Measuring the compositionality of NV expressions in Basque by means of distributional similarity techniques

Antton Gurrutxaga[1], **Iñaki Alegria**[2]

[1]Elhuyar Foundation, Zelai Haundi kalea 3, Osinalde Industrialdea, 20170 Usurbil, Basque Country
[2]IXA group, UPV/EHU, Informatika Fakultatea 649 posta-kutxa 20080 Donostia, Basque Country
[1]a.gurrutxaga@elhuyar.com, [2]i.alegria@ehu.es

Abstract

We present several experiments aiming at measuring the semantic compositionality of NV expressions in Basque. Our approach is based on the hypothesis that compositionality can be related to distributional similarity. The contexts of each NV expression are compared with the contexts of its corresponding components, by means of different techniques, as similarity measures usually used with the Vector Space Model (VSM), Latent Semantic Analysis (LSA) and some measures implemented in the Lemur Toolkit, as Indri index, tf-idf, Okapi index and Kullback-Leibler divergence. Using our previous work with cooccurrence techniques as a baseline, the results point to improvements using the Indri index or Kullback-Leibler divergence, and a slight further improvement when used in combination with cooccurrence measures such as t-score, via rank-aggregation. This work is part of a project for MWE extraction and characterization using different techniques aiming at measuring the properties related to idiomaticity, as institutionalization, non-compositionality and lexico-syntactic fixedness.

Keywords: MWEs, idioms, collocations, compositionality, distributional similarity

1. Introduction

Idiomaticity is considered the key feature to define the concept of Multiword Expressions or Phraseological Units, and is usually described as a non-discrete magnitude, whose "value", according to recent investigations (Granger and Paquot, 2008; Baldwin and Kim, 2010; Fazly and Stevenson, 2007), has turned out to depend on a complex combination of features such as institutionalization, non-compositionality and lexico-syntactic fixedness.

Semantic non-compositionality is a prominent characteristic of many MWEs. The idea underlying this phenomenon is the *Principle of Compositionality*, which states that "the meaning of a whole is a function of the meaning of the parts and of the way they are syntactically combined." (Partee, 1995). According to that, an MWE is non-compositional when its "meaning cannot be inferred from the meaning of its parts" (Cruse, 1986).

The compositionality, and hence the idiomaticity, of MWEs appears rather as a continuum than as a scale of discrete values (Sinclair, 1996). Thus, the classification of MWEs into discrete categories is a difficult task. A very schematic classification that has achieved a relative agreement among experts distinguishes two main types of phraseological units at phrase-level: idioms and collocations.

Relating to compositionality, idioms are defined as non-compositional units, and divided by some scholars (Cowie, 1998; Melcuk, 1998) into opaque and figurative idioms, being the meaning of figurative units more "deductible" or easer to decode. Collocations are often considered as semi-compositional units, in which one component (the base, in our case the noun) preserves its literal meaning, and the other (the base, in this case the verb) is desemantized (light or support verb constructions) or has a meaning specific to the combination with the noun (nevertheless, some collocations are considered compositional, and their idiomaticity is a consequence of other phenomena, like lexico-syntactic fixedness and institutionalization). Free combinations are fully compositional. Some examples for Basque:

- Opaque idiom: 'adarra jo' $\neq$ 'adarra' $+$ 'jo' (*to pull someone's leg*, lit. 'to play the horn')
- Figurative idiom: 'burua hautsi' $\approx$ 'burua' $+$ 'hautsi' (*to racks one's brain(s)*, lit. 'to break one's head')
- Collocation: 'atentzioa eman' $=$ 'atentzioa' $+$ 'eman'(atentzio)[1] (*to catch/attract someone's attention*, lit. 'to give attention')
- Fully compositional: 'liburua irakurri' $=$ 'liburua' $+$ 'irakurri' (*to read a book*)

Even non-compositionality has been considered one of the central features of idiomaticity (Manning and Schütze, 1999), the standard techniques to extract MWEs automatically from text have been based until recently on cooccurrence data, a phenomenon mostly related to institutionalization, or "statistical idiosyncrasy". However, in the last decade, a growing effort has been devoted to the automatic measurement of compositionality from text data. The central concept to characterize compositionality is the hypothesis of distributional similarity (Lin, 1999). As proposed in Baldwin and Kim (2010), "the underlying hypothesis is that semantically idiomatic MWEs will occur in markedly different lexical contexts to their component words."

