate syntactic flexibility. The last four lines of Table 3 show the FS alternants. The annotators used a manual developed for this purpose.

Label	Phenomenon	Example-Litteral Translation
iwo	connector+vehicle+property	san pupulo elafris (like feather light)
ixp-w	word (not tenor) after property	elafri poli san pupulo (light much like feather)
ixp-n	tenor (only noun) after property	elafri fagito san pupulo (light food like feather)
ixp-creative	words $>$ 1 after property	elafris san puli ochi san pupulo (light like bird not like feather)
mod	modifier of vehicle	elafris san pupulo chielidoniu (light like feather swallow.poss)
empm	morphological emphasis on property	pan-alafro san pupulo (emp-light like feather)
empp	phrasal emphasis on property	poli elafris san pupulo (very light like feather)
mwo	FS combined with other MWE	elafri san pupulo to choma pu ton skepazi[4] (light like feather the soil that him covers)
agr	vehicle agrees with plural tenor	skies elafries san pupula (shadow.pl light like feather.pl)
det	determiner before vehicle	elafris san to pupulo (light like the feather)
var	lexical variation	aspros/lefkos san pani (white like cloth)
comp	πιο adjective (και/κι) από noun	pio elafris ki apo pupulo (more light and from feather)
toso	τόσο adjective όσο/σαν det+noun	toso elafris oso/san to pupulo (as light as the feather)
ixp-punc	punctuation after property	elafris, san pupulo (light, like feather)
constr	**the deterninerless normative form**	**elafris san pupulo** (light like feather)

Table 3: The labels for syntactic flexibility annotation.

To check inter-annotator agreement 630 instances (12,9% of the data), representing 6 FS (FS19, FS18, FS12, FS11, FS9, FS8), were annotated by 3 linguists. Krippendorff's alpha coefficient was equal to 0.91.

3.3.2 Semantic annotation

WordNet supersenses were used to semantically annotate the tenors (Schneider et al., 2013; Schneider and Smith, 2015). No sufficient WordNets are available for Modern Greek, so we translated Greek tenors into English. In case of pronouns or pro-drop phenomena, the tenor was induced from the context.

To check inter-annotator agreement 2400 instances (49% of the data), representing 4 FS (FS20, FS19, FS18, FS1), were annotated by 5 linguists. Krippendorff's alpha coefficient was equal to 0.95.

Figure 1 shows the distribution of semantic categories of tenors in the resource. We see that FS (at least, the ones we studied) mainly describe humans and that they select a small set of semantic categories (Mpouli and Ganascia, 2015), which partly explains the high inter-annotator agreement.

3.4 The Flat-Folia resource

To make the annotated resource[5] machine-readable, extensible and reusable, we converted it to Folia format (van Gompel and Reynaert, 2013), an XML-based format for linguistic annotation. We use the

[4]MWE1: elafri to choma pu ton skiepazi, lit. *light the soil that covers him* (it is said for those who have passed away), FS: elafris san pupulo 'as light as a feather'.

[5]The resource can be accessed at http://glotta.ntua.gr/Similes/.

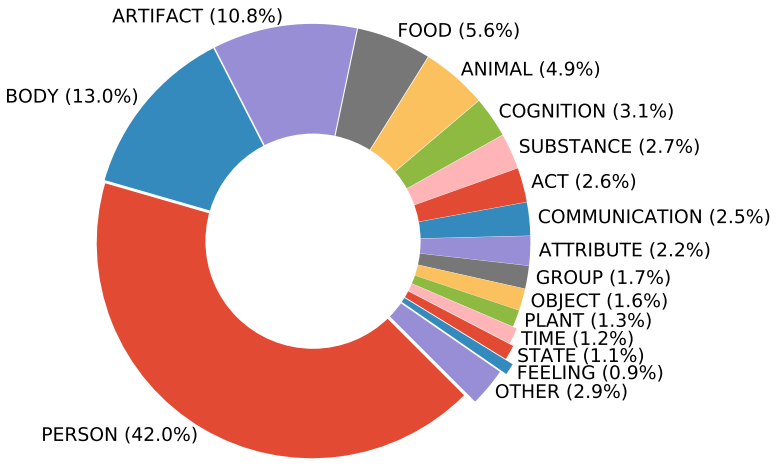


Figure 1: Distribution of semantic categories of tenors in the resource.

Flat - FoliA Linguistic Annotation Tool,[6] which allows users to view and edit Folia annotated documents.

