Johann-Mattis List
Philipps-University (Marburg)

Investigating the impact of sample size on cognate detection

The paper deals with the question of how many words are needed to successfully apply different methods for cognate detection. In order to investigate this question, a large gold standard consisting of 550 concepts translated into 4 languages (English, German, Dutch, and French) was compiled and divided into subsets of increasing sample size. Applying automatic methods for cognate detection on this gold standard shows that the accuracy of language-specific cognate detection methods clearly depends on the sample size. However, given that sample size depends on various different factors such as the genetic closeness of the languages or the degree of contact between the languages under investigation, no general lower or upper bound can be determined from the analysis.

Keywords: Comparative method, lexicostatistics, etymology, computational linguistics.

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, which is 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: 6 f.). Being a non-automatic method which was never really laid out in a strict algorithmic way, there are many parameters which were never specified in the methodological literature. It is left open

(a) how many languages should be compared,
(b) whether or not the genetic relatedness between these languages should have been already proven,
(c) whether or not the cognate sets to be identified should be restricted to semantically similar words, and
(d) how many word pairs of all languages should be included in the survey (henceforth referred to as *sample size*).

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 consid-

ering the amount of cognate words in these word lists for some genetically related languages such as Armenian, English, French, and German (see Table 1): 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. In order to prove the genetic relationship of Armenian with English, French, and German, it is unavoidable to expand the sample.[1] 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.

Table 1: Number (lower triangle) and proportion (upper triangle) of cognates within Swadesh wordlists of 200 items for four Indo-European languages. Cognate counts are based on the data given in Kessler (2001).

	Armenian	English	French	German
Armenian	200 \ 1.0	0.07	0.10	0.10
English	14	200 \ 1.0	0.23	0.56
French	20	46	200 \ 1.0	0.23
German	20	111	46	200 \ 1.0

However, as can also be seen from the examples of shared cognate percentages in Table 1, the question of how many items constitute the "ideal sample size" for the successful application of the comparative method also depends on the genetic closeness of the languages under investigation: While Swadesh lists may not be enough to prove genetic relationship between Armenian and, say, German, they surely provide enough evidence to prove the genetic relationship between German and English. In cases of remote relationship, however, it becomes increasingly difficult to find enough initial pairs of cognate words to establish sound correspondences rigorously (Starostin 2013: 57–65).

But this is again only part of the whole problem, since cognate words are not the only *historical similarities* that can be detected when applying the comparative method. That similarities arising from language contact can seriously influence the results of the comparative method has been noticed by linguists for a long time. Increasing the sample size also increases the chance of finding contact-induces similarities. In cases of heavy contact, only the rigorous stratification of cognate candidates and proposed sound correspondences can help to disentangle borrowed from inherited traits. As a result, the sample size does not only have a theoretical minimum, but also a theoretical maximum. If the sample is too small, one may fail to detect the relevant similarities between genetically related languages. If the sample is too large, one may detect similarities which are not the result of genetic inheritance.

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 of cognate detection. Stating that a word list of 200 items is not enough for

[1] Note that "sample size" here means all words-comparisons that are needed to establish an initial set of sound correspondences between two languages. Thus, what I to address here is the question of the size of the "additional material" which the Moscow school of historical-comparative linguistics requires as a backbone to establish genetic relationship with help of Swadesh's (1955) list of 100 items (Dybo and Starostin 2008, Starostin 2013: 30–44).

the comparative method to successfully prove the genetic relationship between Armenian and English does not really solve the question. We still don't know how many words are needed for a successful application of the comparative method, 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. We can, for example, use methods 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. However, 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-Independent Approaches to Cognate Detection.

Most of the current automatic approaches to cognate detection employ a language-independent notion of similarity. The method by Turchin et al. (2010), for example, builds on Dolgopolsky's (1964) idea of *sound classes*. All words passed to the method are first converted to their respective Dolgopolsky sound classes and all words whose first two consonant classes match are assumed to be cognate. As an example, consider the two words English *mouth* [maʊð] and German *Mund* [mʊnt] "mouth". Converting the words into their Dolgopolsky sound classes (vowels

being ignored), this yields the two strings “MT” and “MNT”. Since the first two consonant classes do *not* match, the method by Turchin et al. (2010) assumes that the words are not cognate.