In a previous paper (Gurrutxaga and Alegria, 2011), we faced the task of extracting NV combinations from corpora based on association measures (AM). The evaluation was

[1] The notation 'eman'(atentzio) is used to convey the fact that the sense adopted by the verb *eman* is specific to its cooccurrence with *atentzio*.

designed on the distinction between MWEs and free combinations. In the present work, we are interested in the differentiation between idioms, collocations and free combinations.

2. Related work

One of the first methods was developed by Berry-Rogghe (1974), who proposed a measure named R-value to compute the compositionality of verb-particle constructions (VPCs), by dividing the overlap between the sets of collocates associated with the particle by the total number of collocates of the VPC. Wulff (2010) proposes two extensions to the R-value in her investigation on VNP-constructions. Basically, Wulff experiments with two methods for combining and weighting individual R-values of each component (noun and verb). Besides, those extensions of R-value are calculated taking into account different percentages of the most significant collocates and selecting them according to the Fisher-Yates exact test (instead of the original version by Berry-Rogghe, who used z-score to that end).

LSA (Latent Semantic Analysis) is used in several studies. Schone and Jurafsky (2001) compute semantic vectors for every proposed word n-gram and subcomponents. They report modest gains in performance. Baldwin et al. (2003) test the model over English noun-noun compounds and verb-particles and evaluate its correlation with similarities and hyponymy values in WordNet. Katz and Giesbrecht (2006) present experiments for German that show that low cosine similarity using LSA correlate with non-compositionality. They use Infomap.

The Vector Space Model is applied, among others, by Garrao et al. (2006) and Fazly and Stevenson (2007), who use the cosine as a similarity measure between vectors. In the first study, VSM is applied on MWEs in Portuguese and the context is the whole paragraph. The second one deals with light verb constructions (LVCs) in English, and uses as context a window of ±15 nouns.

More recently, Korkontzelos and Manandhar (2009) use graph-based sense induction in order to decide compositionality. The shared task Distributional Semantics and Compositionality (DiSCo) at ACL HLT 2011 shows a variety of techniques for this task, mainly association measures and VSM. Resources including MWEs in English and German are provided, a summary of which is given by Biemann and Giesbrecht (2011).

An open discussion is how to evaluate the results. Lin (1999), Baldwin et al. (2003) and Schone and Jurafsky (2001) use as their gold standard either idiom dictionaries or WordNet. Katz and Giesbrecht (2006) and the DiSCo'2011 shared task use a careful manual annotation of a database as a gold standard.

The results are calculated in several ways: precision, recall and accuracy are used in some studies, as well as other proposals e.g., in the DiSCo'2011 shared task the score is calculated as the distance between the system responses and the gold standard.

3. Experimental setup

We are interested in the extraction and characterization of NV expressions in Basque. The contexts of each NV expression are compared with the contexts of its corresponding components, by means of different techniques, as basic VSM, LSA and IR similarity indexes as Indri, tf-idf, Okapi and KL-divergence. We have carried out some experiments in order to have a comparative basis between different distributional similarity approaches, and to compare them with the previous results obtained using AMs to process cooccurrence data.

The corpus used, its pre-processing and the evaluation set are the same that those used in Gurrutxaga and Alegria (2011). We use a journalistic corpus from two sources: (1) Issues published in 2001-2002 by the newspaper *Euskaldunon Egunkaria* (28 Mw); and (2) Issues published in 2006-2010 by the newspaper *Berria* (47 Mw). So, the overall size of the corpus is 75 Mw. The corpus is annotated with lemma, POS, case and number information using EUSTAGGER developed by the IXA group of the University of the Basque Country (Aduriz et al., 1996).

3.1. Context generation

We extract the context words of each bigram from the sentences with contiguous cooccurrences of the components (window span = ± 1). For noun-verb cooccurrence of lemmas to be considered as a bigram, the noun must occur in the grammatical case in which it has been defined after bigram normalization.[2] This is necessary if we intend to be able to differentiate the compositionality of combinations like *kontuan hartu* ('take into account') $\neq$ *kontu hartu* ('to ask for an explanation'). The lemma of the first member of the Basque expressions is always *kontu*, whose POS is "noun", being *kontu* the indefinite form of *kontu* in the absolutive case, and *kontuan* the singular of *kontu* in the inesive case ("in").

