Towards Semantic Language Classification: Inducing and Clustering Semantic Association Networks from Europarl

Steffen Eger[1], **Niko Schenk**[2] **and Alexander Mehler**[1]
[1]Text Technology Lab
[2]Applied Computational Linguistics Lab
Goethe University Frankfurt am Main
{steeger,nschenk,amehler}@em.uni-frankfurt.de

Abstract

We induce semantic association networks from translation relations in parallel corpora. The resulting semantic spaces are encoded in a single reference language, which ensures cross-language comparability. As our main contribution, we cluster the obtained (cross-lingually comparable) lexical semantic spaces. We find that, in our sample of languages, lexical semantic spaces largely coincide with genealogical relations. To our knowledge, this constitutes the first large-scale quantitative lexical semantic typology that is completely unsupervised, bottom-up, and data-driven. Our results may be important for the decision which multilingual resources to integrate in a semantic evaluation task.

1 Introduction

There has been a recent surge of interest in integrating multilingual resources in natural language processing (NLP). For example, Snyder et al. (2008) show that jointly considering morphological segmentations across languages improves performance compared to the monolingual baseline. Bhargava and Kondrak (2011) and Bhargava and Kondrak (2012) demonstrate that string transduction can benefit from supplemental information provided in other languages. Analogously, in lexical semantics, Navigli and Ponzetto (2012) explore semantic relations from Wikipedia in different languages to induce a huge integrated lexical semantic network.

In this paper, we also focus on multilingual resources in lexical semantics. But rather than *integrating* them, we investigate their *(dis-)similarities*. More precisely, we cluster (classify) languages based on their semantic relations between lexical units. The outcome of our classification may have direct consequences for approaches that integrate diverse multilingual resources. For example, from a linguistic point of view, it might be argued that integrating very heterogeneous/dissimilar semantic resources is *harmful*, e.g., in a monolingual semantic similarity task, because semantically unrelated languages might contribute semantic relations unavailable in the language for which semantic similarity is computed. Alternatively, from a statistical point of view, it might be argued that integrating heterogeneous/dissimilar resources is *beneficial* due to their higher degree of uncorrelatedness. In any case, either of these implications necessitates knowledge of a typology of lexical semantics.

In order to address this question, we provide a translation-based model of lexical semantic spaces. Our approach is to generate association networks in which the weight of a link between two words depends on their degree of partial synonymy. To measure synonymy, we rely on translation data that is input to a statistical alignment toolkit. We define the degree of synonymy of two words to be proportional to the number of common translations in a reference language, weighted by the probability of translation. By pivoting on the reference language, we represent semantic associations among words in different languages by means of the synonymy relations of their translations in the *same target language*. This approach ensures cross-language comparability of semantic spaces: Greek and Bulgarian are compared, for example, by means of the synonymy relations

*Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM 2015)*, pages 127–136,
Denver, Colorado, June 4–5, 2015.

that are retained when translating them into the same pivot language (e.g., English).

This approach does not only address proximities of pairs of words shared among languages (e.g., MEAT and BEEF, MOUTH and DOOR, CHILD and FRUIT – cf. Vanhove et al. (2008)). By averaging over word pairs, it also allows for calculating *semantic distances* between pairs of languages.

The *Sapir-Whorf Hypothesis* (SWH) (Whorf, 1956) already predicts that semantic relations are not universal. Though we are agnostic about the assumptions underlying the SWH, it nevertheless gives an evaluation criterion for our experiment: if the SWH is true, we expect a clustering of translation-based semantic spaces along the genealogical relationships of the languages involved. However, genealogy is certainly not the sole principle potentially underlying a typology of lexical semantics. For example, Cooper (2008) finds that French is semantically closer to Basque, a putatively non-Indoeuropean language, than to German. To the best of our knowledge, a large-scale quantitative typological analysis of lexical semantics is lacking thus far and we intend to make first steps towards this target.

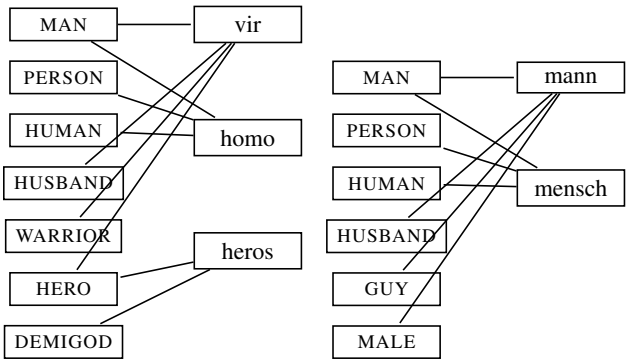
The paper is structured as follows. Section 2 outlines related work. Section 3 presents our formal model and Section 4 details our experiments on clustering semantic spaces across selected languages of the European Union. We conclude in Section 5.