As an alternative, alignment algorithms can be used to calculate the *edit distance* between two words. The edit distance between two words is defined as the smallest number of *edit operations* (*deletion*, *insertion*, *substitution*) needed to transform one word into the other (Levensthein 1965). This is equivalent to the Hamming distance of the alignment of two words (Hamming 1950). It can further be normalized by dividing it by the length of the longer word. Once the pairwise normalized edit distance (NED) is computed for a given pair of words, one can define a specific threshold below which the words are judged to be cognate. As an example, consider again the two words English *mouth* [maʊð] and German *Mund* [mʊnt] “mouth”. The optimal alignment of both words is:

m	au	–	θ
m	ʊ	n	t

and the edit distance between both words is thus 3 (since they differ in three positions in the alignment). The normalized edit distance (NED) is 3 / 4 = 0.75. Assuming a threshold of 0.6, the NED approach will also assume that both words are not cognate.

2.3. Language-Specific Approaches to Automatic Cognate Detection.

LexStat (List 2012a) is a new method for automatic cognate detection based on language-specific similarities. The method is implemented as part of a larger Python library for quantitative tasks in historical linguistics (List and Moran 2013) and can be downloaded from http://www.lingpy.org. LexStat 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,
(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 (List 2012b). As was shown in List (2012a), LexStat largely outperforms alternative methods that rely on language-independent similarities, such as the above-mentioned sound-class-based method proposed by Turchin et al. (2010), or alignment-based methods, such as normalized edit distance (NED). Given that LexStat closely mimics the comparative method regarding the underlying notion of word similarity, it seems to be a good 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 (Key & Comrie 2007). 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 and Seebold 2002, Meyer-Luebke 1911, Pfeifer 1993, Vaan 2008, Wodtko 2008). Borrowings were coded in two ways: In the first coding, borrowed words were assigned to separate cognate sets. In the second coding, borrowings were assigned to the cognate sets to which they would belong if they were *not* borrowed. The second coding procedure is common in evolutionary biology where the term *homology* is used to indicate that two genes share a common history without specifying whether this common history is due to vertical inheritance or lateral transfer (Fitch 2000). For our experiment, it may be interesting to code borrowed words as cognates, since it may give us some hints whether and to which degree borrowing influences the results of language-specific cognate detection algorithms. For the downloadable gold standard, see Supplementary materials.

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, the method by Turchin et al. (2010), and the NED method (see List 2012a for details). In contrast to the NED and the LexStat method, the method by Turchin et al. (2010) does not need to be passed a specific threshold for cognate detection, since the threshold (two matching consonant classes) is inherent in the method itself. Choosing optimal thresholds for automatic cognate detection methods is not trivial, and no methods to automatically infer optimal thresholds are available. In order to apply a consistent criterion for threshold selection, the thresholds for NED and LexStat were calibrated on the results of the Turchin method. This was done by applying LexStat and NED to the largest sample, using several varying thresholds. Of all results, those thresholds were picked in which the number of false positives proposed by LexStat and NED came closest to the results of the Turchin method. This calibration procedure yielded an "optimal" threshold of 0.65 for NED, and a threshold of 0.625 for LexStat.

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. Precision and recall can be summarized by calculating their harmonic mean, the so-called *F-scores*, using the formula $2 \times (P \times R) / (P + R)$, where *P* is the precision and *R* is the recall.

Among different evaluation measures which have been proposed to estimate the accuracy of automatically induced cognate judgments (see Bergsma & Kondrak 2007), *B-Cubed* scores were chosen. 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, and Bergsma & Kondrak (2007) showed that they are very useful to estimate the performance of cognate detection algorithms.

