Associazione Italiana di Linguistica Computazionale

Emerging Topics at the
Third Italian Conference on Computational Linguistics
and EVALITA 2016

CONTENTS

Distributed Representations of Lexical Sets and Prototypes in Causal Alternation Verbs

Edoardo Maria Ponti*
Universitá di Cambridge

Elisabetta Jezek**
Università degli Studi di Pavia

Bernardo Magnini†
Fondazione Bruno Kessler

Lexical sets contain the words filling an argument slot of a verb, and are in part determined by selectional preferences. The purpose of this paper is to unravel the properties of lexical sets through distributional semantics. We investigate 1) whether lexical set behave as prototypical categories with a centre and a periphery; 2) whether they are polymorphic, i.e. composed by subcategories; 3) whether the distance between lexical sets of different arguments is explanatory of verb properties. In particular, our case study are lexical sets of causative-inchoative verbs in Italian. Having studied several vector models, we find that 1) based on spatial distance from the centroid, object fillers are scattered uniformly across the category, whereas intransitive subject fillers lie on its edge; 2) a correlation exists between the amount of verb senses and that of clusters discovered automatically, especially for intransitive subjects; 3) the distance between the centroids of object and intransitive subject is correlated with other properties of verbs, such as their cross-lingual tendency to appear in the intransitive pattern rather than transitive one. This paper is noncommittal with respect to the hypothesis that this connection is underpinned by a semantic reason, namely the spontaneity of the event denoted by the verb.

1. Introduction

The arguments of a verb are the "slots" that have to be filled to satisfy its valency (subject, object, etc.). Hence, verbs display so-called selectional restrictions over the possible fillers occupying these slots (Séaghdha and Korhonen 2014), which play a major role in determining the verb meaning. Moreover, the selection of the fillers happens in accordance with the different senses of a verb. Lists of the possible fillers occurring with a certain pattern of a verb can be collected in a corpus-driven fashion: these lists are named "lexical sets" (Hanks 1996).

Several approaches in Computational Linguistics managed to inspect lexical sets and their patterns of variation (Montemagni, Ruimy, and Pirrelli 1995; McKoon and Macfarland 2000), as well as to use them as features for verb classification (McCarthy 2000; Joanis, Stevenson, and James 2008). On the other hand, selectional preferences were employed for many tasks in Natural Language Processing, including Word Sense Disambiguation (Resnik 1997; McCarthy and Carroll 2003, *inter alia*), Metaphor Processing (Shutova, Teufel, and Korhonen 2013), Information Extraction (Pantel et al.

 * E-mail: ep490@cam.ac.uk
** E-mail: jezek@unipv.it
 † E-mail: magnini@fbk.eu

2007), Discourse Relation Classification (Ponti and Korhonen 2017), Dependency Parsing (Mirroshandel and Nasr 2016), and Semantic Role labeling (Gildea and Jurafsky 2002; Zapirain et al. 2013).

We aim at establishing a new method of analysis for lexical sets. In particular, we address the following key questions: are lexical sets Aristotelian categories (yes-no membership) or prototypical categories (graded membership)? If the latter holds true, are the fillers arranged homogeneously or do they cluster around some sub-categories? Finally, is the relation between lexical sets from different patterns of a same verb informative about this verb's meaning? We address these questions under a distributional semantics perspective. In fact, distributional semantics provides a mapping between each filler and a vector lying in a continuous multi-dimensional space. By virtue of this, lexical sets can be treated as continuous categories, where members can be either central or peripheral. Moreover, this allows to quantify the distance between vectors with spatial measures.

In order to test the relevance for linguistic theory of this approach, we focus on a case study, namely verbs undergoing the causative-inchoative alternation. Solving the above-mentioned issues may help clarifying some of the vexed questions about this class of verbs. These show both transitive and intransitive patterns: the object of the former and the subject of the latter play the same semantic role of patient. Based on the cross-lingual frequency of each of these patterns and the direction of morphological derivation, it is possible to sort these verbs onto a scale (Haspelmath 1993; Samardzic and Merlo 2012; Haspelmath et al. 2014), which can be possibly interpreted semantically as the "spontaneity" of the corresponding event (see § 2.2). We investigate the existence of any asymmetry between the lexical sets of the transitive object and the intransitive subject, and if so the connection of this asymmetry with the spontaneity scale.

The structure of the paper is as follows. In § 2, we define the core notions of this study, including lexical sets, causative-inchoative verbs, and distributional semantics. § 3 presents the method and the data, whereas § 4 reports the results of the experiments. Finally, § 5 draws the conclusions and § 6 proposes possible future lines of research.

2. Definitions and Previous Work

In this section, we describe in detail the notions underlying the subject matter of the research (lexical sets, § 2.1), the case study (causative-inchoative verbs, § 2.2), and the method (distributional semantics, § 2.3). At the same time, we present the previous work concerning each of these aspects.

2.1 Lexical Sets

A lexical set can be defined as the set of words that occupy a specific argument position within a single verb sense, such as {*gun*, *bullet*, *shot*, *projectile*, *rifle*...} for the sense 'to shoot' of *to fire*, or {*employer*, *teacher*, *attorney*, *manager*...} for its sense 'to dismiss'. The notion of lexical set was firstly introduced by Hanks (1996). Its purpose is explaining how the semantics of verbs is affected by the patterns of complements they are found with. Hanks' approach is justified by the pervasiveness of patterns in corpora: these patterns are instantiated by specific lexical items typically occurring in the argument positions. These items, called fillers, form sets belonging to different patterns of meaning. Hanks and Pustejovsky (2005) and Hanks and Jezek (2008) propose an ontology where fillers are clustered into semantic types, i.e. categories such as [[Location]], [[Event]], [[Vehicle]], [[Emotion]]. These form a hierarchy that branches into more specific types.

However, these categories are problematic, as lexical sets tend to "shimmer" (Jezek and Hanks 2010): their membership tends to change according to the verb they associate with. The shimmering nature of lexical sets is not an accidental phenomenon. Rather, it stems from the fact that verb selectional restrictions may cut across conceptual categories due to the specific predication introduced by the verb. For example, both *wash* and *amputate* typically select [[Body Part]] as their direct object. Nevertheless, they select different prototypical members of the set, as can be seen in the examples below from Jezek and Hanks (2010) where only shared members are underlined:

(1) [[Human]] *wash* [[Body Part]]
    where [[Body Part]]: {<u>hand</u>, hair, face, foot, mouth...}

(2) [[Human]] *amputate* [[Body Part]]
    where [[Body Part]]: {leg, limb, arm, <u>hand</u>, finger...}

Lexical sets can be extracted from corpora automatically. This operation hinges upon traditional techniques for the acquisition of subcategorisation frames and selectional restrictions. The former allow to capture the syntactic pattern (Brent 1991; Schulte Im Walde 2000) or semantic frame pattern (Baker, Fillmore, and Lowe 1998) in which each verb is found. The latter create probabilistic models of preferences over fillers that are evaluated intrinsically through human judgments (Brockmann and Lapata 2003) and extrinsically through disambiguation tasks (Van de Cruys 2014).