2 Related work

A field related to our research is *semantic relatedness*, in which the task is to determine the degree of semantic similarity between pairs of words, such as *tiger* and *cat*, *sex* and *love*, etc. Classically, semantic word networks such as WordNet (Fellbaum, 1998) or EuroWordNet (Vossen, 1998) have been used to address this problem (Jiang and Conrath, 1997), and, more recently, taxonomies and knowledge bases such as Wikipedia (Strube and Ponzetto, 2006). Hassan and Mihalcea (2009) define the task of *cross-lingual semantic relatedness*, in which the goal is to determine the semantic similarity between words from different languages, and Navigli and Ponzetto (2012) have combined WordNet with Wikipedia to construct a multi-layer semantic network in which computation of cross-lingual semantic relatedness may be performed. Most recently, neural network-based distributed semantic representations focusing on cross-language similarities between words and larger textual units have become popular (Chandar A P et al. (2014), Hermann and Blunsom (2014), Mikolov et al. (2013)).

There have been (a) few different computational approaches to *semantic language classification*. Mehler et al. (2011) test whether languages are genealogically separable via topological properties of semantic (concept) graphs derived from Wikipedia. This approach is top-down in that it assumes that the genealogical tree is the desired output of the classification. Cooper (2008) computes semantic distances between languages based on the curvature of translation histograms in bilingual dictionaries. While this results in some interesting findings as indicated, the approach is not applied to language classification, but focuses on computing semantically similar languages for a given query language. Vanhove et al. (2008) construct so-called semantic proximity networks based on monolingual dictionaries, and envision to use them for semantic typologies. They do not apply their methodology to the multilingual setup, however, which a typology necessitates.

Orthographic, phonetic and *syntactic* similarity of languages have received considerably more attention than *semantic* similarity, as we focus on. Classical approaches in determining orthographic/phonetic relatedness of languages are based on lexico-statistical comparisons of items in standardized word lists (Campbell, 2003; Rama and Borin, 2015), such as the Swadesh lists (Swadesh, 1955). Rama and Borin (2015) study the impact of different string similarity measures on orthographic language classification. Ciobanu and Dinu (2014) measure orthographic similarity between Romanian and related languages. They also indicate applications of (knowledge of) similarity values between languages, such as serving as a guide for machine translation (Scannell, 2006). Koehn (2005) produces a genealogical clustering of the languages in Europarl based on ease of translation, as measured in BLEU scores, between any two languages (which, putatively, yields a syntactic similarity indication). This results in an imperfect reproduction of the ge-



(a) English-Latin (b) English-German

Figure 1: Excerpts of bilingual dictionaries as bipartite graphs with links between words if and only if one is a translation of the other. Data from www.latin-dictionary.net and dict.leo.org.

nealogical language tree for the languages involved.

3 Model

We start with motivating our approach by example of bilingual *dictionaries* before we formally generalize it in terms of probabilistic translation relations. Bilingual dictionaries, or the bipartite graphs that represent them (cf. Figure 1), induce lexical semantic association networks in any of the languages involved by *placing a link between two words of the same language if and only if they share a common translation in the other language* (cf. Figure 2).

Since translations provide partially synonymous expression in the target language, the latter links can be seen to denote semantic relatedness (in terms of synonymy) of the interlinked words. Further, the more distant two words in such a lexical semantic association network, the lower the degree of their partial synonymy: the longer the path from one word to another, the higher the loss of relatedness among them (cf. Eger and Sejane (2010)).

Note that association networks derived from bilingual dictionaries represent semantic similarities of words of the source language R subject to semantic relations of their translations in the target language L. The reason is that whether or not a link is established between two words α and β in R depends on associations of their translations present in L. To illustrate this, consider the association networks outlined in Figure 2, induced from the bilingual dictionaries outlined in Figure 1, which match between $R =$ English and $L =$ Latin and $L =$ German, respectively. When L is classical Latin, the semantic field centered around (the English word) MAN is partially different from the semantic field around MAN when L is German. For example, under $L =$ Latin, MAN is directly linked with HERO and WARRIOR (indirectly with DEMIGOD) – these semantic associations are not present when German is the language L.

By fixing R and varying L, we can create different lexical semantic association networks, each encoded in language R, and each representing the semantic relations of L.[1] Analyzing and contrasting such networks may then allow for clustering languages due to shared lexical semantic associations.

As mentioned above, we generalize the model outlined so far to the situation of probabilistic translation relationships derived from corpus data, rather than from bilingual dictionaries. Working on corpus data has both advantages and disadvantages compared to using human compiled and edited dictionaries. On the one hand,

- the translation relations induced from corpus data are *noisy* since their estimation is partially inaccurate due to limitations of alignment toolkits such as GIZA++ (Och and Ney, 2003) as employed by us. Implications of this inaccuracy are outlined below.
- By using unannotated corpora, we cannot straightforwardly distinguish between cases of polysemy and homonymy. The problem is that homonymy should (ideally) not contribute to generating lexical semantic association networks as considered here. However, homonymy is apparently a rather rare phenomenon, while polysemy, which we expect to underlie the structure of our networks, is abundant (cf. Löbner (2002)).