4. Results

4.1 Sample Size and General Accuracy of Automatic Cognate Detection.

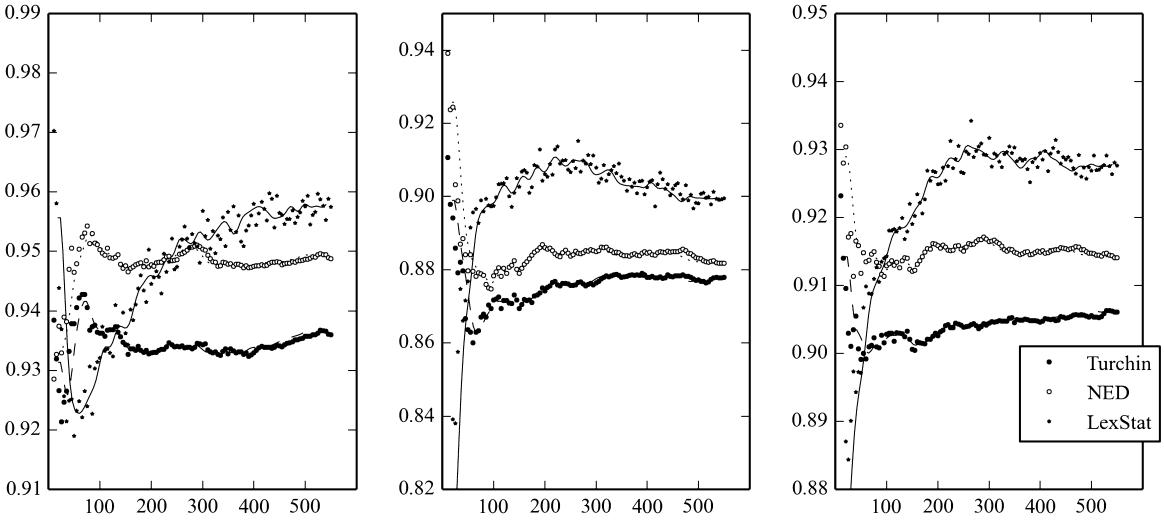
The results of the general analysis (precision, recall, and F-scores for the cognate detection task) are plotted in Figure 1. As can be seen from the figure, the results of the two language-independent methods are quite similar regarding their tendency. After an initial phase of scattered results in those tests where the sample size is low, they stabilize 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. Both recall and F-Scores grow logistically and converge around a sample size of 200 items and 300 items, respectively. 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 and F-scores for small word lists. This changes dramatically once the sample size increases. Comparing the scores for a sample size of 50 items (F-score ca. 0.90) with those of 100 items (F-Score ca. 0.915), an increase of about 0.015 points can be attested, and between 100 and 300 items (F-Score cs. 0.93), there is still an increase of more than 0.02 points.

The scores for precision seem also to show a logistical growth, although it is not possible to determine a definite point of convergence for the given range of sample sizes. The drastic initial decrease of precision is a relic of the B-Cubes measure: If no cluster decision is being made, i.e. if all words are assigned to different cognate sets, the B-Cubed precision is 1, since no erroneous cluster decisions have been made. Since LexStat tends to leave most of the words unclassified if not enough evidence can be found to assign them to the same cluster, it automatically commits only a few erroneous decisions when dealing with small samples.