4 Investigating the resource

4.1 The relation between syntactic flexibility and tenor semantics

Shannon entropy is used as a measure of the syntactic and semantic diversity (Manning and Schütze, 1999; Lebani et al., 2015). For the syntactic entropy, each FS instance is represented as a vector including binary values of the syntactic features defined in Table 3 (presence or absence of a feature). Thus, syntactic entropy is measured according to the distinct vectors and their frequency. Semantic entropy is measured by the number of occurrences of each semantic category (Figure 1).

Pearson correlation coefficient (Benesty et al., 2009) showed a strong positive linear relationship between the syntactic and semantic diversity (equal to 0.84 with $p - value = 3.2 \cdot 10^{-6}$), across the FS for the syntactic features *constr, comp, toso, ixp-punc, ixp-creative, ixp-n, empp, mwo*. We observe a positive relation between fixedness and semantic idiomaticity, also observed for V+N MWEs (Fazly and Stevenson, 2007). Correlation improved for the aforementioned syntactic features and deteriorated for the features *iwo, ixp-w, mod, var, det, empm* (Table 3). This result is reminiscent of the (Kay and Sag, 2012, 4) observation that syntactic contexts may or may not interfere with MWE semantics (§2). The relation between the syntactic and semantic diversity of each FS is depicted in Figure 2.

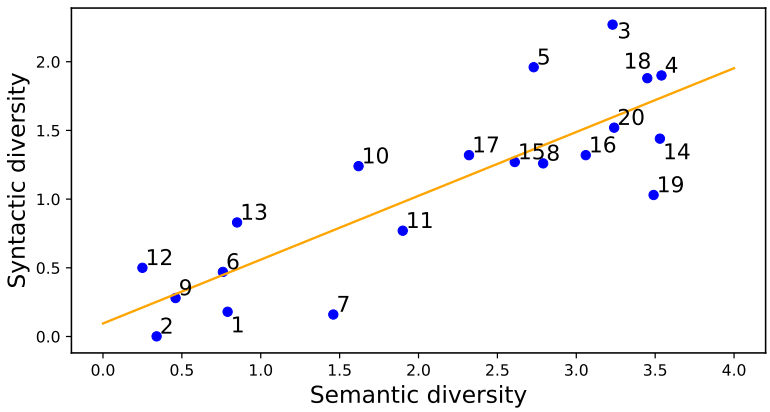
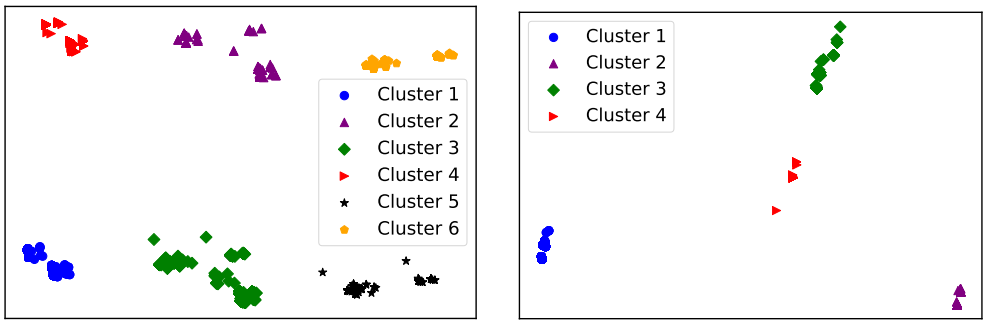


Figure 2: The linear relation between the syntactic and semantic diversity of the similes (the diversity is measured in terms of Shannon entropy).

[6]FLAT is an open source software developed at the Centre of Language and Speech Technology, Radboud University Nijmegen. It can be obtained from https://github.com/proycon/flat

4.2 Classification of FS by syntactic flexibility

Clustering over the vectors including binary values of the syntactic features (presence or absence of a feature) failed to produce an FS classification by syntactic flexibility (such as the one proposed for all MWEs in Sag et al. (2002)); rather, it produced interesting FS instance clusters. We applied Logistic Principal Component Analysis (LPCA) (Landgraf and Lee, 2015) to achieve dimensionality reduction and to visualize the FS instances in two dimensions. The k-Means algorithm[7] is used to separate the data in k clusters of FS instances, where the instances in each cluster are expected to have similar characteristics (Berkhin, 2006). Although the instances could be analyzed in more clusters employing more complicated clustering algorithms, in this initial approach, we identify significant knowledge in a more general separation, as we describe below.