On the other hand,

- classical dictionaries can be very heterogeneous in their scope and denomination of translation links between words (see, e.g., Cooper (2008)), making the respective editors of the bilingual dictionaries distorting variables.

[1] Each network represents the semantic relations of *both* languages R and L, but since we keep R fixed and vary L, each association network inherits the same properties from R.

- Corpus data allows for inducing probabilities of translation relations of words, which indicate weighted links more accurately than ranked assignments provided by classical dictionaries.
- Corpus data allows for dealing with real language use by means of comparable excerpts of natural language data.

Network generation Assume that we are given different natural languages $L_1, \ldots, L_M, R$ and bilingual translation relations that map from language L_k to language R, for all $1 \leq k \leq M$. We call the language R *reference language*.[2] In our work, we assume that the translation relations are probabilistic. That is, we assume that there exist probabilistic 'operators' P_k that indicate the probabilities – denoted by $P_k[\alpha|z]$ – by which a word z of language L_k translates into a word α of language R. Our motivation is to induce M different lexical semantic networks that represent the lexical semantic spaces of the languages $L_1, \ldots, L_M$, each encoded in language R, which finally allows for comparing the semantic spaces of the M different source languages. To this end, we define the weighted graphs $G_k = (V_k, W_k)$, where the nodes V_k of G_k are given by the vocabulary R^{voc} of language R, i.e. $V_k = R^{\text{voc}}$. We define the weight of an edge $(\alpha, \beta) \in \left(R^{\text{voc}}\right)^2$ as

$$W_k(\alpha, \beta) = \sum_{z \in L_k^{\text{voc}}} P_k[\alpha|z] P_k[\beta|z] p[z], \quad (1)$$

where $p[z]$ denotes the (corpus) probability of word $z \in L_k^{\text{voc}}$. Since each G_k is spanned using the same subset of the vocabulary of the reference language R, we call it the *L_k(-based) network version of R*.

Eq. (1) can be motivated by postulating that W_k is a joint probability. In this case we can write

$$\begin{aligned} W_k(\alpha, \beta) &= \sum_{z \in L_k^{\text{voc}}} W_k(\alpha, \beta, z) = \sum_{z \in L_k^{\text{voc}}} W_k(\alpha, \beta | z) W_k(z) \\ &\approx \sum_{z \in L_k^{\text{voc}}} W_k(\alpha|z) W_k(\beta|z) W_k(z), \end{aligned} \quad (2)$$

where the first equality is marginalization ('summing out over the possible states of the world'), and the third step is an approximation which would be accurate if α and β were conditionally independent given z. By inserting the conditional probabilities $P_k[\alpha|z]$, $P_k[\beta|z]$ (whose existence we assumed above) and the corpus probability $p[z]$ into Eq. (2), we obtain Eq. (1). Note that in the special case of a bilingual dictionary of L_k and R, where $P_k[\alpha|z]$ can be defined as 1 or 0 depending on whether α is a translation of z or not,[3] $W_k(\alpha, \beta)$ is proportional to the *number of words z (in language L_k) whose translation is both α and β*; i.e., assuming that $p[z]$ is a constant in this setup, Eq. (1) simplifies to:

$$W_k(\alpha, \beta) \propto \sum_{z \in L_k^{\text{voc}} : z \text{ translates into } \alpha \text{ and } \beta} 1.$$

Clearly, the more common translations two words have in the target language, the closer their semantic similarity should be, all else being equal.[4] Eq. (1) generalizes this interpretation by non-uniformly 'prioritizing' the translations of z.

Network analysis In order to compare the network versions $G_1, \ldots, G_M$ of language R that are output by network generation, we first define the vector representation of node v^k in graph $G_k = (V_k, W_k)$ as the probability vector of ending up in any of the nodes of G_k when a random surfer starts from v^k and surfs on the graph G_k according to the normalized weight matrix $\mathbf{W}_k = [W_k(\alpha, \beta)]_{(\alpha,\beta) \in V_k \times V_k}$. Note that the higher $W_k(\alpha, \beta)$, the higher the likelihood that the surfer takes the transition from α to β. More precisely, we let the meaning $[\![v^k]\!]$ of node v^k in graph G_k be the vector $\mathbf{v}^k$ that results as the limit of the iterative process (see, e.g., Brin and Page (1998), Gaume and Mathieu (2008), Kok and Brockett (2010)),

$$\mathbf{v}_{N+1}^k = d \mathbf{v}_N^k \mathbf{A}^{(k)} + (1 - d) \mathbf{v}_0^k,$$

where each $\mathbf{v}_N^k$, for $N \geq 0$, is a $1 \times |R^{\text{voc}}|$ vector, $\mathbf{A}^{(k)}$ is obtained from $\mathbf{W}_k$ by normalizing all rows such that $\mathbf{A}^{(k)}$ is row-stochastic, and d is a damping factor that describes preference for the starting vector $\mathbf{v}_0^k$, which is a vector of zeros except for index

[2] Alternative names for the concept we have in mind might, e.g., be *pivot language*, *tertium comparationis* or *interlingua*.