Separately, the contexts of the corresponding noun and verb are extracted from single occurrences. For example, in the case of the bigram *erabakia hartu* ('to make/take a decision'), the contexts of *erabaki* ('decision') come from sentences where *hartu* ('to take') does not occur, or occur at a distance greater than 1 (non-contiguous cooccurrence); likewise for the contexts of the verb (*hartu*).

Only content-bearing lemmas are included in the contexts (nouns, verbs and adjectives).

3.2. Context processing and methods

We process the contexts in two different ways, depending on the techniques and tools used to measure distributional similarity.

Firstly, a VSM model is constructed and the contexts are represented as vectors. As similarity measures between the vector of a given bigram and those of its members, we use Berry-Roghe's R-value (R_{BR}) and its two extensions proposed by Wulff (R_{W1} and R_{W2}), Jaccard index and cosine; as for cosine, different AMs have been tested for vector weights [f, t-score, log-likelihood ratio (LL), pointwise mutual information (PMI), and Fisher's exact test]; AMs are calculated using the Ngram Statistics Package by Ted

[2] For more detailed information on the normalization of different Basque bigram forms belonging to the same noun_lemma+noun_case+verb_lemma key, see Gurrutxaga and Alegria (2011).

Pedersen (http://www.d.umn.edu/ tpederse/nsp.html). Besides, we use Lemur cosine implementation (which uses idf values for weights). For the versions of R-values, and our implementations of Jaccard index and cosine, we experimented with different percentages of the vector of collocates (100%, 75% and 50%), using the aforementioned measures to rank the collocates.

Secondly, the same contexts have been represented as documents, and compared using the Lemur Toolkit, by means of different indexes (Allan et al., 2003). The central idea is to use the contexts of the bigrams as queries against a document collection that includes the context-documents of all the members of the bigrams. This idea can be implemented in three different ways:

- Lemur_1: Similarly as with vectors, the contexts of a bigram are included in a single query document, and the same for the contexts of its members
- Lemur_2: Each context sentence is included in a different document. Thus, for a given bigram we created as much query documents as occurrences of the bigram. Using the same criteria, each context sentence of a member of a bigram is a different document in the index
- Lemur_3: The context sentences of bigrams are treated as individual documents, but the contexts of each one of its members are represented in two separate documents

Due to processing reasons, the number of context sentences used in Lemur to generate documents is limited to a maximum of 2,000, and randomly selected from the whole set of contexts.[3] We use default settings for smoothing parameters. In each query, the number of similar documents retrieved is 200.

A slight different approach has been adopted in LSA. The contiguous cooccurrences of the members of each bigram have been represented as single tokens in the corpus to be processed by Infomap. Only content-bearing words are included in the corpus. Using Infomap, a matrix of 30.000 rows and 2.000 content-bearing word is created, and then SVD is applied. Infomap uses the cosine measure to calculate the similarities between items in the rows of the matrix. As in Lemur, we retrieved the 200 most similar words of each bigram for evaluation. Using this information, we rank the bigrams in two ways: a) according to the average value of the cosines between the bigram and each of its members; b) according to the average similarities between the lists of 200 "neighbors" corresponding to the bigram and each of its members (we use cosine to measure those similarities). Finally, instead of retrieving similar words, Infomap brings the possibility to retrieve the similar documents of a given query, which can also be a given document. Thus, we can directly compare the context documents of a bigram and those of their components, and use Infomap in a similar way as in the Lemur_1 type experiment.

3.3. Evaluation

As an evaluation reference, we use a subset of 600 combinations selected randomly from a larger evaluation set (4,334) extracted from the corpus as defined in Gurrutxaga and Alegria (2011). This set of 4,334 bigrams is the result of merging the 2,000-best candidates of each AM ranking from the w = ± 1 and $f > 30$ extraction set.

The subset has been manually classified by three lexicographers into three categories: idioms, collocations and free combinations. Annotators were provided with an evaluation manual, with explanatory information about the evaluation task and the guidelines that had to be followed to differentiate between idioms, collocations and free combinations, based on the criteria mentioned in section 1. Illustrative examples are included.[4]

The agreement among evaluators was calculated using the Fleiss's κ statistics, and obtained a value of 0.54. Although this level of agreement is relatively low when comparing with (Krenn et al., 2004; Fazly and Stevenson, 2007), it is comparable to the one reported by Pecina (2010), who attributed his "relatively low" value to the fact that "the notion of collocation is very subjective, domain-specific, and also somewhat vague". Cases when agreement is two or higher have been automatically adopted, and the remaining cases have been classified after discussion. 10 combinations that do not belong to the NV category were removed. Finally, the evaluation set includes 590 items, out of which 46 are idioms (either opaque or figurative), 153 collocations and 391 free combinations.

