Investigating the impact of sample size on cognate detection

Johann-Mattis List

Investigating the Impact of Sample Size on Cognate Detection

Johann-Mattis List (Philipps-University Marburg)

2013-01

1 Cognate Detection in Historical Linguistics

In historical linguistics, the problem of cognate detection is traditionally approached within the framework of the comparative method (Trask 2000: 64-67, Fox 1995). The most important aspects of this traditional method for cognate detection are a language-specific notion of word similarity, derived from previously identified regular sound correspondences, and the iterative character of the method, by which proposed lists of cognates and sound correspondences are constantly refined and updated (Durie 1996: 6f). Being a manual method which was never really laid out in a strict algorithmic way, there are many parameters which were never really specified in the methodological literature. It is left open how many languages should be compared (a), or whether the genetic relatedness between these languages should have been already proven (b). It is also not specified whether the cognate sets to be identified should be restricted to semantically similar words or whether words that greatly diverge semantically should also be included in the comparison (c). Furthermore, the sample size, i.e. the size of the word lists in which linguists search for cognates, is left undefined (d).

For the successful application of the method it is irrelevant whether the first three parameters (a, b, and c) are specified or not. The method is indifferent regarding the number of languages being compared, it has its own procedure to determine genetic relatedness between languages, and semantically different but formally similar words have seldom posed a problem for historical linguists. The last parameter (d), the size of the word lists, however, is of crucial importance for the method, although nobody has so far been able to determine how many items a word list should at least contain in order to be applicable. That the popular Swadesh-200 word lists (Swadesh 1952) are surely not enough when questions of remote relationship have to be solved can be easily demonstrated when considering the amount of cognate words in these word lists for some genetically related languages such as as Armenian, English, French, and German: Given that there are maximally 20 cognates between Armenian and the other three languages, it is hardly possible that these cognates are enough to set up a satisfying set of sound correspondences between these languages. It might also be questioned whether the number of cognates attested between French and the Germanic languages is enough for a rigorous application of the comparative method.

2 Sample Size and Cognate Detection

Given that sample size is crucial for the success of the comparative method, it would be desirable to have at least a rough estimate regarding the lower bound of how many words are needed for the task

	Albanian	English	French	German
Albanian		0.07	0.10	0.10
English	14		0.23	0.56
French	20	46		0.23
German	20	111	46	

Table 1: Number and proportion of cognates within Swadesh-200 word lists of four Indo-European languages. Cognate counts are based on the data given in Kessler (2001).

of cognate detection. Stating that a word list of 200 items is not enough for the comparative method to successfully prove the genetic relationship between Albanian and English does not really solve the question of how many words are needed, neither in general, nor in this specific case. Such an estimate would, of course, depend on the genetic closeness of the languages being compared, and it would surely vary accordingly. Nevertheless, it would be helpful to know how many items one needs at least in order to successfully compare languages as divergent as, say, German and French. Given the manual character of the comparative method, it is not easy to investigate the problem by simply applying the method to randomly varying sizes of a given word list. Not only would it be too time-consuming to conduct all the analyses, it would also be difficult to maintain objectivity when having the same sample of languages being investigated again and again by the same scholar. Fortunately, there are alternative ways to investigate the impact of sample size on cognate detection which do not rely on a manual application of the comparative method. Since the reason, why the comparative method relies so heavily on sample size is its language-specific similarity notion, it is enough to employ an automatic method for cognate detection that closely mimics the comparative method regarding the underlying notion of word similarity, and apply it to varying samples of a large gold standard containing cognate judgments taken from the literature.