2.2 Causative-Inchoative Verbs

In principle, lexical sets can be constructed for every verb. In this work, however, we limit our inquiry to a specific subset of verbs, namely causative-inchoative verbs in Italian. This provides a testbed for our methods of analysis, which can be easily extended to other classes of verbs and alternations. The choice of this specific subset is due to the fact that understanding the internal structure of lexical sets and their relations seems to be crucial to solve the problems surrounding this class of verbs, including asymmetries between transitive objects and intransitive subjects and their relation with the spontaneity scale (see below).

Causative-inchoative verbs alternate. In other terms, they appear in two patterns, either as transitive or intransitive. In the former, an agent brings about a change of state affecting a patient; in the latter, the change of the same patient is presented as spontaneous (e.g. *break*, as in "Mary broke the key" vs. "the key broke"). The verbs in the two patterns can be expressed by either a same form or two distinct forms cross-lingually. In the second case, the forms can be morphologically asymmetrical: one has a derivative affix and the other does not. Otherwise, the forms are suppletive, being completely unrelated (e.g. *kill/die*). The members of causative-inchoative verbs that retain a same form or are morphologically related in both patterns vary cross-lingually (Montemagni, Ruimy, and Pirrelli 1995). Alternations regarding physical change-of-state and manner-of-motion are found in English, whereas they are limited to psychological and physical changes-of-state in Italian. In Japanese and Salish languages, also verbs like *arrive* and *appear* do alternate (Alexiadou 2010).

From a semantic point of view, Italian causative-inchoative verbs are required to be telic and have an inanimate patient (Cennamo 1995). Morpho-syntactically, they are generated from an asymmetrical derivation, called "anti-causativisation." The intransitive form is sometimes marked with the pronominal clitic *si*: its presence is mandatory,

optional or forbidden according to verb-specific rules (Cennamo and Jezek 2011). Because of this, many different categorisations of Italian causative-inchoative verbs were attempted (Folli 2002; Jezek 2003).

Causative-inchoative verbs in general are endowed with peculiar properties. Haspelmath (1993) claims that verbs with a cross-lingual preference for a marked causative form denote a more "spontaneous" situation. Spontaneity is intended by the author as the likelihood that the occurrence of the event denoted by the verb does not require the intervention of an agent. In this way, a correlation between the form and the meaning of these verbs was borne out. Moreover, Samardzic and Merlo (2012) and Haspelmath et al. (2014) demonstrated that verbs appearing more frequently (intra- and cross-lingually) in the inchoative form tend to morphologically derive the causative form. Here, the correlation bridges between form and frequency. Vice versa, situations entailing an agentive participation prefer to mark the inchoative form and occur more frequently in the causative form.

However, what remains uncertain is whether spontaneous and agentive variants of the same verbs differ in their lexical sets. Atkins and Levin (1995) and Levin and Hovav (1995) argued that selectional restrictions for spontaneous verbs (named internal causation verbs) are stricter because their event unfolds due to some inherent property of the patient: referents without this capability are excluded, contrary to what happens with agentive verbs (named external causation verbs). This capability is defined "teleological" by (Copley and Wolff 2014). However, McKoon and Macfarland (2000) reported contradicting results from corpus-based analyses that did not find any significant difference in the content of lexical sets of spontaneous and agentive verbs (although from a sample of less than 100 sentences and only 5 categories). Instead, they reported a difference between transitive objects and intransitive objects in that the former contained a larger amount of abstract nouns compared to the latter.

2.3 Distributional Semantics

Once established the domain, we need to provide a reliable method of inquiry. Previous works based on set theory treated lexical sets as Aristotelian categories, of which a filler is either a member or not. For instance, Montemagni, Ruimy, and Pirrelli (1995) collected lexical sets manually and employed set intersection as a measure of similarity between two sets. Research in psychology, however, has long since demonstrated that the members of a linguistic set are found in a radial continuum where the most central element is the prototype for its category, and those at the periphery are less representative (Rosch 1973; Lakoff 1987).[1]

The full exploitation of the semantic information inherent to the argument fillers of verbs can take advantage of some recent developments in distributional semantics. Efficient algorithms have been devised to map each word of a vocabulary into a corresponding vector of n real numbers, which can be thought as a sequence of coordinates in a n-dimensional space (Mikolov et al. 2013a, *inter alia*). This mapping is yielded by unsupervised machine learning, based on the assumption that the meaning of a word can be inferred by its context, i.e. its neighbouring words in texts. This is known as Distributional Hypothesis (Harris 1954; Firth 1957). Distributed models have some relevant properties: first, the geometric closeness of two vectors corresponds to the similarity

---

[1] For previous work on lexical sets considering prototypicality in the context of the notion of shimmering, see Jezek and Hanks (2010).

in meaning of the corresponding words. Moreover, its dimensions possibly retain a semantic interpretation such that non-trivial analogies can be established among words.

Word vectors allow to capture the spatial continuum implied by the notion of prototype. Previous works showed that word vectors can be clustered to imitate linguistic categories, because each cluster captures the 'semantic landscape' of a word (Hilpert and Perek 2015). The center of these clusters can be interpreted as the prototype of the corresponding category, and the proximity of the cluster members to the center as the degree of their prototypicality. In fact, the cluster members are not scattered randomly, but rather are arranged according to the internal structure of the cluster (Dubossarsky, Weinshall, and Grossman 2016). The prototypicality of the cluster members provides an explanation about linguistic phenomena, such as the resistance to the diachronic change of meaning (Geeraerts 1997; Dubossarsky et al. 2015).

In this work, we extend the usage of the notion of prototypicality from meaning-based categories derived through vector quantization to grammatically defined categories (i.e. lexical sets) derived through dependency parsing. Moreover, we go beyond the estimation of the distance of each word from the centroid of its category: in particular, we propose new methods to assess the internal diversity in terms of sub-categories and the distance between lexical sets themselves.

Table 1
List of 20 causative-inchoative verbs and count of their fillers for each argument slot.