[3] More correctly, one could define $P_k[\alpha|z] = \frac{1}{f_z}$, whenever α is a translation of z, and $P_k[\alpha|z] = 0$, otherwise, where f_z is the number of translations of word z. This would lead to an analogous interpretation as the given one.

[4] This reasoning ignores cases of homonymy, which weaken the semantic argument. See our discussion above.

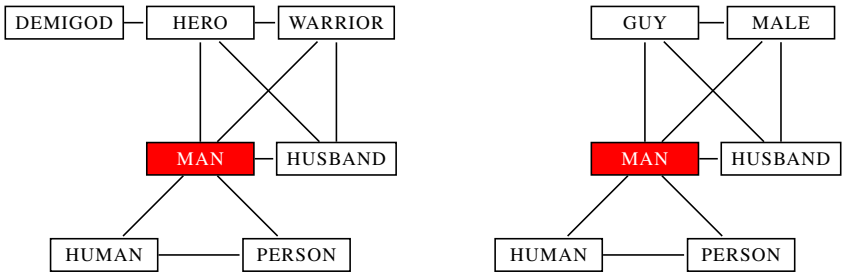


Figure 2: Lexical semantic association networks derived from bilingual dictionaries, given in Figure 1, by linking two English words if and only if they have a common translation in Latin (left) or German (right). The node for MAN is highlighted in both networks.

position of word v^k, where $\mathbf{v}_0^k$ has value 1.[5] Subsequently, we can contrast words v and w (or, rather, their meanings) in the same network version of reference language R, by considering, for instance, the cosine similarity or vector distance of their associated vectors. More generally, we can contrast the lexical semantic meanings $\mathbf{v}^k$ and $\mathbf{w}^j$ of any two language R words v and w, across two languages L_k and L_j, by, e.g., evaluating,

$$\mathbf{v}^k \cdot \mathbf{w}^j \quad \text{(scalar product, cosine similarity)}$$

or

$$||\mathbf{v}^k - \mathbf{w}^j|| \quad \text{(vector distance)}.$$

Finally, the lexical semantic distance or similarity between two *languages* L_k and L_j can be determined by simple averaging,

$$D(L_k, L_j) = \frac{1}{|R^{\text{voc}}|} \sum_{v \in R^{\text{voc}}} S(\mathbf{v}^k, \mathbf{v}^j), \qquad (3)$$

where S is a distance or similarity function.

Discussion We mentioned above that toolkits like GIZA++ cannot perfectly estimate translation relationships between words in different languages. Thus, we have to face situations of 'noisily' weighted links between words in the same network version of reference language R. Typically, a higher chance of mismatch occurs in the case of bigrams. To illustrate, consider the French phrase *êtres chers* ('beings loved'/'loved ones'). Here, GIZA++ typically assigns positive weight mass to $P_{\text{fr}}[\text{LOVE}|\text{être}]$ although, from a point of view of a classical dictionary, translating *être* into *love* is clearly problematic. Since it is likely that, e.g., $P_{\text{fr}}[\text{HUMAN}|\text{être}]$ and $P_{\text{fr}}[\text{BEING}|\text{être}]$ will also be positive, we can expect weighted links in the French network version of English between HUMAN and LOVE as well as between BEING and LOVE. Thus, besides 'true' semantic relations, our approach also captures, though unintentionally, co-occurrence relations.

[5] We always set d to 0.8 in our experiments.

4 Experiments

We evaluate our method by means of the Europarl corpus (Koehn, 2005). Europarl documents the proceedings of the European parliament in the 21 official languages of the European Union. This provides us with sentence-aligned multi-texts in which each tuple of sentences expresses the same underlying meaning.[6] Using GIZA++, this allows us to estimate the conditional translation probabilities $P[A|B]$ for any two words A, B from any two languages in the Europarl corpus. In our experiment, we focus on the approx. 400,000 sentences for which translations in all 21 languages are available. To process this data, we set all words of all sentences to lower-case. Ideally, we would have lemmatized all texts, but did not do so because of the unavailability of lemmatizers for some of the languages. Therefore, we decided to lemmatize only words in the reference language and kept full-forms for all source languages.[7] We choose

[6] In a tuple of sentences, one sentence is the source of which all the other sentences are translations.