In order to compare the results of the different techniques, we base our evaluation on the rankings provided by each measure. If we had an ideal measure, the set of bigram categories ('id', 'col' and 'free') would be an ordered set, with 'id' values on the top of the rank, 'col' in the middle part, and 'free' in the bottom. Thus, the idea is to compute the distance between a rank derived from the ideally ordered set, which contains a high amount of ties, and the rank derived from the set of categories yielded by each measure. To this end, we use Kendall's τ_B as a rank-correlation measure, using the Perl module Statistics-RankCorrelation-0.1203 by Gene Boggs (http://search.cpan.org/~gene/Statistics-RankCorrelation-0.1203/). Statistical significance of the Kendall's τ_B correlation coefficient is tested by the Z-test, computed according to Bolboacă and Jäntschi (2006).

In addition to that, average precision values (AP) have been calculated for each ranking.

In the case of association measures, similarity measures applied to VSM, and Infomap, the bigrams are ranked by means of the values of the corresponding measure (using the average value of the similarities between bigram-noun and bigram-verb). In the case of experiments with Lemur, the information used to rank the bigrams are the positions of the documents corresponding to each member of the bigram

[3] In order to make the results of Lemur and VSM experiments comparable, the same criteria has been used to generate VSM vectors.

[4] In addition, we made a classification that differentiated between opaque and figurative idioms, but was discarded due to the low proportion of opaque idioms (only two items). Thus, opaque and figurative idioms have been joined together in a single category.

	Measure	τ_B	AP
	random rank	(0.06593)	0.33230
AM	f	0.15104	0.43524
	t-score	0.14794	0.44269
	LL	0.11637	0.42787
	PMI	(-0.08274)	0.30641
VSM	R_{BR} (t-score)	0.25037	0.49538
	R_{W1} (t-score)	0.26152	0.50213
	R_{W2} (t-score)	(0.06277)	0.30819
	Jaccard (PMI)	(-0.00762)	0.27990
	cosine (t-score)	0.17267	0.35724
Lemur_1	Indri_rank	0.28690	0.53497
	tf-idf_rank	0.18964	0.45041
	KL_rank	0.28449	0.54251
	cosine (idf)	0.25343	0.51412
Lemur_3	Indri_hit	0.30143	**0.57135**
	Indri_rank	0.28421	0.54678
	KL_hit	**0.30303**	0.56666
	KL_rank	0.28790	0.55131
LSA	Infomap	(0.10042)	0.39467
	Infomap_neigh	0.14999	0.44427
	Infomap_doc	0.22994	0.50009

Table 1: Kendall's τ_B rank-correlations relative to an ideal compositionality ranking and average precisions (AP), obtained by different AMs and distributional similarity measures; non-significant values of τ_B in parentheses ($p > 0.05$).

in the document list retrieved for the different queries. For the experiments in which the context sentences have been distributed in different documents, average positions are calculated and weighted taking into account the amount of documents for each bigram analysis. In addition to that, the total number of documents in the list (or "hits") is weighted similarly.

Finally, precision curves are calculated for idioms, collocations and overall MWE extraction.

4. Results

The results for Kendall's τ_B and AP values are summarized in Table 1 (only the experiments with most remarkable results are included).

These first experiments are exploratory, and must be analyzed with caution. In any case, the use of measures such as the Indri index and KL-divergence inside Lemur brings a noticeable improvement in the results, even with respect to a baseline established by the best measures of association, as f and t-score. The best results for the different similarity measures in the VSM implementation were obtained including 50% of the collocates and selecting them according to their t-score weights in the case of R-values and cosine similarity, and PMI in the case of Jaccard index. The results obtained in Lemur_2 are not included in the table, as they are poorer than in the other two modalities, which show similar performances. The results of the Okapi index were clearly disappointing and have not been included; a specific tuning of its different parameters will be needed in future experiments. LSA-Infomap results are surprisingly low, and do not meet expectations, except in the case of the document-retrieval experiment (Infomap_doc,) whose results, at a given extent, are close to the ones obtained with other contexts-as-documents experiments with Lemur.