Lemma	Translation	S	O
chiuder(si)	to close	289	606
aprir(si)	to open	195	1337
aumentare	to improve	534	998
romper(si)	to break	83	419
riempir(si)	to fill	58	166
raccoglier(si)	to gather	85	505
connetter(si)	to connect	39	134
divider(si)	to split	129	246
finire	to stop	1092	721
uscire/portare fuori	to go/put out	325	638
alzar(si)	to arise/raise	75	304
scuoter(si)	to rock	10	69
bruciare	to burn	75	174
congelare	to freeze	10	30
girare	to spin	155	243
seccare	to dry	15	14
svegliar(si)	to awake/wake	68	89
scioglier(si)	to melt	94	143
(far) bollire	to boil	2	2
affondare	to sink	18	74

3. Data and Method

We sourced the lexical sets from a sample of ItWac, a wide corpus gathered by crawling texts from the Italian domain in the web using medium frequency vocabulary as seeds (Baroni et al. 2009). This sample was further enriched with morpho-syntactic information through the graph-based MATE-tools parser (Bohnet 2010). We trained and evaluated this parser on the Italian treebank inside the collection of Universal Dependencies v1.3 (Nivre et al. 2016). The evaluation on gold standard data suggests how many errors we expect to affect the predictions on the new data, i.e. the ItWac corpus: these errors are then propagated to the following steps for the extraction of lexical sets. According to the LAS metric, the relevant dependency relations scored 0.751 for *dobj* (direct object), 0.719 for *nsubj* (subject), and 0.691 for *nsubjpass* (subject of a passive verb). A target group of 20 causative-inchoative verbs was taken from Haspelmath et al. (2014): they are listed in Table 1, together with the count of the extracted lexical sets for the relevant semantic macro-roles (see below).

Once the sentences were parsed, the target verbs were identified inside the dependency trees. The lemmas of these verbs and the forms of their arguments were stored in a database. Argument fillers were grouped according to the semantic macro-roles defined by Dixon (1994), rather than their dependency relations: subjects of transitive verbs (A), subjects of intransitive verbs (S) and objects (O). In particular, the subjects of verb forms accompanied by the *si*-clitic and those without an object depending on the same verb were treated as S.[2] These operations resulted in a database structured as follows: in each row, a verb is alongside of the fillers it occurs with in a specific sentence. For example, compare an original sentence and its corresponding entry in Example 3:

(3) **Plinio** il Vecchio non **cita** più il **Po** di Adria
Pliny the Elder doesn't mention anymore the Po of Adria
*perché l' **argine** dell' Adige si era **rotto** ed era*
because the bank of the Adige had broken and had
confluito *nella Filistina.*
merged with the Filistina.

Verb	A	S	O
citare	Plinio	_	Po
rompere		argine	
confluire	_	_	_

Because of the nature of vector models, we made the following design choices to deal with special linguistic phenomena. We discarded everything but the head of multi-word nouns, such as *Plinio* of *Plinio il Vecchio*, to preserve the one-to-one mapping between words and vectors. We did not distinguish proper and common nouns, such as *Po* and *argine*, since their representations lie in the same multi-dimensional space. Subjects in ellipsis or co-reference were neglected, since no pragmatic annotation of the sentences was available: for instance, *Adige* should appear as S in the entry for *confluire*, which is left blank instead. Finally, polysemous words and homonyms were

---

2 Subjects of verbs inflected in the passive voice were treated as O, instead.

represented by a single form, hence a single vector, since their representations conflate all the relevant meanings. For instance, *Adige* is both a river and a location, but these are not distinguished in the vector model.

The database was later collapsed by verb lemma so that each of them became associated to three sets of fillers (one per macro-role). Each of these sets is a corpus-based lexical set. Compared to manually picked lexical sets, they are more noisy but less sparse: the vastness of the data mitigates the errors in the parsing step. Moreover, the automation in lexical set extraction allows to access the fillers of virtually every verb: resources based on manual selection like T-PAS (Jezek et al. 2014), on the other hand, are limited to a small amount of verbs.

Afterwards, each of the argument fillers was mapped to a vector according to three different pre-trained models. The vectors are generated unsupervisedly back-propagating the gradient from the loss of a task to update randomly initialized embeddings. Each model, however, relies on different tasks:

- **CBOW** stands for Continuous Bag of Words and is part of the Word2Vec suite (Mikolov et al. 2013b). The Italian model was developed by Dinu, Lazaridou, and Baroni (2015) through negative sampling. 300-dimensional representations were obtained by training a binary classifier that discriminates whether a pair of a target word w and a context c belongs to an actual text or not. True contexts are obtained from a window of 5 words on either side of the target word, false contexts are drawn randomly from a noise distribution. Moreover, infrequent words are pruned. On the other hand, a preprocessing step called subsampling deletes words from windows whose frequency f is higher than an threshold with a probability $p = 1 - \sqrt{\frac{t}{f}}$. The text from which word-context pairs are sampled is the full ItWac corpus, consisting in 1.6 billion tokens. The representations resulting from this algorithm are organised by topical similarity or relatedness.

- **fastText** embeddings were trained on the dump of the Italian Wikipedia through a character-level skip-gram (Bojanowski et al. 2017). The skip-gram algorithm predicts the following element given a sequence of previous elements. The elements of Bojanowski's version are characters: each word is represented by a bag of character n-grams and its representation consists in the sum of the representations of each of them. The vector dimensionality is 300; random sequences are drawn with a ratio of 5 to 1 compared to true sequences, with a probability proportional to the square root of the uni-gram frequency. The size of the context window is uniformly sampled from values between 1 and 5. The rejection threshold for subsampling is 10^{-4}. The usage of character n-grams makes this vector model sensitive to morphological similarity.

- **Polyglot** vectors result from a classifier distinguishing between sequences taken from the dump of the Italian Wikipedia and corrupted sequences (Al-Rfou, Perozzi, and Skiena 2013). The window of phrases was 5, the dimensionality of vectors 64. Subsampling discarded words not appearing in the raking of the 100 thousand most frequent ones.

In order to measure spatial distances inside these vector models, many different metrics are available, including pure geometrical (Euclidean) distance. In this work, we

rely on the popular metrics of cosine distance (Cha 2007). Assume that **a** and **b** are vectors, and that $\mathbf{a}_i$ and $\mathbf{b}_i$ are their i^{th} components, respectively. The cosine similarity between **a** and **b** is then defined as follows:

$$\cos(\theta) = \frac{\sum_{i=1}^{n} \mathbf{a}_i \mathbf{b}_i}{\sqrt{\sum_{i=1}^{n} \mathbf{a}_i^2}\sqrt{\sum_{i=1}^{n} \mathbf{b}_i^2}} \tag{1}$$

The opposite, namely cosine distance d, is simply defined as $1 - \cos(\theta)$. As for the values that $d(\mathbf{a},\mathbf{b})$ can assume, the minimum is at 0 (angles completely overlap) and the maximum is at 1 (orthogonal vectors).