2.1 Language-Specific and Language-Independent Similarities

It is useful to make a distinction between language-specific and language-independent notions of word similarity. Language-specific similarity is hereby understood as similarity between words which is reflected in regular sound correspondences. Lass (1997: 130) calls this kind of similarity *genotypic* as opposed to *phenotypic similarity*, which is based on surface resemblances of phonetic segments, but the most crucial aspect of this kind of similarity is that it is language-specific. It is never defined in general terms but always with respect to the language systems which are being compared. Correspondence relations can therefore only be established for individual languages, they can never be taken as general statements. As an example, consider the two words English *mouth* [maʊð] and German *Mund* [mʊnt] “mouth”. From a language-specific perspective, these two words are maximally similar, since all correspondences, which are reflected in the alignment of the words, occur regularly, even the null-correspondence German [n] ≈ English [-] (Starostin 2010: 95. From a language-independent perspective, however, there are phonetically much more similar candidates to compare in both languages, such as, e.g., English *mount* [maʊnt], or German *Maus* [maus] “mouse” . In contrast to language-independent phenotypic similarities, language-specific similarities can never be proposed by relying on one word pair alone. This is the reason why the comparative method so heavily relies on the sample size: The smaller a sample is, the greater the possibility that it does not contain enough cognate words that make it possible to detect these specific similarities.

2.2 Language-Specific Automatic Cognate Detection

LexStat (List 2012a) is a publicly available method for automatic cognate detection based on language-specific similarities. The method takes multilingual (usually semantically aligned) word lists in IPA transcription as input and returns the same list with additional cognate judgments as output. The basic working procedure of the method consists of five stages: (1) sequence conversion, (2) preprocessing, (3) scoring-scheme creation, (4) distance calculation, and (5) sequence clustering. In stage (1), the input words are converted into tuples consisting of sound classes and prosodic strings (cf. List 2012b regarding the idea behind sound classes and prosodic strings). In stage (2), a simple language-independent method is used to derive preliminary cognate sets. In stage (3), a Monte-Carlo permutation test is used to create language-specific log-odds scoring schemes for all language pairs. In stage (4) the pairwise distances between all word pairs, based on the language-specific scoring schemes, are computed. In stage (5), the sequences are clustered into cognate sets whose average distance is beyond a certain threshold. In addition to these five stages, all cognate sets detected by the method are aligned, using the SCA method for multiple phonetic alignment (ibid.). As was shown in (List 2012a), LexStat largely outperforms alternative methods that rely on language-independent similarities, such as the sound-class-based method proposed by Turchin et al. (Turchin et al. 2010), or alignment-based methods, such as normalized edit distance (NED), or sound-class-based alignment distance (SCA, List 2012b). Given that LexStat closely mimics the comparative method regarding the underlying notion of word similarity, it seems to be an ideal candidate to test the impact of sample size on cognate detection.

3 Testing the Impact of Sample Size

3.1 Gold Standard

In order to test to which degree language-specific methods for cognate detection depend on the samples size, an analysis of different, randomly created partitions taken from a newly compiled large gold standard was carried out. The gold standard consists of 550 items translated into four languages (German, English, Dutch, and French) which were taken from the Intercontinental Dictionary Series (IDS). The orthographic entries in the original were converted into IPA transcriptions by the author, relying on one dictionary source for each language in order to maintain consistency. Cognate judgments were applied manually by consulting the respective literature (KLUGE, REW, OREL, PFEIFER, VAAN, NIL). [1]

3.2 Test Samples

With its 550 glosses translated into four languages, this gold standard is much larger than other publicly available datasets with respect to sample size. The data for the test was created as follows: Starting from the basic gold standard containing all 550 items, 550 new subsets of the data were created by randomly deleting 5, 10, 15, etc. items from the original dataset and taking 5 different samples for each distinct number of deletions. This process yielded 550 datasets, covering the whole range of possible sample sizes between 5 and 550 in steps of 5. These datasets were then analyzed, using the LexStat method and the three above-mentioned language-independent methods (Turchin, NED, SCA, see List 2012a for details).

[1] The dataset is not yet published, but the author will gladly share it upon request.

3.3 Evaluation Measures

In applications of information retrieval it is common to evaluate algorithms by calculating their precision and recall. Precision refers to the proportion of items in the test set that also occur in the reference set. Recall refers to the proportion of items in the reference set that also occur in the test set (Witten and Frank 2005: 171). In the context of automatic cognate detection, a high precision is equivalent to a low proportion of false positives, and a high recall is equivalent to a high proportion of correctly identified cognates. Since the main interest of our experiment was to test the impact of sample size on cognate detection, we calculated the average B-Cubed recall of all five subsets for each sample size. B-Cubed scores were originally introduced as part of an algorithm by (Bagga and Baldwin 1998), but (Amigó et al. 2009) could show that they are especially apt as a clustering evaluation measure.[2]