Measure combination	τ_B	AP
t_Indri(Lemur_1)_0.5	0.28896	0.57272
t_KL(Lemur_1)_0.5	0.28924	0.59326
t_KL(Lemur_1)_0.7	0.30625	**0.59413**
t_KL(Lemur_3)_0.5_hit	0.29650	0.56431
t_KL(Lemur_3)_0.7_hit	**0.32044**	0.57934
t_KL(Lemur_3)_0.7_rank	0.30592	0.55939

Table 2: Kendall's τ_B rank-correlations and average precisions (AP) obtained combining the ranking results of some association and distributional similarity measures.

Bearing in mind the possibility that the combination of measures could result in greater accuracy, we performed several trials with rank-aggregation using Borda's method (Dwork et al., 2001). As can be observed in Table 2, slight improvements are obtained; specifically, with t-score on the part of AMs, and Indri or KL-divergence as similarity measures (0.5 indicates equal weights for the measures aggregated; 0.7 indicates 3:7 ratio for the weights of t-score and the distributional measure).

Figure 1 shows the precision curves for the extraction of MWEs by some of the measures in Table 1 and 2. All the measures of distributional similarity outperform the best cooccurrence measures. The index combining t-score and KL-divergence in the Lemur_3 experiment (based on the weighted number of hits) [t_KL(Lemur_3)_0.7_hit] is slightly the best measure, even though the difference with the best distributional measures is hardly significant.

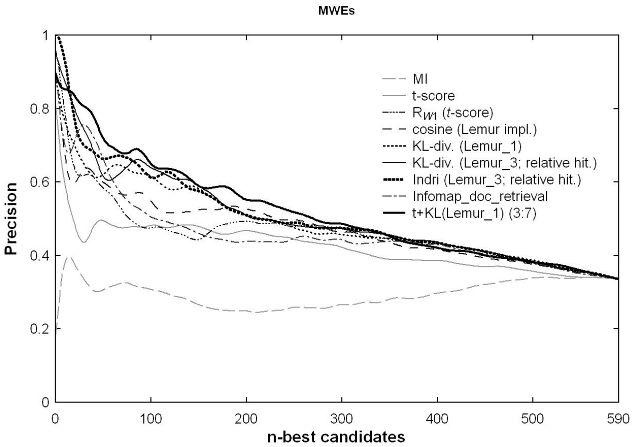

Figure 1: Precision results for the compositionality rankings of MWEs.

In Figure 2 and 3, we present separately the precision curves for idioms and collocations. The most remarkable point is that distributional similarity measures, specially the Indri index and KL-divergence, obtains significantly better ranks for idiomatic expressions than cooccurrence measures. Being idioms the least compositional expressions, this is the result expected, which supports the hypothesis that semantic compositionality can be better characterized using measures of distributional similarity than using association measures. Another interesting result is that PMI is in this case not significantly worse than other AMs, unlike precision graphs in Figure 1 and 3.

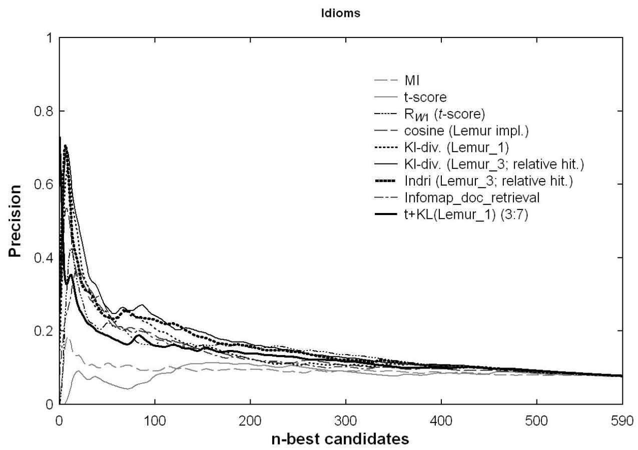

Figure 2: Precision results for the compositionality rankings of idioms.