(i) All syntactic features.

(ii) Syntactic features correlating with tenor semantics only.

Figure 3: Clustering FS instances.

When all syntactic features are used, we obtain six clusters (Figure 3(i)) with prominent features *det*, *var*, *comp* and *constr*. Only *constr*'s and *comp*'s syntactic diversity correlates with tenor's semantic diversity (§4.1). The most frequent features by cluster are: cluster 1, *constr* (the normative form); cluster 2, *det*, *var*; cluster 3, *det*; cluster 4, *var*; cluster 5, *det*, *comp*, *constr*; cluster 6, *det*, *comp*, *constr*, *var*.

When only the syntactic flexibility features that contribute positively to the flexibility-idiomatic semantics correlation are used (§4.1), we obtain four clusters (Figure 3(ii)) with prominent features *constr*, *comp*, *ixp-punc* and *toso*. The most frequent features by cluster are: cluster 1, *constr* (the normative form, i.e. absence of any syntactic feature); cluster 2, *comp*, *constr*; cluster 3, *ixp-punc*, *constr*; cluster 4, *toso*, *constr*. The feature *constr* occurs in the clusters 2, 3 & 4, so these clusters can be characterized by their other features *comp*, *ixp-punc* and *toso*. The four structures are the FS alternatives (§3) that more or less partition the set of FS instances (only *ixp-punc* co-occurs with the other FS alternatives, very rarely).

Geeraert et al. (2017), using eyetracking, identified lexical variation (corresponding to our feature *var*) as one of the "easiest" MWE variations for English in terms of comprehension. This might indicate that MWE variations that are "easy" in terms of comprehension are, at the same time, independent of MWE idiomaticity and "more prominent" than other syntactic structures in the speakers' output.

4.3 Predicting syntactic flexibility

To better understand the relation between semantic and syntactic diversity, we sought a quantitative definition of idiomaticity. Intuitively, since FS function as MWE adjectives (§2), rather than defining idiomaticity as the distance of the context of the constituents of the FS from the context of the FS itself (Fazly and Stevenson, 2007), we defined it as the similarity between the tenor semantics and the semantics of the NPs selected by the free property of the FS. For instance, we calculated the similarity between the semantics of the tenors of FS1 άσπρος σαν το πανί (aspros san to pani) 'as white as a sheet' and the semantics of the NPs selected by the free occurrences of the adjective άσπρος (aspros) 'white'. We defined two measures of semantic similarity:

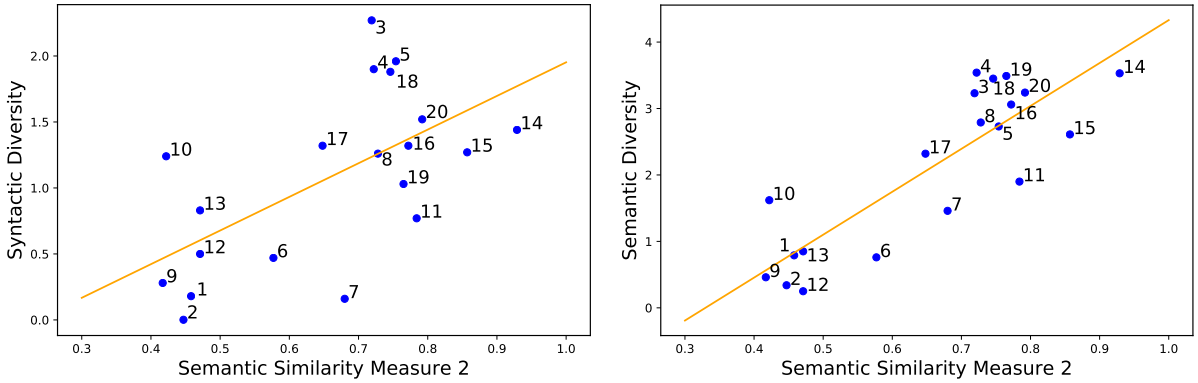
[7] The popular k-means algorithm is sufficient because the obtained clusters have clear limits.

Semantic similarity measure 1 (SSM1): Cosine similarity (Manning and Schütze, 1999) based on frequency is applied to the vectors of the two sets of supersenses described above, which include the frequency of each supersense, for instance, the frequency of the supersense person in the semantics of FS1 and in the semantics of the NPs selected by the FS's free property. To ensure comparability of the vectors, frequencies have been normalized in the range zero to one.