4 Results

The results of this analysis are plotted in Figure 1. As can be seen from the figure, the results of the three language-independent methods are quite similar regarding their tendency. Only the degrees of the scores differ. The scores themselves show only marginal variations and remain constant regardless of the sample size. The results for the language-specific LexStat analysis, on the other hand, clearly depend on the sample size, growing logistically, until converging around a sample size of 200 items. This nicely reflects the language-specific character of the LexStat method: If the word lists fed to the algorithm are too small, no language-specific similarities can be inferred, and no cognates can be detected, as reflected by the low recall for small word lists. This changes dramatically once the sample size is increased. Comparing the scores for a sample size of 50 items (90.88) with those of 100 items (93.89), an increase of about 3 points can be attested, and between 100 and 200 items (95.02), there is still an increase of more than 1 point (see Table 2).

Items	B-Cubed Recall			
	Turchin	NED	SCA	LexStat
50	86.10	85.55	92.44	90.88
100	86.55	85.77	92.20	93.89
200	86.88	86.61	92.68	95.02
300	87.13	86.64	92.90	95.05
400	87.14	86.81	92.89	94.94
500	87.07	86.77	92.75	94.90

Table 2: B-Cubed recall of the four different automatic methods in randomly created subsamples of varying sample size extracted from the gold standard.

5 Conclusion

One might wonder whether the fact that the scores converge at a sample size of 200 allows to conclude that 200 words are enough for the preliminary stages of language comparison. Since, to my knowledge, the gold standard presented in this study is the only available one covering more than 500 items, it

[2] For details on how the scores are calculated for the evaluation of cognate judgments, see (Hauer and Kondrak 2011).

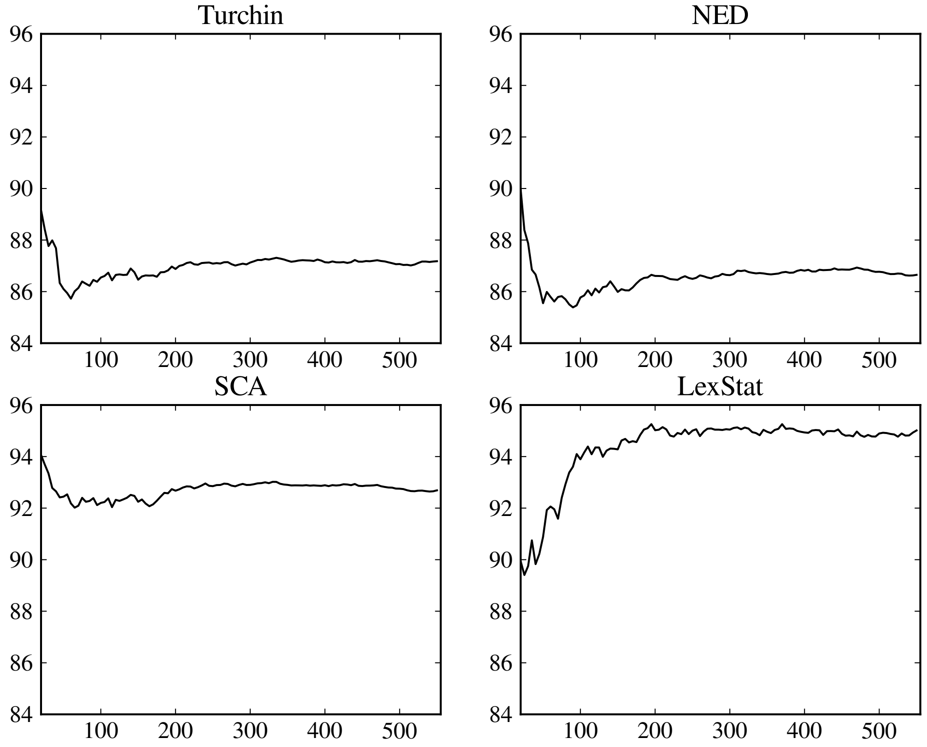

Figure 1: Performance of the methods in dependence of the sample size (number of vocabulary items per word list). The figures show the B-Cubed recall for the four methods.

may be questioned whether the data is representative enough to draw general conclusions regarding the necessary size of word lists for automatic cognate detection methods. Nevertheless, what the results of the analysis show is that word list size indeed has an impact on the results. Thus, when using language-specific methods, there is no use in taking word lists with less than 100 items. 200 words, however, are surely a good start for languages as closely related as German and French. However, whether 200 words are enough for cases of remote relationship remains questionable. More analyses on larger samples are needed to shed light on this question.