[7] Lemmatization tools and models are taken from the TreeTagger (Schmid, 1994) home page `www.cis.uni-muenchen.de/~schmid/tools/TreeTagger`

English as the reference language.[8] In all languages, we omitted all words whose corpus frequency is less than 50 and excluded the 100 most frequent (mostly function) words.[9] In the reference language, we also ignored all words whose characters do not belong to the standard English character set.

Figure 3 shows subgraphs centered around the seed word WOMAN in five network versions of English. All subgraphs are constructed using the Europarl data. Apparently, the network versions of English diverge from each other. For instance, the semantic association between WOMAN and WIFE appears to be strongest in the French and in the Spanish version of English, while in the Finnish version there does not even exist a link between these nodes. In contrast, the weight of the link between WOMAN and LESBIAN is highest in the Czech version of English, while that between WOMAN and GIRL is strongest in the Finnish version. All in all, the wiring and the thickness of links clearly differ across language networks, indicating that the languages differ in terms of semantic relations of their translations.

Table 1 shows network statistics of the graphs G_k. All network versions of English consist of exactly 5,021 English (lemmatized) words. The networks show a high cluster value, indicating that neighbors of a word are probably interlinked (i.e., semantically related) (cf. Watts and Strogatz (1998)). Average path lengths and diameters are low, that is, distances between words are short, as is typically observed for semantic networks (cf. Steyvers and Tenenbaum (2005)). The density of the networks (measured by the ratio of existing links and the upper bound of theoretically possible links) varies substantially for the language networks. For instance, in the Hungarian network version of English, only 2.56% of the possible links are realized, while in the Dutch version, 8.45% are present. This observation may hint at the 'degree of analyticity' of a language: the more word forms per lemma there are in a language, the less likely they are linked by means of Eq. (1).

[8]Due to the limited availability of lemmatizers, not all languages could have served as a reference language. Although we posit that the choice of reference language has no (or minimal) impact upon the resulting language classification as outlined below, this would need to be experimentally verified in follow-up work.

[9]The threshold of 50 serves to reduce computational effort.

	# nodes	CV	GD	D	density (%)
cs	5,021	0.39	1.96	4	4.51
da	5,021	0.43	1.95	5	5.35
nl	5,021	0.50	1.85	4	8.45 (9.22)
et	5,021	0.37	1.98	5	3.81 (4.57)
fi	5,021	0.35	1.99	4	3.28 (6.63)
fr	5,021	0.44	1.91	4	6.37 (8.23)
de	5,021	0.43	1.96	5	5.03 (5.81)
el	5,021	0.36	2.00	5	3.79
hu	5,021	0.33	2.07	5	2.56
it	5,021	0.45	1.87	4	7.41 (9.53)
lv	5,021	0.41	1.94	4	5.29
lt	5,021	0.41	1.94	4	5.08
pl	5,021	0.39	1.94	4	4.84 (6.56)
pt	5,021	0.40	1.97	4	4.74
ro	5,021	0.39	2.00	5	4.22
sk	5,021	0.36	1.99	5	3.73 (5.23)
sl	5,021	0.38	1.97	4	4.13
es	5,021	0.40	1.98	5	4.67 (5.80)
sv	5,021	0.43	1.94	5	5.69

Table 1: Number of nodes, cluster value (CV), geodesic distance (GD), diameter (D) and density of different network versions of English. Links are binarized depending on whether their weights are positive or not. In brackets: values of lemmatized versions of L_k.

Note that since the density of a network may have substantial impact on random surfer processes as applied by us, and since analyticity is a morphological rather than a semantic phenomenon, it may be possible that the classification results reported below are in fact due to syntagmatic relations – in contrast to our hypothesis about their semantic, paradigmatic nature. We address this issue below.

Semantic similarity Before proceeding to our main task, the clustering of semantic spaces, we measure how strongly our semantic association networks capture semantics. To this end, we compute the correlation coefficient between the semantic similarity scores of the word pairs in the WordSimilarity-353 (Finkelstein et al., 2001) English word relatedness dataset and the similarity scores, for the same word pairs, obtained by our method. The WordSimilarity-353 dataset consists of 353 word pairs annotated by the average of 13 human experts, each on a scale from 0 (unrelated) to 10 (very closely related or identical). We evaluated only on those word pairs for which each word in the pair is contained in our set of 5,021 English words, which amounted to 172 word pairs. To be more