Semantic similarity measure 2 (SSM2): Cosine similarity based on binary vectors: the vectors of both supersence sets which include only the presence or absence (i.e. 1 or 0, respectively) of each supersense. For instance, FS1 selects 7 semantic categories and the respective free adjective selects 17; 5 semantic categories occur in both sets.

Simile	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
SSM 1	.09	.23	.67	.80	.74	.14	.99	.66	.61	.89	.28	.99	.16	.87	.27	.43	.51	.56	.79	.58
SSM 2	.46	.45	.72	.72	.75	.58	.68	.73	.42	.42	.78	.47	.47	.93	.86	.77	.65	.75	.77	.79

Table 4: The *Semantic Similarity Measure 1 (SSM1)* and the *Semantic Similarity Measure 2 (SSM2)* of the semantics of the 20 similes and the supersences of the respective free adjectives.



(i) Relation between *syntactic diversity* and *SSM2*.

(ii) Relation between *semantic diversity* and *SSM2*.

Figure 4: The linear relation between *syntactic diversity* or *semantic diversity* and *semantic similarity measure 2 (SSM2)*.

For each property of the 20 FS, we retrieved the first 200 unique instances from the HNC. We excluded MWEs involving the adjectives, for instance, for the property κόκκινος (kokinos) 'red' we excluded the occurrences of κόκκινη γραμμή (kokini grami) 'red line'[8]. The NPs selected by the adjectives in question were annotated semantically with the WordNet supersenses. The results of both semantic similarity measures are reported in Table 4.

We found that only $SSM2$ is correlated with syntactic diversity and semantic diversity (only syntactic flexibility features correlating with syntactic diversity were used (§4.1)). Pearson correlation coefficient between syntactic diversity and $SSM2$ is equal to 0.61 with $p - value = 4.4 \cdot 10^{-4}$ (Figure 4(i) depicts the correlation) and between *semantic diversity* and $SSM2$, it is equal to 0.85 with $p - value = 2 \cdot 10^{-6}$ (Figure 4(ii) depicts the strong linear correlation). The positive linear correlation between syntactic diversity and $SSM2$ is less strong than between semantic diversity and $SSM2$ (Figure 4). This might indicate that some similes tend to be used in particular syntactic configurations, which interfere with the semantically sensitive syntactic alternants; this is a hypothesis that will be checked with future research.

Because cosine similarity by the semantic categories selected by the FS and its free property correlates well with the syntactic and semantic diversity of FS, it could be used as a predictor of the semantic idiomaticity of FS.

[8]Κόκκινη γραμμή (kokini grami) denotes a boundary or limit which should not be crossed.

5 Conclusion

The quantitative study of the Modern Greek fixed simile (FS) resource adds interesting new aspects to the understanding of the relation between MWE syntactic flexibility and idiomatic semantics and paves the way to further research.

We found that cosine similarity by the semantic categories selected by the FS and its free property, but not cosine similarity by frequency, correlates well with both the syntactic and semantic diversity of FS and it could be used as a predictor of FS syntactic flexibility and idiomaticity. Furthermore, syntactic deviations from the normative form can be split in "FS semantics sensitive " and "FS semantics insensitive" ones. The latter introduce noise in the correlation of idiomaticity and syntactic flexibility and in their correlation with cosine similarity by semantic categories (*SSM2*). Finally, the clustering of FS instances as vectors of syntactic flexibility features showcases some "semantics insensitive syntactic deviations" as dominant features. The same constructions have been shown by V+N MWE comprehension studies to require less comprehension effort than other syntactic constructions. If future research shows that a correlation exists, it may be inferred that, along with idiomatic semantics, syntactic flexibility depends on certain cognitive factors/skills.

Aknowledgements

We thank Katerina Selimi, Dimitra Stasinou, Vasiliki Moutzouri and Maria Chantou for their help at the annotation phase of this work.

References

Ron Artstein and Massimo Poesio. Inter-coder agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596, December 2008.

Timothy Baldwin and Su Nam Kim. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second edition*, pages 267–292. CRC Press, Boca Raton, 2010.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.

Pavel Berkhin. A survey of clustering data mining techniques. In *Grouping multidimensional data*, pages 25–71. Springer, 2006.