References

Amigó, E., J. Gonzalo, J. Artiles, and F. Verdejo (Aug. 2009). "A comparison of extrinsic clustering evaluation metrics based on formal constraints". In: *Information Retrieval* 12.4, 461–486.

Bagga, A. and B. Baldwin (1998). "Entity-based cross-document coreferencing using the vector space model". In: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*. "COLING-ACL '98" (Montréal, Quebec, Canada, Aug. 10–14, 1998). Association of Computational Linguistics, 79–85.

Durie, M., ed. (1996). *The comparative method reviewed. Regularity and irregularity in language change*. With an intro. by M. D. Ross and M. Durie. New York: Oxford University Press.

Fox, A. (1995). *Linguistic reconstruction. An introduction to theory and method*. Oxford: Oxford University Press.

Hauer, B. and G. Kondrak (2011). “Clustering semantically equivalent words into cognate sets in multilingual lists”. In: *Proceedings of the 5th International Joint Conference on Natural Language Processing*. (Chiang Mai, Thailand, Nov. 8–13, 2011). AFNLP, 865–873.

Kessler, B. (2001). *The significance of word lists. Statistical tests for investigating historical connections between languages*. Stanford: CSLI Publications.

Lass, R. (1997). *Historical linguistics and language change*. Cambridge: Cambridge University Press.

List, J.-M. (2012a). “LexStat. Automatic Detection of Cognates in Multilingual Wordlists”. In: *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*. (Avignon, France, Apr. 23–24, 2012). Association for Computational Linguistics, 117–125.

– (2012b). “SCA. Phonetic alignment based on sound classes”. In: *New directions in logic, language, and computation*. Ed. by M. Slavkovik and D. Lassiter. LNCS 7415. Berlin and Heidelberg: Springer, 32–51.

Starostin, G. (2010). “Preliminary lexicostatistics as a basis for language classification: A new approach”. In: *Journal of Language Relationship* 3, 79–116.

Swadesh, M. (1952). “Lexico-statistic dating of prehistoric ethnic contacts. With special reference to North American Indians and Eskimos”. In: *Proceedings of the American Philosophical Society* 96.4, 452–463. JSTOR: `3143802`.

Trask, R. L., comp. (2000). *The dictionary of historical and comparative linguistics*. Edinburgh: Edinburgh University Press.

Turchin, P., I. Peiros, and M. Gell-Mann (2010). “Analyzing genetic connections between languages by matching consonant classes”. In: *Journal of Language Relationship* 3, 117–126.

Witten, I. H. and E. Frank (2005). *Data mining. Practical machine learning tools and techniques*. 2nd ed. Amsterdam et al.: Elsevier.

Dictionaries and Databases

IDS M. R. Key and B. Comrie, eds. (2007). *IDS – The Intercontinental Dictionary Series*. URL: `http://lingweb.eva.mpg.de/ids/`.

KLUGE F. Kluge, found. (2002). *Etymologisches Wörterbuch der deutschen Sprache*. Cont. by E. Seebold. 24th ed. Berlin: de Gruyter.

NIL D. Wodtko, B. Irslinger, and C. Schneider, eds. (2008). *Nomina im Indogermanischen Lexikon*. Heidelberg: Winter.

OREL V. Orel, comp. (2003). *A handbook of Germanic etymology*. Leiden: Brill.

PFEIFER W. Pfeifer, ed. (1993). *Etymologisches Wörterbuch des Deutschen*. 2nd ed. 2 vols. Berlin: Akademie. URL: `http://www.dwds.de/`.

REW W. Meyer-Lübke, comp. (1911). *Romanisches etymologisches Wörterbuch*. Sammlung romanischer Elementar- und Handbücher 3.3. Heidelberg: Winter.

VAAN M. Vaan (2008). *Etymological dictionary of Latin and the other Italic languages*. Leiden Indo-European Etymological Dictionary Series 7. Leiden and Boston: Brill.