4. Experiments

In order to answer the questions outlined in the introduction, we devised three experiments. In all of them, we assume that the S and O lexical sets might not overlap, as borne out by Montemagni, Ruimy, and Pirrelli (1995) and McCarthy (2000). The first (§ 4.1) investigates the internal structure of lexical sets and how their members aggregate around a prototype. Based on psycholinguistic theories (see § 2.3), we expect vectors to lie on a continuum. As for causative-inchoative verbs in particular, objects were shown to be more homogeneous than subjects (McKoon and Macfarland 2000), hence they should form denser clusters. The second experiment (§ 4.2) deals with polymorphism of lexical sets, i.e. the amount of distinct sub-categories they contain. We expect the number of sub-clusters to be proportional to the number of verb senses. Finally, the third experiment (§ 4.3) is aimed at studying how lexical sets of different arguments of the same verb are related. In particular, we predict that the distance between intransitive subject and transitive object varies depending on the spontaneity of the causative-inchoative verb (see § 2.2). In fact, intransitive subjects and transitive objects of spontaneous verbs should be limited to referents with "teleological capability" (Atkins and Levin 1995; Levin and Hovav 1995), hence creating dense groups possibly located far from each other. Fillers of agentive verbs instead should show a higher degree of overlap because of their looser constraints.

4.1 Prototypicality: Distance from Centroid

Once the fillers have been mapped to their corresponding vectors, a lexical set appears as a group of points in a multi-dimensional model. The centre of this group is the Euclidean mean among the vectors, which is a vector itself and is called centroid. In the first experiment, we measure the cosine distance of every vector member of a lexical set from the centroid estimated from all the other members.[3] This leave-one-out setting aims at avoiding biases due to outliers (such as phraseological usages or misspellings). In semantic terms, this measure should correspond to assessing how far a filler is from its prototype.

We obtained two sets (S and O) of cosine distance values for each verb: these can be plotted as boxes and whiskers, like in Figure 1. The cosine distances are represented as a scatter plot: on its side, a box informs about the average (horizontal line), the median[4]

---

3 Every filler was weighted proportionally to its absolute frequency.
4 The median is the value separating the higher half of the ordered values from the lower half.
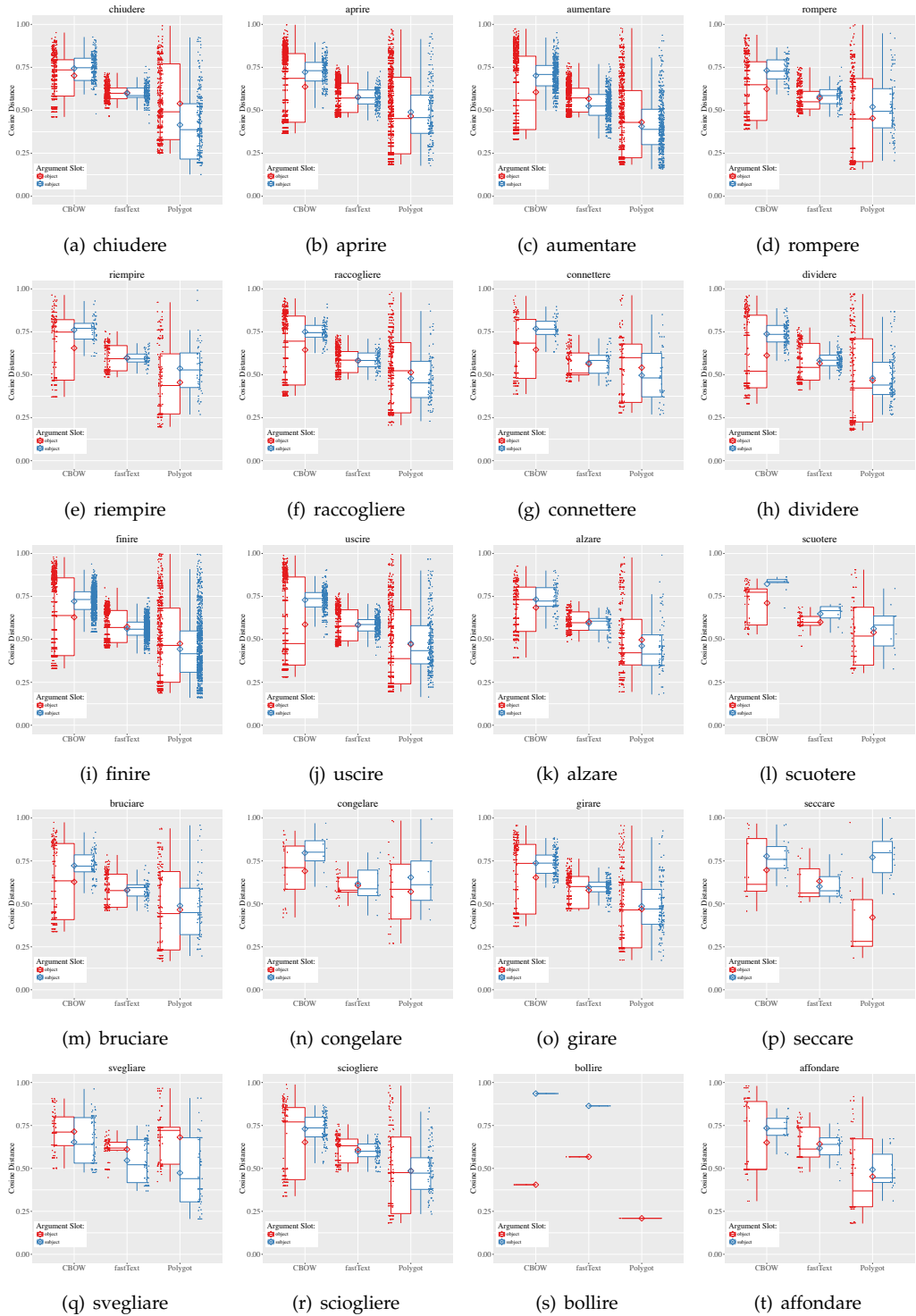
Figure 1
Boxes and whiskers of vector distances from centroid. Lexical sets for objects are in red, for subjects in blue. The columns in each subplot represent different models: from left to right, CBOW, fastText, and Polyglot.

(diamond), the second and third quartiles (rectangle), and the extremes (bars) of the distribution.

Firstly, we observe a same trend for both intransitive subjects and objects that depends on the algorithms. The whole bars tend to hover around higher values for CBOW, and lower values for Polyglot. fastText instead lies somewhere in between them, and its central quartiles cover a short span of values. Secondly, there is a systematic gap between the medians for S and O. Excluding *bollire*, for which data are insufficient, the object is lower 18 out of 19 times for CBOW. On the other hand, for Polyglot the preference is very mild (11 out of 18) and no clear pattern emerges for fastText.