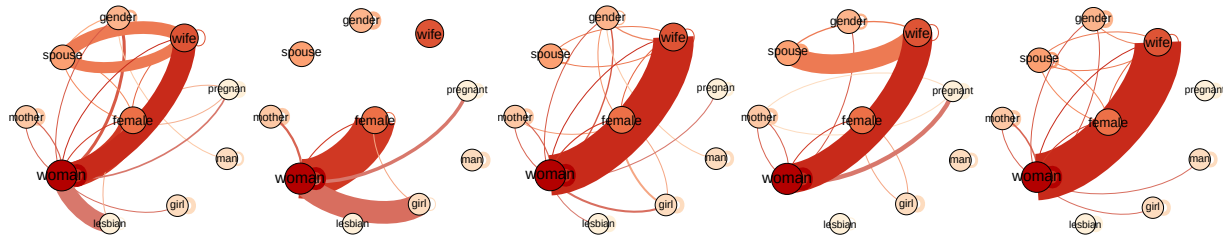


Figure 3: From left to right: Czech, Finnish, French, German, and Spanish networks. Thickness of edges indicates weights of links. Links with weights below a fixed threshold are ignored for better graphical presentation.

precise on the computation of semantic relatedness, for each word pair (u, v) in the WordSimilarity-353 dataset, we computed the semantic similarity of the word pair in the language L_k version of English by considering the cosine similarity of $\mathbf{u}^k$ and $\mathbf{v}^k$, that is, by means of the semantic meanings of u and v generated by the random surfer process on network G_k. Doing so for each language L_k gives 20 different correlation coefficients, one for each network version of English, shown in Table 2.

it	0.34678	⋮	⋮
pt	0.32249	sl	0.25720
es	0.31990	bg	0.25372
ro	0.31204	hu	0.24910
nl	0.30885	et	0.24212
da	0.30715	lt	0.24207

Table 2: Sample Pearson correlation coefficients between human gold standard and our approach for different network versions of English.

We first note that the correlation coefficients differ between network versions of English, where the Italian version exhibits the highest correlation with the (English) human reference, and the Lithuanian version the lowest. Note that Hassan and Mihalcea (2009) obtain a correlation coefficient of 0.55 on the whole WordSimilarity-353 dataset, which is considerably higher than our best score of 0.34. However, first note that our networks, which consist of 5,021 lexical units, are quite small compared to the data sizes that other studies rely on, which makes a comparison highly unfair. Secondly, one has to see that we compute the semantic relatedness of English words *from the semantic point of view of two languages:* the reference language and the respective source language (e.g., the Italian version of English), which, by our very postulate, differs from the semantics of the reference language. According to Table 2, the semantics of English is apparently better represented by the semantics of Italian, Portuguese, Spanish, Romanian, and Dutch, than, e.g., by the one of Bulgarian, Hungarian, Estonian, and Lithuanian – at least subject to the translations provided by the Europarl corpus.[10]

Clustering of semantic spaces Finally, we cluster semantic spaces by comparing the network versions of the English reference language. To determine the semantic distance between two languages L_k and L_j, we plug in each pair of languages in Eq. (3) – with $S(\mathbf{v}^k, \mathbf{v}^j)$ as vector distance – thus obtaining a symmetric 20×20 distance matrix. Figures 4 and 5 show the results when feeding this distance matrix as input to k-means clustering (a centroid based clustering approach) and to hierarchical clustering using default parameters. As can be seen, both clustering methods arrange the languages on the basis of their semantic spaces along genealogical relationships. For instance, both clustering algorithms group Danish, Swedish, Dutch and German (Germanic), Portuguese, Spanish, French, Italian, Romanian (Romance), Bulgarian, Czech, Polish, Slovak, Slovene (Slavic), Finnish, Hungarian, Estonian (Finno-Ugric), and Latvian, Lithuanian (Baltic). Greek, which is genealogically isolated in our selection of languages, is in our classification associated with the Romance languages, but constitutes an outlier in this group. All in all, the clustering appears highly non-random and almost a

[10] Table 2 also suggests that the Romance languages are semantically closer to English in our data than, e.g., the Germanic, which may be considered a deviation from, e.g., genealogical language similarity.

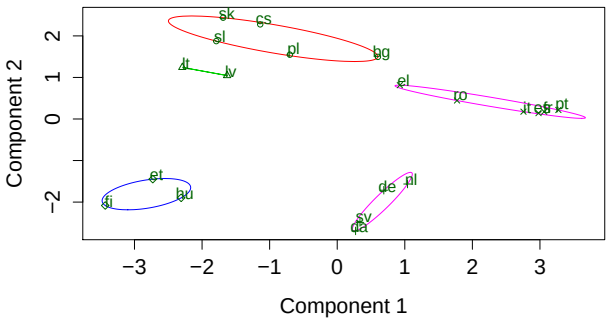


Figure 4: k-means cluster analysis of the 20 Europarl languages. Optimal number of clusters $k = 5$ determined by sum of squared error analysis.