4.2 Optimal Sample Size and Genetic Closeness.

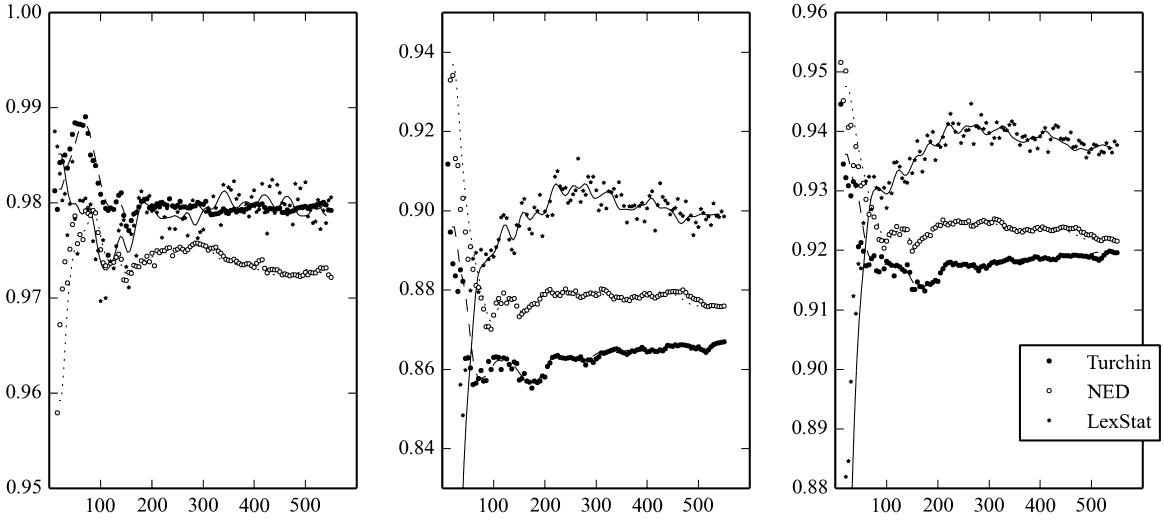
Figure 2 shows the results of the analyses for the Germanic languages in the sample. Basically, the results show a similar tendency as was observed for the analysis of all four languages. However, the increase in accuracy for the LexStat method is accelerated, and the convergence of the F-scores is reached at about 250 items (in contrast to 300 items in the full analysis). On the one hand, this illustrates the trivial fact that sample size directly depends on genetic closeness. On the other hand, it may seem surprising that the difference between the Germanic and the full sample is rather small (250 vs. 300 items in the F-Scores). One might argue that this is due to the fact that the Germanic languages also constitute the majority of the full sample. However, even when comparing further subsets like, for example, Dutch and Ger-

Figure 1: Comparing the performance of the methods for the cognate detection task. Y-axis shows the scores of the analyse, *X*-axis shows the size of the sample (number of basic vocabulary items).



(a) Precision (b) Recall (c) F-Score

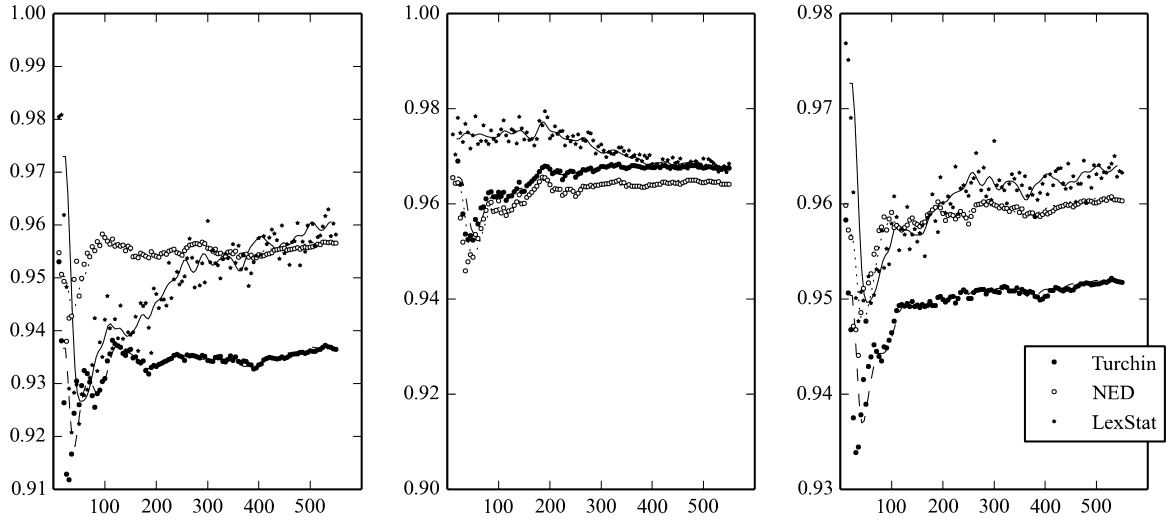
Figure 2: Comparing the performance of the methods on the subset of Germanic languages (English, German, Dutch).



(a) Precision (b) Recall (c) F-Score

man with English and French, the number of items when LexStat reaches convergence does not differ too much (250 vs. 300). It is possible that the number of 550 test items is still too small to conduct realistic tests on the impact of sample size on cognate detection. The current analyses might have missed interesting results which only show up when further increasing the sample size. Nevertheless, a striking difference between the full analysis and the Germanic subset is the increase in precision: While the precision of LexStat steadily increases in the full analysis along with the increase in sample size, it does not show this tendency in the Germanic subset where it quickly (at around 150–200 items) reaches a rather steady state. Since an increase in precision points to a decrease in false positives, this shows that for genetically close languages a much smaller sample suffices to achieve stable results.