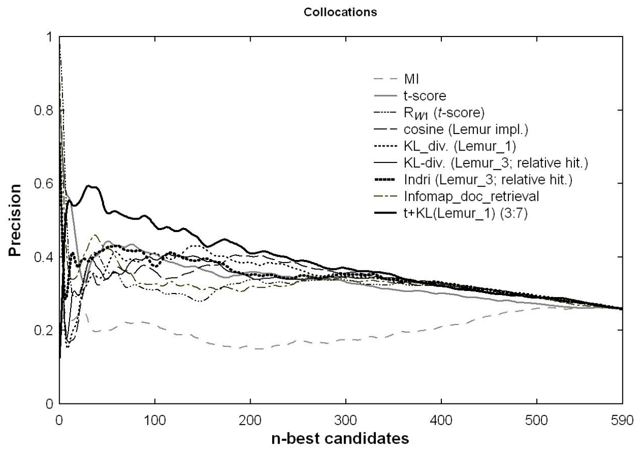

Figure 3: Precision results for the compositionality rankings of collocations.

Regarding the precision for collocations in Figure 3, association measures as t-score outperform distributional measures only in a narrow portion at the beginning of the ranking ($n < 50$), and thereafter, their values fluctuate within the same range (0.3-0.4), except for PMI and Infomap, whose precisions are quite worse.

In Figure 2, the combined measure t_KL(Lemur_3)_0.7 performs lower than simple distributional similarity measures, probably due to the poor contribution of t-score to the extraction of idioms. In contrast, it is the best measure for collocation extraction for $n > 25$.

Table 3 and 4 display the values of average precision separately for idioms and collocation rankings. In accordance with the precision curves in Figure 2, idioms are better ranked using exclusively distributionality measures as KL-divergence, without combining them with AMs. A for collocations, KL-divergence can hardly beat AMs as f or t-score, but their aggregation in equal parts (0.5) yields a clearly better average precision.

	Measure	Idioms	Collocations
	random rank	0.08381	0.25766
AM	f	0.08160	0.36902
	t-score	0.09136	0.36319
	LL	0.10417	0.33354
	PMI	0.10064	0.21893
VSM	R_{BR} (t-score)	0.23399	0.32112
	R_{W1} (t-score)	0.19570	0.34382
	R_{W2} (t-score)	0.14878	0.20491
	Jaccard (PMI)	0.10040	0.19851
	cosine (t-score)	0.11879	0.25661
Lemur_1	Indri_rank	0.23343	0.36127
	tf-idf_rank	0.13633	0.33570
	KL_rank	0.24631	0.36714
	cosine (idf)	0.21541	0.35449
Lemur_3	Indri_hit	0.27331	**0.36997**
	Indri_rank	0.26524	0.35103
	KL_hit	**0.30231**	0.35637
	KL_rank	0.29058	0.34686
LSA	Infomap	0.18512	0.25747
	Infomap_neigh	0.27375	0.26749
	Infomap_doc	0.18604	0.34692

Table 3: Average precisions (AP), obtained by different AMs and distributional similarity measures for idioms and collocations

Measure combination	Idioms	Collocations
t_Indri(Lemur_1)_0.5	0.15995	0.43512
t_KL(Lemur_1)_0.5	0.16302	**0.45475**
t_KL(Lemur_1)_0.7	0.18851	0.43438
t_KL(Lemur_3)_0.5_hit	0.18313	0.40676
t_KL(Lemur_3)_0.7_hit	0.22172	0.39551
t_KL(Lemur_3)_0.7_rank	0.21052	0.38315

Table 4: Average precisions (AP) for idioms and collocations obtained combining the ranking results of some association and distributional similarity measures.

5. Conclusions and Future work

The results obtained for Kendall's τ_B show that, in the task of ranking the candidates according to their semantic compositionality, the Indri index and KL-divergence outperform the other distributional similarity measures tested, as well as the association measures. In the case of distributional similarity measures, further research should be undertaken to corroborate this outcome. In comparison with AMs, this is mostly due to the fact the Indri index or KL-divergence obtained much better results than AMs in the characterization of idioms as non-compositional combinations. In the case of collocations, no such claim can be made looking at the average precision results, and we conclude that, for the extraction of collocations, statistical idiosyncrasy is a property as significant as compositionality when used separately. Even though the slight improvement obtained with rank-aggregation could be hardly taken as statistically significant, these first trials create expectations that the combination of the different features involved in idiomaticity could provide better results (Fazly and Stevenson, 2007). These expectations are clearly justified in the case of collocations, whose AP results improve noticeably when combining t-score and KL-divergence with equal weights.