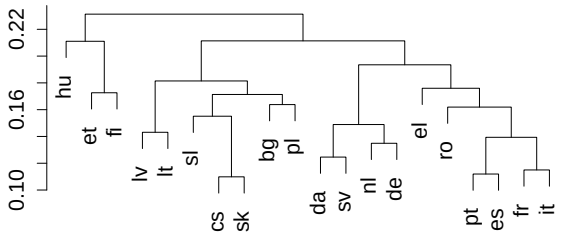


Figure 5: Dendrogram of hierarchical clustering of the 20 non-lemmatized Europarl languages.

perfect match of what is genealogically expected.

To address the question of whether morphological principles are the driving force behind the clustering of the semantic spaces generated here, we lemmatized the reference language English and all source languages L_k for which lemmatizers were freely available in order to conduct the same classification procedure. This included 10 languages: Bulgarian, Dutch, Estonian, Finnish, French, German, Italian, Polish, Slovak, and Spanish. This procedure leads to an assimilation of density values in the graphs G_k as shown in Table 1: for the 10 languages, the relative standard deviation in network density decreases by about 23%. However, the optimal groupings of the languages do not change in that k-means clustering determines the five groups Spanish, French, Italian; Bulgarian, Slovak, Polish; German, Dutch; Finnish; Estonian, *irrespective of whether the named ten languages are lemmatized or not.*[11]

Integrated networks Lastly, we address the derivative question raised in the introduction, viz., whether the integration of heterogeneous/dissimilar multilingual resources may be harmful or beneficial. To this end, we consider *integrated networks* $G^{(S)}$ in which the weight of a link $(\alpha, \beta) \in E^{(S)}$ is given as the *average* (arithmetic mean) link weight of all link weights in the networks for a selection of languages S. Using our optimal number of $k = 5$ clusters (and the clusters themselves) derived above, we thus let S range over the union of all the languages in the $2^k - 1$ possible subsets of clusters.[12] For each so resulting network $G^{(S)}$, we determine semantic similarity between any pair of words exactly as above and then compute correlation with the WordSimilarity-353 dataset. Results are given in Table 3. The numbers appear to support the hypothesis that, in the given monolingual semantic similarity task for English, integrating semantically similar languages (and, putatively, languages whose semantic similarity to English itself is closer) leads to better results than integrating heterogeneous languages. For example, the average network consisting of the Romance languages has a roughly 2% higher correlation than the network consisting of all languages. Interestingly, however, the very best combination result is achieved when we integrate the Romance, Germanic and the three non-Indoeuropean languages Finnish, Hungarian and Estonian.

R+G+F	0.34402	⋮	⋮
R+G	0.34376	S+B	0.27496
R+F	0.33743	S	0.27462
R	0.33719	B+F	0.27424
⋮	⋮	F	0.26074
R+G+F+B+S	0.31670	B	0.25904

Table 3: Sample Pearson correlation coefficients between human gold standard and our approach for different integrated network versions. Language cluster abbreviations: **R**omance (it, fr, pt, es, ro, el), **G**ermanic (sv, nl, de, da), **S**lavic (bg, cz, pl, sk, sl), **B**altic (lv, lt), **F**inno-Ugric (fi, hu, et).

[11] The clustering based on 10 languages slightly differs in that Finnish and Estonian are assigned to distinct clusters.

[12] Ideally, we would have let S range over all possible $2^n - 1$ nonempty subsets of n languages, but this would have required $2^{20} - 1 > 1$ million comparisons.

5 Conclusion

We have encoded lexical semantic spaces of different languages by means of the same pivot language in order to make the languages comparable. To this end, we introduced association networks in which links between words in the reference language depend on translations from the respective source language, weighted by probability of translation. Our methodology is closely related to analogous approaches in the paraphrasing community which interlink paraphrases by means of their translations in other languages (e.g., Bannard and Callison-Burch (2005), Kok and Brockett (2010)), but our application scenario is different and we also describe a principled manner to generate *weighted links* between lexical units from multilingual data. Using random walks to represent similarities among words in the association networks, we finally derived similarity values for pairs of languages. This allowed us to perform several cluster analyses to group the 20 source languages. Interestingly, in our data sample, semantic language classification appears to be almost perfectly correlated with genealogical relationships between languages. To the best of our knowledge, our translation-based lexical semantic classification is the first large-scale quantitative approach to establishing a lexical semantic typology that is completely unsupervised, 'bottom-up', and data-driven.[13]

In future work, we intend to delineate specific lexical semantic fields in which particular languages differ, which can easily be accomplished within our approach. Also, it must be investigated whether our association networks can capture semantic similarity in a competitive manner once they are scaled up appropriately. Finally, applying our methodology to a much larger set of languages is highly desirable.

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with Bilingual Parallel Corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 597–604, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aditya Bhargava and Grzegorz Kondrak. 2011. How Do You Pronounce Your Name?: Improving G2P with Transliterations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 399–408, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aditya Bhargava and Grzegorz Kondrak. 2012. Leveraging Supplemental Representations for Sequential Transduction. In *HLT-NAACL*, pages 396–406. The Association for Computational Linguistics.