Figure 3: Comparing the performance of the methods on the subset of English and French.



(a) Precision (b) Recall (c) F-Score

Figure 3 shows the specific results for the subset of English and French. These results are rather surprising: While precision steadily increases, recall shows a specific downtrend which starts at around 250 items. A decrease in recall corresponds to an increase in false negatives. With an increasing sample size, LexStat gets better in avoiding false positives, but at the same time gets worse in finding true cognates. One might think that this trend is somehow related to the increase of noise introduced by the large amount of borrowings from French into English. However, the same downtrend can also be observed when comparing the results for German and French, where the number of borrowings is much lower. Unfortunately, no further explanation for the results can be given at the moment.

4.3 Undetected Borrowings.

Given that undetected borrowings can yield a set of "wrong correspondences", they need to be identified and filtered out when applying the comparative method. That borrowings have a definite influence on the results of automatic cognate detection analyses is illustrated in Table 2. Here, the evaluation scores of the application of the three methods to the largest sample (550 items) of the gold standard are given in two "flavors". The first flavor is the performance on the traditional cognate detection task. The second flavor is the performance on the *homolog detection task*. In contrast to pure cognate detection, borrowings are explicitly included in this task, and the failure of a method to correctly identify borrowed words as homologs is penalized. As can be seen from the table, all three methods achieve a higher precision for the homolog detection task and a higher recall for the cognate detection task. This shows that both methods yield less false positives but more false negatives when no difference between borrowings and cognates is being made. The difference in precision, however, is much smaller for the LexStat method, and the F-Scores are higher for the cognate than for the homolog detection task, while they are identical in case of the NED and the Turchin method. This shows that the LexStat method handles noise arising from language contact much better than the NED method. Nevertheless, it also shows that LexStat can definitely be betrayed by large amounts of borrowings in the data. In terms of concrete numbers, of 176 borrowings with a direct donor in one of the languages, LexStat wrongly identifies 104 as cognates, NED 117, and the Turchin

Table 2: Comparing the performance of NED, Turchin, and LexStat in the cognate and the homolog detection task.

Method	Task	Precision	Recall	F-Score
NED	Cognates	0.95	0.88	0.91
	Homologs	0.98	0.86	0.91
Turchin	Cognates	0.94	0.88	0.91
	Homologs	0.97	0.86	0.91
LexStat	Cognates	0.97	0.90	0.93
	Homologs	0.98	0.86	0.92

method 140. This shows that borrowing definitely constitutes a problem for automatic cognate detection analyses.

An interesting question is whether stratification can make a difference in automatic cognate detection. In order to test this, a further test with the LexStat method was carried out. While the original LexStat method draws the attested distribution of possible sound correspondences from the *whole sample* it is given, the initial sample for the attested distribution was now restricted to basic vocabulary items drawn from the 100 and the 200 concept list proposed by Swadesh (1955 and 1952). The results for these analyses were compared to random trials. In these trials, the sample size was also restricted to 100 and 200 concepts, but the selection of concepts was carried out at random. The trials were repeated 50 times each, and the average of the results were compared with those obtained for the analyses based on sound correspondences derived from a stratification of the data.

Table 3: Comparing the impact of stratification on erroneous classification of English borrowings.

Items	Stratification	F-Score	Erroneously classified borrowings
100	random	0.85	0.28
	basic	0.86	0.17
200	random	0.88	0.35
	basic	0.87	0.19

Table 3 gives the proportion of missclassified French borrowings in English in the two analyses. As can be seen clearly, the number of erroneously classified borrowings is much lower in the analyses in which the initial sample was based on proper "basic vocabulary" than for randomly selected words pairs. This seems to indicate that stratification can indeed make a difference, also in automatic cognate detection. However, comparing the low F-Scores with those obtained for analyses in which the full sample was used also shows that a lot of interesting signal is lost. Further research is needed to find the right balance between signal loss resulting from stratification and unwanted noise resulting from large samples.