4.2 Polymorphism: Sub-Clusters

The distance from the prototype examined in § 4.1 does not account for the complex internal structure of a lexical set, which aggregates several sub-categories. Figure 2 shows a visual example of this sub-organisation by plotting the heatmaps of the density of vectors reduced to 2 dimensions through t-SNE (Maaten and Hinton 2008): it is possible to observe spots in isolation and aggregation. In order to assess the polymorphism of

(a) aumentare CBOW (b) aumentare fastText (c) aumentare Polyglot

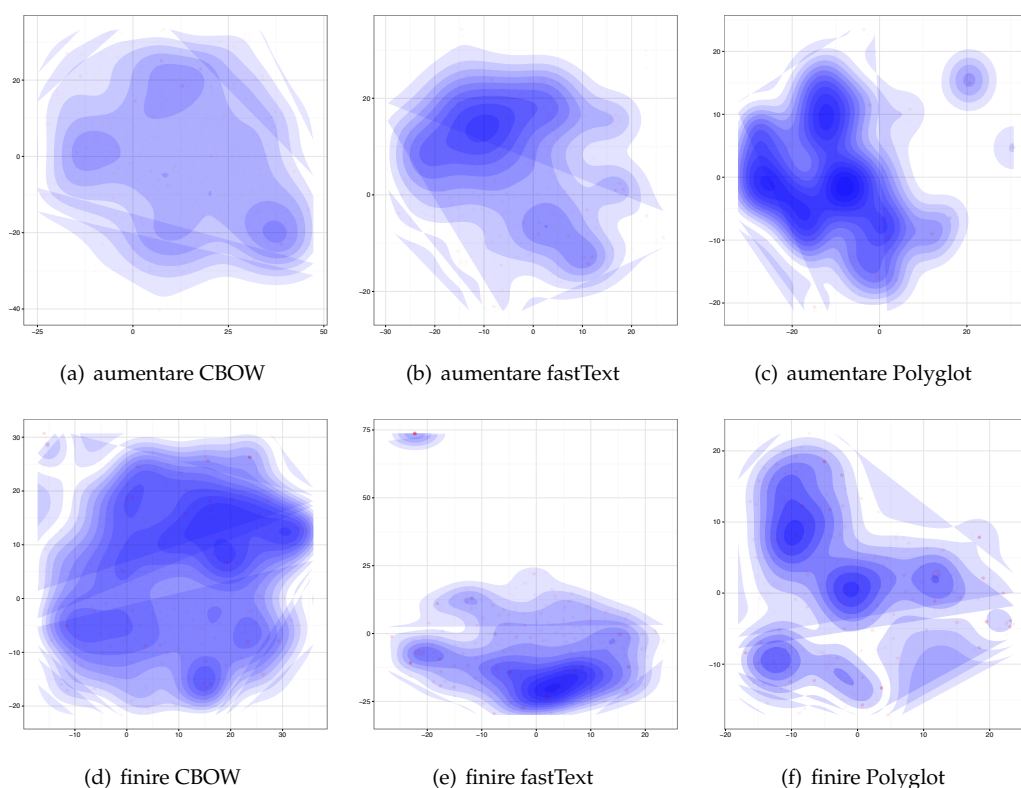(d) finire CBOW (e) finire fastText (f) finire Polyglot

Figure 2
Heatmaps of the density (low is white and high is blue) of vectors reduced to 2 dimensions through t-SNE. They belong to the S lexical sets of *aumentare* (4 senses) and *finire* (12 senses).

lexical sets, i.e. the amount of different shades of meaning they contain, we cluster the vectors in each of them and contrast this measure with the number of the verb senses.

Vector clustering is performed through the algorithm X-Means (Pelleg and Moore 2000), an extension of the simpler algorithm K-Means (MacQueen 1967). The latter performs vector quantisation: it assigns vectors in a space to a pre-specified amount of k clusters. The algorithm converges into a local optimum by initialising k arbitrary means inside the model. Then it carries on iteratively two steps: firstly it assigns each vector to these means minimising a measure of variance. Secondly, it calculates the new means based on the newly obtained clusters. Thus, for vectors $\mathbf{v}$ clusters C and centroids μ, its objective can be formalised as:

$$\underset{C}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{v} \in C_i} \|\mathbf{v} - \mu_i\| \tag{2}$$

In addition to these steps, X-Means can perform a further operation, namely deciding if and where splitting clusters in two to create new clusters. Iterations start from $k = 2$ up to a pre-specified upper bound (in our case, $k = 20$): after a split, X-Means estimates through a test whether the old cluster or the two new clusters fit the data better: in our case, this test consists in the Bayesian Information Criterion, which is defined as follows: given a split M, the maximised value of the log-likelihood $\hat{L}$ of the vectors V, the number of parameters k, and the number of data points n, $\mathrm{BIC}(M) = \hat{L}(V) - \frac{k}{2} \ln n$.

The number of clusters provided by X-Means upon convergence is displayed in Table 2. Moreover, we provide the number of senses of each verb according to WordNet for Italian (Artale, Magnini, and Strapparava 1997) in a separate column. At first glance, the results show that lexical sets tend to have a similar number of clusters across the algorithms, which is surprising considering the different natures of these representations. However, this might be due to a bias: in fact, X-Means is possibly inclined to make similar decisions for sets of identical cardinality and with a low upper bound.

We estimated the Pearson's correlation between the number of clusters and verb senses, which is reported in the bottom column: values of its coefficient mirror the strength of the correlations, ranging from -1 (negative), across 0 (absent), to 1 (positive). The p-value instead stands for the confidence by which we can exclude the null hypothesis (absence of correlation). Results for CBOW and fastText reveal a mild positive correlation. This is especially evident for subjects, which appear to be a better cue for predicting the number of verb senses. On the other hand, results are inconclusive for Polyglot, as the p-value is not significant enough.

4.3 Spontaneity: Distance between Centroids

In §§ 4.1–4.2, lexical sets of the same verb have been considered independently from each other. We now assess whether any relation holds between them by gauging the cosine distance between the centroid of S and the centroid of O for each verb. This operation is aimed at finding to which extent these two lexical sets overlap and unveiling possible asymmetries. In order to estimate whether spontaneity (see § 2.2) affects this degree of overlap, we compared the ranking of our sample of verbs according to the ratio of the cross-lingual frequency of their transitive and intransitive forms (Haspelmath et al. 2014) and a ranking based on their centroid distances.

Table 2
Number of verb senses and clusters inside lexical sets upon convergence of X-Means.