Sergey Brin and Lawrence Page. 1998. The Anatomy of a Large-scale Hypertextual Web Search Engine. *Comput. Netw. ISDN Syst.*, 30(1-7):107–117, April.

Lyle Campbell. 2003. How to show Languages are related: Methods for Distant Genetic Relationship. In *The Handbook of Historical Linguistics*. Blackwell.

Sarath Chandar A P, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An Autoencoder Approach to Learning Bilingual Word Representations. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1853–1861. Curran Associates, Inc.

Alina Maria Ciobanu and Liviu P. Dinu. 2014. An Etymological Approach to Cross-Language Orthographic Similarity. Application on Romanian. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1047–1058.

Martin C. Cooper. 2008. Measuring the Semantic Distance between Languages from a Statistical Analysis of Bilingual Dictionaries. *Journal of Quantitative Linguistics*, 15(1):1–33.

Steffen Eger and Ineta Sejane. 2010. Computing Semantic Similarity from Bilingual Dictionaries. In *Proceedings of the 10th International Conference on the Statistical Analysis of Textual Data (JADT-2010)*, pages 1217–1225, Rome, Italy, June. JADT-2010.

Steffen Eger. 2012. Lexical Semantic Typologies from Bilingual Corpora — A Framework. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 90–94. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

[13] But see also the first author's preliminary investigations on semantic language classification in Sejane and Eger (2013), based on freely available (low-quality) bilingual dictionaries, and Eger (2012).

Lev Finkelstein, Gabrilovich Evgenly, Matias Yossi, Rivlin Ehud, Solan Zach, Wolfman Gadi, and Ruppin Eytan. 2001. Placing Search in Context: The Concept Revisited. In *Proceedings of the Tenth International World Wide Web Conference*.

Bruno Gaume and Fabien Mathieu. 2008. PageRank Induced Topology for Real-World Networks. *Complex Systems*, page (on line).

Samer Hassan and Rada Mihalcea. 2009. Cross-lingual Semantic Relatedness Using Encyclopedic Knowledge. In *EMNLP*, pages 1192–1201. ACL.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual Models for Compositional Distributed Semantics. *CoRR*, abs/1404.4641.

Jay J. Jiang and David .W. Conrath. 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proc. of the Int'l. Conf. on Research in Computational Linguistics*, pages 19–33.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: The Tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Stanley Kok and Chris Brockett. 2010. Hitting the Right Paraphrases in Good Time. In *HLT-NAACL*, pages 145–153. The Association for Computational Linguistics.

Sebastian Löbner. 2002. *Understanding Semantics*. Oxford University Press, New York.

Alexander Mehler, Olga Pustylnikov, and Nils Diewald. 2011. Geography of social ontologies: Testing a variant of the Sapir-Whorf Hypothesis in the context of Wikipedia. *Computer Speech & Language*, 25(3):716–740.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-coverage Multilingual Semantic Network. *Artificial Intelligence*, 193(0):217 – 250.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Comput. Linguist.*, 29(1):19–51, March.

Taraka Rama and Lars Borin. 2015. Comparative evaluation of string similarity measures for automatic language classification. In *Sequences in Language and Text*. De Gruyter Mouton.

Kevin Scannell. 2006. Machine translation for closely related languages. In *Proceedings of the Workshop on Strategies for Developing Machine Translation for Minority Languages*, pages 103–107.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Ineta Sejane and Steffen Eger. 2013. Semantic typologies by means of network analysis of bilingual dictionaries. In Lars Borin and Anju Saxena, editors, *Approaches to Measuring Linguistic Differences*, pages 447–474. De Gruyter.

Benjamin Snyder, Tahira Naseem, Jacob Eisenstein, and Regina Barzilay. 2008. Unsupervised Multilingual Learning for POS Tagging. In *EMNLP*, pages 1041–1050. ACL.

Mark Steyvers and Josh Tenenbaum. 2005. The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive Science*, 29(1):41–78.

Michael Strube and Simone P. Ponzetto. 2006. WikiRelate! Computing Semantic Relatedness using Wikipedia. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1419. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Morris Swadesh. 1955. Towards Greater Accuracy in Lexicostatistic Dating. *International Journal of American Linguistics*, 21:121–137.

Martine Vanhove, Bruno Gaume, and Karine Duvignau. 2008. Semantic Associations and Confluences in Paradigmatic Networks. In *From Polysemy to Semantic Change: Towards a Typology of Lexical Semantic Associations*, pages 233–264. John Benjamins.

Piek Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Norwell, MA, USA.

Duncan. J. Watts and Steven H. Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):409–10.

Benjamin Whorf. 1956. *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. MIT Press, Cambridge.