5. Discussion

The results reported in this study may be a bit disappointing, since it is not clear what they actually tell us. We still don't know the lower bound of words needed for a successful applica-

tion of the comparative method. We also don't find direct evidence for an upper bound, not to speak of the specific results of the analyses, which are generally difficult to explain. However, what the results definitely show is that word list size definitely *has* an impact on the results and that stratification *cannot be ignored*. More research with larger samples (both regarding the number of languages and the number of test items) is needed to shed light on the problems that were discussed in this study.

Supplementary materials are available from:
- http://jolr.ru/article.php?id=134
- https://gist.github.com/LinguList/8235795

The zip-archive includes:
- readme.md, a short description of the data-format;
- ids.qlc, the gold standard in QLC-format.

Literature

AMIGÓ, E., J. GONZALO, J. ARTILES, and F. VERDEJO (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.

BERGSMA, S. and G. KONDRAK (2007). "Multilingual cognate identification using integer linear programming". In: *RANLP Workshop on Acquisition and Management of Multilingual Lexicons*. Ed. by The International Conference on Recent Advances in Natural Language Processing. Borovets, Bulgaria.

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.

DYBO, A. and G. STAROSTIN (2008). "In defense of the comparative method, or the end of the Vovin controversy". In: *Aspekty komparativistiki*. Vol. 3. Ed. by I. S. Smirnov. Orientalia et Classica XI. Moscow: RGGU, 119–258. http://starling.rinet.ru/Texts/compmeth.pdf.

FITCH, W. M. (2000). "Homology. A personal view on some of the problems". In: *Trends in Genetics* 16.5, 227–231.

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

HAMMING, R. W. (1950). "Error detection and error detection codes". In: *Bell System Technical Journal* 29.2, 147–160.

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

KEY, M. R. and B. COMRIE, eds. (2007). *IDS — The Intercontinental Dictionary Series*: http://lingweb.eva.mpg.de/ids/.

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

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.

LIST, J.-M. (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.

LIST, J.-M. and S. MORAN (2013). "An open source toolkit for quantitative historical linguistics". In: *Proceedings of the ACL 2013 System Demonstrations. (Sofia, Bulgaria, Aug. 4–9, 2013)*. Association for Computational Linguistics, 13–18. http://aclweb.org/anthology/P/P13/P13–4003.pdf.

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

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

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

STAROSTIN, G. S. (2013). *Jazyki Afriki.Opyt postroenija leksistatističeskoj klassifikacii* [*The languages of Africa. Experience in establishing a lexicostatistical classification*]. Vol. 1: *Metodologija. Kojsanskie jazyki* [*Methodology. Khoisan languages*]. Moscow: Jazyki Russkoj Kul'tury

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.

SWADESH, M. (1955). "Towards greater accuracy in lexicostatistic dating". In: *International Journal of American Linguistics* 21.2, 121–137.

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.

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

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

WODTKO, D., B. IRSLINGER, and C. SCHNEIDER, eds. (2008). *Nomina im Indogermanischen Lexikon*. Heidelberg: Winter.

Й.-М. *Лист*. К вопросу о влиянии размера лексической выборки на обнаружение этимологических когнатов.

В статье исследуется вопрос об оптимальном размере словарного списка, на котором можно было бы апробировать различные методы детекции этимологических когнатов. Чтобы получить ответ на этот вопрос, был разработан «золотой стандарт» из 550 концептов, переведенных на 4 языка (английский, немецкий, голландский, французский); внутри этого списка было выделено несколько последовательно увеличиваемых подмножеств. Применение автоматических методов детекции когнатов к этому стандарту показывает, что степень точности методов, разработанных для конкретных языковых типов, явно зависит от размера списка. Учитывая, однако, что оптимальный размер зависит от столь различных факторов, как степень генетической близости языков и масштаб ареальных контактов между сравниваемыми языками, нельзя сказать, что анализ позволяет определить универсальную верхнюю или нижнюю границу списка.

Ключевые слова: сравнительный метод, лексикостатистика, этимология, компьютерная лингвистика.