Lemma	Senses	CBOW		fastText		Polyglot	
		S	O	S	O	S	O
chiudere	9	10	13	15	13	10	13
aprire	5	10	16	12	16	12	12
aumentare	4	16	14	15	14	16	16
rompere	6	9	14	10	14	10	13
riempire	3	2	10	10	15	8	14
raccogliere	6	2	15	8	14	2	14
connettere	3	2	8	9	10	3	7
dividere	2	12	11	13	13	12	11
finire	12	16	15	16	16	16	14
uscire	12	19	15	14	14	15	13
alzare	6	6	12	10	16	11	13
scuotere	1	1	10	1	10	1	8
bruciare	11	8	13	13	14	2	11
congelare	4	2	2	4	2	2	2
girare	6	11	11	12	13	13	13
seccare	4	2	5	8	7	2	3
svegliare	1	7	8	9	7	10	8
sciogliere	7	14	10	16	12	14	11
bollire	2	-	-	-	-	-	-
affondare	6	5	6	7	12	6	7
Pearson's correlation		0.572	0.493	0.596	0.458	0.326	0.389
p-value		0.011	0.032	0.007	0.049	0.173	0.100

In Figure 3, we plot the ranking based on cross-lingual frequency against the cosine distance: it emerges that the latter tends to increase for more spontaneous verbs, i.e. verbs with a preference for the intransitive pattern. This tendency is straightforward for all the vector models, as the LOESS line suggests. Moreover, after ranking verbs based on cosine distance between S and O lexical sets, we estimated two correlation metrics with respect to the frequency-based ranking: Spearman and Kendall. Their coefficients with the corresponding p-values are reported in Table 3: they demonstrate a mild-strong positive correlation for both CBOW and Polyglot. However, the p-value does not allow to exclude the null hypothesis (absence of correlation) for fastText with reasonable certainty.

5. Discussion

The questions at the heart of these experiments were: how are lexical set structured? In particular, do their elements distribute uniformly in the space, or rather gather together (near or far from the prototype)? Are they polymorphic, i.e. composed by more than one sub-category? Moreover, which is the degree of overlap between lexical sets of different
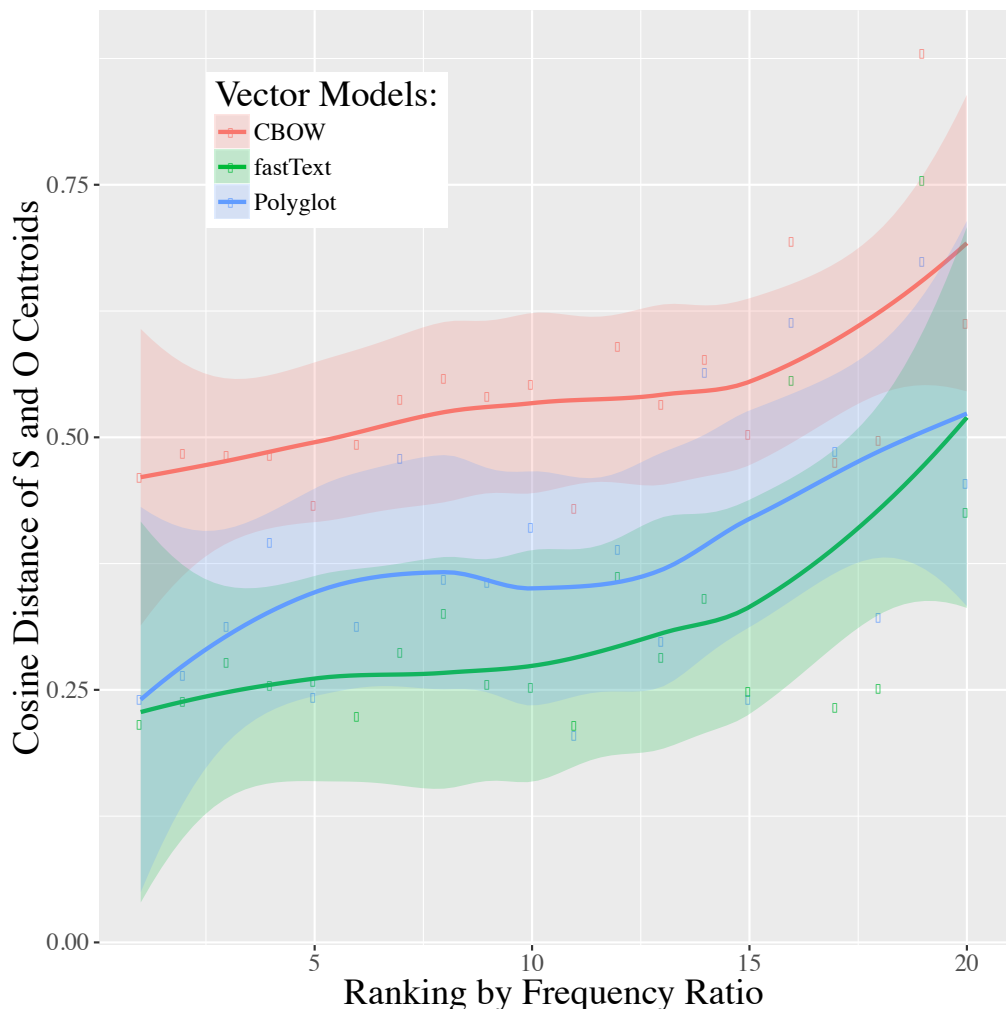
Figure 3
On the x-axis, ranking based on cross-lingual form frequencies as reported by Haspelmath et al. (2014). On the y-axis, cosine distances between centroid of S and O lexical sets in Italian. Lines are LOESS regressions, and shaded areas their confidence regions.

argument slots? In this section, we discuss the answers inferred from the results, and analyse the specific behaviour of every vector model.

In the first experiment (§ 4.1), the members of O lexical sets are scattered from the centre to the periphery. On the other hand, the members of S lexical sets lie in a more compact range of distances, mostly farther from the centroid. This implies that O behaves more similarly to a radial category, whereas S just populates the periphery. From a linguistic point of view, this means that the content of O is more homogeneous, whereas S is more heterogeneous: this finding is in line with previous work (McKoon and Macfarland 2000). In the second experiment (§ 4.2), we observed a mild positive correlation between the number of clusters and of the verb senses. Hence, it is possible

Table 3
Spearman and Kendall correlations between a ranking based on lexical set distance in Italian and another based on cross-lingual frequency ratio.

|	CBOW	fastText	Polyglot
Spearman	0.560	0.420	0.490
p-value	0.010	0.069	0.028
Kendall	0.411	0.305	0.358
p-value	0.012	0.064	0.030

to conclude that polymorphic verbs accept more categories of referents as their possible argument fillers. This holds true especially for S.