As for a whole account of idiomaticity, the next steps would be to integrate measures of lexico-syntactic flexibility into the system, and to explore the application of machine learning to automatically detect and characterize the idiomaticity of MWEs (Katz and Giesbrecht, 2006).

6. Acknowledgments

This research was supported in part by the Spanish Ministry of Education and Science (OpenMT-2, TIN2009-14675-C03-01) and by the Basque Government (Berbatek: Tools and Technologies to promote Language Industry. Etortek - IE09-262; KONBITZ project, Saiotek 2011). Our colleagues Ainara Estarrona and Larraitz Uria from the IXA group are kindly acknowledged for providing their expertise as linguists in the manual evaluation process. Maddalen Lopez de la Calle and Iñaki San Vicente from Elhuyar Foundation and Oier Lopez de la Calle from the IXA group have contributed with their expertise to the design of the experiments with Lemur and Infomap. Finally, special thanks goes to Yosu Yurramendi from the University of the Basque Country, for his advice on the statistics, specially in the evaluation step.

7. References

I. Aduriz, I. Aldezabal, I. Alegria, X. Artola, N. Ezeiza, and R. Urizar. 1996. Euslem: A lemmatiser/tagger for Basque. *Proc. of EURALEX96*, pages 17–26.

J. Allan, J. Callan, K. Collins-Thompson, B. Croft, F. Feng, D. Fisher, J. Lafferty, L. Larkey, T.N. Truong, P. Ogilvie, et al. 2003. The Lemur Toolkit for language modeling and information retrieval.

T. Baldwin and S.N. Kim. 2010. Multiword expressions. *Handbook of Natural Language Processing, second edition. Morgan and Claypool.*

T. Baldwin, C. Bannard, T. Tanaka, and D. Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, page 96.

G.L.M. Berry-Rogghe. 1974. Automatic identification of phrasal verbs. *Computers in the Humanities*, pages 16–26.

C. Biemann and E. Giesbrecht. 2011. Distributional semantics and compositionality 2011: Shared task description and results. *Workshop on Distributional semantics and compositionality 2011. ACL HLT 2011*, page 21.

S.D. Bolboacă and L. Jäntschi. 2006. Pearson versus Spearman, Kendall's tau correlation analysis on structure-activity relationships of biologic active compounds. *Leonardo Journal of Sciences*, (9):179–200.

A.P. Cowie. 1998. *Phraseology: Theory, analysis, and applications*. Oxford University Press, USA.

D.A. Cruse. 1986. *Lexical semantics*. Cambridge Univ Pr.

C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. 2001. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622.

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

M. Garrao, C. Oliveira, M. de Freitas, and M. Dias. 2006. Corpus-based compositionality. *Computational Processing of the Portuguese Language*, pages 268–271.

S. Granger and M. Paquot. 2008. Disentangling the phraseological web. *Phraseology. An interdisciplinary perspective*, pages 27–50.

A. Gurrutxaga and I. Alegria. 2011. Automatic extraction of NV expressions in Basque: basic issues on cooccurrence techniques. *ACL HLT 2011*, page 2.

G. Katz and E. Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19. Association for Computational Linguistics.

I. Korkontzelos and S. Manandhar. 2009. Detecting compositionality in multi-word expressions. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 65–68. Association for Computational Linguistics.

B. Krenn, S. Evert, and H. Zinsmeister. 2004. Determining intercoder agreement for a collocation identification task. In *Proceedings of KONVENS*, pages 89–96.

D. Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 317–324. Association for Computational Linguistics.

C. D. Manning and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

I. Melcuk. 1998. Collocations and lexical functions. *Phraseology. Theory, Analysis, and Applications*, pages 23–53.

B.H. Partee. 1995. Lexical semantics and compositionality. *An invitation to cognitive science: Language*, 1:311–360.

P. Pecina. 2010. Lexical association measures and collocation extraction. *Language resources and evaluation*, 44(1):137–158.

P. Schone and D. Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem. In *Proc. of the 6th Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*, pages 100–108. Citeseer.

J. Sinclair. 1996. The search for units of meaning. *Textus*, 9(1):75–106.

S. Wulff. 2010. *Rethinking Idiomaticity*. Corpus and Discourse. Continuum International Publishing Group Ltd, New York.