Despina Chila-Markopoulou. *Modern Greek comparative constructions: A syntactic analysis of adjectival and adverbial comparatives*. PhD thesis, National and Kapodistrian University of Athens, 1986. (In Greek).

Afsaneh Fazly and Suzanne Stevenson. Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 9–16. Association for Computational Linguistics, June 2007.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103, 2009. ISSN 0891-2017. URL `http://aclweb.org/anthology/J09-1005`.

Kristina Geeraert, R. Harald Bayen Bayen, and John Newman. Understanding idiomatic variation. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, page 80–90, April 2017.

Patrick Hanks. Similes and sets: The English preposition "like". In *Languages and Linguistics: Festschrift for Professor Fr. Čermák*. Philosophy Faculty of the Charles University, Prague, 2005.

Michael Israel, Jennifer Riddle Harding, and Vera Tobin. On simile. In Michel Achard and Suzanne Kemmer, editors, *Language, Culture and Mind*, pages 123–135. CSLI Publications, 2004.

Paul Kay and Ivan A. Sag. A lexical theory of phrasal idioms. Available at: www1.icsi.berkeley.edu/∼kay/idioms-submitted.pdf, 2012.

Andrew J. Landgraf and Yoonkyung Lee. Dimensionality reduction for binary data through the projection of natural parameters. *arXiv preprint arXiv:1510.06112*, 2015.

Eric Laporte. Choosing features for classifying multiword expressions. In Manfred Sailer and Stella Markantonatou, editors, *Multiword expressions: Insights from a multi-lingual perspective*, Phraseology and Multiword Expressions. Language Science Press, 2018.

Gianluca E. Lebani, Marco Senaldi, and Alessandro Lenci. *Modeling idiom variability with entropy and Distributional Semantics*. Universitätsbibliothek Tübingen, 2015.

Christopher D. Manning and Hinrich Schütze. *Foundations of statistical Natural Language Processing*. MIT press, 1999.

Stella Markantonatou, Panagiotis Kouris, Katerina Selimi, Dimitra Stasinou, and Yanis Maistros. Ψυχή άσπρη σαν το χιόνι but never ψυχή άσπρη σαν το γάλα: semasio-syntactic comments on the fixed similes of modern greek. In *Proceedings of the 38th Annual Meeting of the Department of Linguistics, School of Philology, Aristotle University of Thessaloniki, (In Memoriam Michalis Setatos*, Thessaloniki, 2016.

Jelena Mitrović and Stella Markantonatou. A cross-linguistic study on greek and serbian mwes and enrichment of lexical resources via crowdsourcing. In Stella Markantonatou and Anastasia Christofidou, editors, *Multiword expressions in Greek*, volume 15 of *Deltio Epistimonikis Orologias ke Neologismon*. Language Science Press, 2018.

Vasileia Mpolla-Mavridou. *A contrastive study of the fixed similes of the Greek and English languages*. PhD thesis, Aristotle University of Thessaloniki, 1996. In Greek.

Suzanne Mpouli and Jean-Gabriel Ganascia. "pale as death" or "pâle come le mort": Frozen similes used as literary clichés. In *EUROPRHAS 2015: Computerised and Corpus-Based Approaches to Phraseology: Monolingual and Multilingual Perspectives*, 2015.

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. Brighter than gold: Figurative language in user generated comparisons. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2008–2018. Association for Computational Linguistics, October 2014.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. Idioms. *Language*, 70(3):491–538, 1994.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*, volume 2276/2010 of *CICLing '02*, pages 1–15, London, UK, 2002. Springer-Verlag. ISBN 3-540-43219-1.

Nathan Schneider and Noah A. Smith. A corpus and model integrating multiword expressions and supersenses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1537–1547. Association for Computational Linguistics, May–June 2015. URL `http://www.aclweb.org/anthology/N15-1177`.

Nathan Schneider, Behrang Mohit, Kemal Oflazer, and Noah A. Smith. Coarse lexical semantic annotation with supersenses: An Arabic case study. In *Proceedings of NAACL-HLT 2013*, pages 661–667. Association for Computational Linguistics, June 2013.

Maarten van Gompel and Martin Reynaert. Folia: A practical xml format for linguistic annotation-a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3:63–81, 2013.

Tony Veale and Yanfen Hao. Learning to understand figurative language: From similes to metaphor to irony. In *Proceedings of the 29th Annual Meeting of the Cognitive Science Society (CogSci 2007)*, 2007.