In the third experiment (§ 4.3), we established the existence of a correlation between the verb ranking based on the cross-lingual ratio of intransitive and transitive verbs and the ranking based on cosine distance between S and O centroids in Italian. From a linguistic point of view, this is possibly due to the constraints on referents of spontaneous verbs (Atkins and Levin 1995; Levin and Hovav 1995), called teleological capability (Copley and Wolff 2014): this makes the sets clear-cut and possibly far from each other. This adds another piece to the puzzle of the so-called spontaneity scale: Figure 4 shows a synopsis of our result in the context of the correlations established in previous works. Solid lines stand for correlations proven based on corpus evidence. The dotted line, on the other hand, suggests the existence of and underlying motivation for the correlations (i.e. spontaneity), which nonetheless remains unproven and undetermined in its nature. Its possible validation is left to future research, but remains tricky due to its purely semantic nature.

The conclusions can be drawn from more than one vector model, although results are not significant for all of them. In particular, fastText does not show any tendency with respect to the distribution of distances, nor it is possible to exclude the null hypothesis for the correlation between distance of S and O centres and frequency of the intransitive pattern. This might be due to the fact that fastText is based on characters, hence on morphological information, rather than capturing topic relatedness. Moreover, there are possible biases that plague the experiments. Firstly, the homogeneity of the O lexical set might be an artifact because the sample of objects is usually wider and hence more representative. Instead, the heterogeneity of the S lexical set is in part due to its method of extraction: sometimes also transitive subjects (A) are treated as S, because of either unexpressed objects or parsing mistakes.

6. Conclusions and Future Work

Our work provided evidence that lexical sets of Italian causative-inchoative verbs are non-uniform categories, whose distribution around the prototype varies to a great extent. This distribution is sensitive to the argument slot: transitive objects display a more uniform distribution of distances from the prototype, whereas the fillers of intransitive subjects lie on the edge of the category. This difference might be due to different selectional restrictions applied to the object. Moreover, the number of verb senses

Spontaneous

?

Frequently Intransitive

Unmarked Intransitive   Distant S and O centres
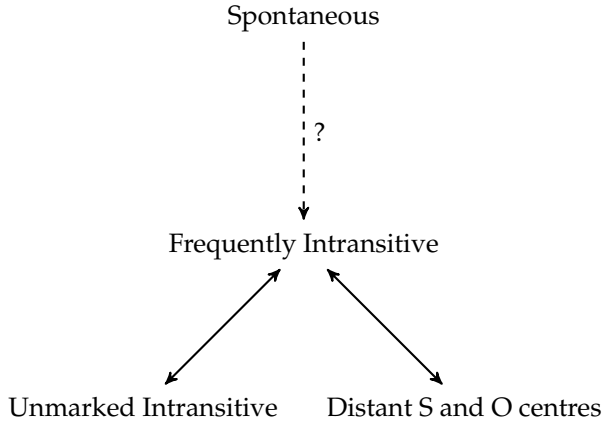
Figure 4
Synopsis of the correlations among features of causative-inchoative verbs. The measures are based on Kendall Tau test (τ) and Spearman's ranking test (ρ).

appeared to play a role with respect to the polymorphism of lexical sets: intuitively, the more the shades of meaning, the more the argument types that match the selectional preferences. Finally, a correlation was discovered between the cosine distance of lexical sets of a given verb in Italian and the cross-lingual behaviour of its translations, i.e. the tendency to appear more frequently as intransitive or as transitive. This finding has to be paired with the previously established correlation between the latter and the cross-lingual tendency to derive morphologically the intransitive form or the transitive one.

To amend the limitations mentioned in § 5 (noise from parsing and extraction, lexical sets not fully representative), further research should: resort to an enhanced database with a wider sample, try to reduce the parsing error with state-of-art parsers, and add sense disambiguation for polysemous word forms (Grave, Obozinski, and Bach 2013). Also, it should choose even more pre-trained vector models, in order to try and replicate these results. In particular, the new vector models could be optimized for similarity through semantic lexica (Faruqui et al. 2015) or based on syntactic dependencies (Séaghdha 2010). The experiments in this work may be extended to other languages, either individually or through a multi-lingual word embedding model (Faruqui and Dyer 2014). In fact, cross-lingually correlations are more clear-cut than those emerging single languages (Haspelmath et al. 2014).

Acknowledgements

References

Al-Rfou, Rami, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria, 8–9 August 2013.

Alexiadou, Artemis. 2010. On the morpho-syntax of (anti–) causative verbs. In Malka Rappaport Hovav, Edit Doron, and Ivy Sichel, editors, *Syntax, lexical semantics and event structure*. pages 177–203.

Artale, Alessandro, Bernardo Magnini, and Carlo Strapparava. 1997. WordNet for Italian and its use for lexical discrimination. In *AI*IA 97: Advances in Artificial Intelligence*, pages 346–356, Rome, Italy, 17–19 September 1997. Springer.

Atkins, Beryl T.S. and Beth Levin. 1995. Building on a corpus: A linguistic and lexicographical look at some near-synonyms. *International Journal of Lexicography*, 8(2):85–114.

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, volume 1, pages 86–90, Montreal, Quebec, Canada, 10–14 August 1998.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Bohnet, Bernd. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 89–97, Beijing, China, 23–27 August 2010.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Brent, Michael R. 1991. Automatic acquisition of subcategorization frames from untagged text. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 209–214, Berkeley, California, 18–21 June 1991.

Brockmann, Carsten and Mirella Lapata. 2003. Evaluating and combining approaches to selectional preference acquisition. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, volume 1, pages 27–34, Budapest, Hungary, 12–17 April 2003.

Cennamo, Michela. 1995. Transitivity and VS order in Italian reflexives. *Sprachtypologie und Universalienforschung*, 48(1-2):84–105.

Cennamo, Michela and Elisabetta Jezek. 2011. The anticausative alternation in italian. In Giovanna Massariello Merzagora and Serena Dal Maso, editors, *I luoghi della traduzione. Le interfacce*. Bulzoni Editore, pages 809–823.

Cha, Sung-Hyuk. 2007. Comprehensive survey on distance/similarity measures between probability density functions. *City*, 1(2):1.

Copley, Bridget and Phillip Wolff. 2014. Theories of causation should inform linguistic theory and vice versa. In Bridget Copley and Fabienne Martin, editors, *Causation in grammatical structures*. Oxford University Press, pages 11–57.

Dinu, Georgiana, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), Workshop Track*, San Diego, California, 7–9 May 2015.

Dixon, Robert M.W. 1994. *Ergativity*. Cambridge University Press.

Dubossarsky, Haim, Yulia Tsvetkov, Chris Dyer, and Eitan Grossman. 2015. A bottom up approach to category mapping and meaning change. In *Word Structure and Word Usage. Proceedings of the NetWordS Final Conference*, pages 66–70, Pisa, Italy, 30 March – 1 April 2015.

Dubossarsky, Haim, Daphna Weinshall, and Eitan Grossman. 2016. Verbs change more than nouns: a bottom-up computational approach to semantic change. *Lingue e linguaggio*, 15(1):7–28.

Faruqui, Manaal, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2015)*, pages 1606–1615, Denver, Colorado, 31 May – 5 June.

Faruqui, Manaal and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden, 26–30 April

2014.

Firth, John R. 1957. A synopsis of linguistic theory, 1930-1955. In *Philological Society of Great Britain*, editor, *Studies in linguistic analysis*. Basil Blackwell, pages 1–32.

Folli, Raffaella. 2002. *Constructing Telicity in English and Italian*. Ph.D. thesis, Oxford University.

Geeraerts, Dirk. 1997. *Diachronic prototype semantics: A contribution to historical lexicology*. Oxford University Press.

Gildea, Daniel and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

Grave, Edouard, Guillaume Obozinski, and Francis Bach. 2013. Hidden Markov tree models for semantic class induction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning (CoNLL)*, pages 94–103, Sofia, Bulgaria, 8–9 August 2013.

Hanks, Patrick. 1996. Contextual dependency and lexical sets. *International Journal of Corpus Linguistics*, 1(1):75–98.

Hanks, Patrick and Elisabetta Jezek. 2008. Shimmering lexical sets. In *Proceedings of the XIII EURALEX International Congress*, pages 391–402, Barcelona, Spain, 15–19 July 2008.

Hanks, Patrick and James Pustejovsky. 2005. A pattern dictionary for natural language processing. *Revue Française de linguistique appliquée*, 10(2):63–82.

Harris, Zelig. 1954. Distributional structure. *Word*, 10(23):146–162.

Haspelmath, Martin. 1993. More on the typology of inchoative/causative verb alternations. In Bernard Comrie and Maria Polinsky, editors, *Causatives and transitivity*, volume 23. pages 87–120.

Haspelmath, Martin, Andreea Calude, Michael Spagnol, Heiko Narrog, and Elif Bamyaci. 2014. Coding causal–noncausal verb alternations: A form–frequency correspondence explanation. *Journal of Linguistics*, 50(03):587–625.

Hilpert, Martin and Florent Perek. 2015. Meaning change in a petri dish: constructions, semantic vector spaces, and motion charts. *Linguistics Vanguard*, 1(1):339–350.

Jezek, Elisabetta. 2003. *Classi di verbi tra semantica e sintassi*. ETS.

Jezek, Elisabetta and Patrick Hanks. 2010. What lexical sets tell us about conceptual categories. *Lexis*, 4(7):22.

Jezek, Elisabetta, Bernardo Magnini, Anna Feltracco, Alessia Bianchini, and Octavian Popescu. 2014. T-PAS: A resource of corpus-derived types predicate-argument structures for linguistic analysis and semantic processing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, 26-31 May 2014.

Joanis, Eric, Suzanne Stevenson, and David James. 2008. A general feature space for automatic verb classification. *Natural Language Engineering*, 14(03):337–367.

Lakoff, George. 1987. *Women, fire, and dangerous things: What categories reveal about the mind*. Cambridge University Press.

Levin, Beth and Malka Rappaport Hovav. 1995. *Unaccusativity: At the syntax-lexical semantics interface*. MIT press.

Maaten, Laurens van der and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

MacQueen, James. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297, Berkeley, California, 21 June 1965 – 7 January 1966.

McCarthy, Diana. 2000. Using semantic preferences to identify verbal participation in role switching alternations. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 256–263, Seattle, Washington, 29 April – 4 May 2000.

McCarthy, Diana and John Carroll. 2003. Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, 29(4):639–654.

McKoon, Gail and Talke Macfarland. 2000. Externally and internally caused change of state verbs. *Language*, 76(4):833–858.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR) Workshop*, Scottsdale, Arizona, 2–4 May 2013.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013 (NIPS 2013)*. Curran Associates, Inc., Lake Tahoe, Nevada, 05–10 December 2013, pages 3111–3119.

Mirroshandel, Seyed Abolghasem and Alexis Nasr. 2016. Integrating selectional constraints and subcategorization frames in a dependency parser. *Computational Linguistics*, 42(1).

Montemagni, Simonetta, Nilda Ruimy, and Vito Pirrelli. 1995. Ringing things which nobody can ring. a corpus-based study of the causative-inchoative alternation in Italian. *Textus*, 8(2):1000–1020.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan T. McDonald, Slav Petrov, Sampo Pyysalo, and Natalia Silveira. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pages 1659–1666, Portoroz, Slovenia, 23–28 May 2016.

Pantel, Patrick, Rahul Bhagat, Bonaventura Coppola, Timothy Chklovski, and Eduard H. Hovy. 2007. ISP: Learning inferential selectional preferences. In *Proceedings of the 2007 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT)*, pages 564–571, Rochester, New York, 22–27 April 2007.

Pelleg, Dan and Andrew W. Moore. 2000. X-means: Extending K-means with efficient estimation of the number of clusters. In *Proceedings of the 17th International Conference on Machine Learning (ICML)*, volume 1, pages 727–734, Stanford, California, 29 June – 2 Jul 2000.

Ponti, Edoardo Maria, Elisabetta Jezek, and Bernardo Magnini. 2016. Grounding the lexical sets of causative-inchoative verbs with word embedding. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016)*, Naples, Italy, 5–6 December 2016.

Ponti, Edoardo Maria and Anna Korhonen. 2017. Event-related features in feedforward neural networks contribute to identifying causal relations in discourse. In *Proceedings of EACL, Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 25–30, Valencia, Spain, 3–7 April 2017.

Resnik, Philip. 1997. Selectional preference and sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, pages 52–57, Washington, D.C., 4 April 1997.

Rosch, Eleanor H. 1973. Natural categories. *Cognitive psychology*, 4(3):328–350.

Samardzic, Tanja and Paola Merlo. 2012. The meaning of lexical causatives in cross-linguistic variation. *Linguistic Issues in Language Technology*, 7(12):1–14.

Schulte Im Walde, Sabine. 2000. Clustering verbs semantically according to their alternation behaviour. In *Proceedings of the 18th conference on Computational linguistics*, volume 2, pages 747–753, Saarbrücken, Germany, 31 July – 4 August 2000.

Séaghdha, Diarmuid O. 2010. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 435–444, Uppsala, Sweden.

Séaghdha, Diarmuid O. and Anna Korhonen. 2014. Probabilistic distributional semantics with latent variable models. *Computational Linguistics*, 40(3):587–631.

Shutova, Ekaterina, Simone Teufel, and Anna Korhonen. 2013. Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353.

Van de Cruys, Tim. 2014. A neural network approach to selectional preference acquisition. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 26–35, Doha, Qatar, 25–29 October 2014.

Zapirain, Benat, Eneko Agirre, Lluis Marquez, and Mihai Surdeanu. 2013. Selectional preferences for semantic role classification. *Computational Linguistics*, 39(3):631–663.