Studia Linguistica Universitatis Iagellonicae Cracoviensis
128 (2011)

KAMIL STACHOWSKI
Jagiellonian University, Cracow

A NOTE ON LEVENSHTEIN DISTANCE VERSUS HUMAN ANALYSIS

Keywords: Levenshtein distance, loanword adaptation, Dolgan, Russian

Abstract

This paper argues that automatic phonetic comparison will only return true results if the languages in question have similar and comparably lenient phonologies. In the situation where their phonologies are incompatible and / or restrictive, linguistic knowledge of both of them is necessary to obtain results matching human perception. Whilst the case is mainly exemplified by Levenshtein distance and Russian loanwords in Dolgan, the conclusion is also applicable to the approach as a whole.

0. Rationale and introductory notes

In Stachowski (2010), I presented a method of quantifying the phonetic adaptation of loanwords, which heavily depends on prior human analysis. It has been suggested to me that it would be more valuable if the requirement could be removed for an *expert analyst to specify the adaptations ahead of time*. This question leads directly to the problem of how much linguistic knowledge, or knowledge of the languages being analyzed, is necessary for the results of an automatized assay to correlate with human (native speakers') perception.

Levenshtein distance (= edit ~) has more than once been shown to be capable of credible results (see e.g. Heeringa et al. 2006), even for *genetically* and *typologically* quite distant languages, as Kipchak Turkic vs. Iranian in van der Ark et al. (2007). However, it seems that this is much more often applied to *phonologically* quite similar languages such as Dutch, English, German or Norwegian dialects. Moreover, most of these languages are phonotactically relatively rich and therefore lenient, which appears to be key here. This is not at all the case with Dolgan, the Russian loanwords in which I have attempted to analyze. Neither is it the case with a great number of different languages, Turkic and other, to which in theory the method can be applied.

Levenshtein distance has also been criticized for its crudity, resulting in the charge that it *so completely misrepresents the nature of language* (Heggarty 2006: 185). A number of refinements have been proposed, and also a number of other algorithms of varying degrees of advantageousness (see e.g. Heeringa et al. 2006 or Nerbonne, Heeringa 2009, etc.). Nevertheless, I have chosen to use the basic version of the method here for its popularity and simplicity, and because it represents quite well the crucial methodological assumptions common to at least the majority of propositions.

I will: **1.** present the results of contrasting Levenshtein distance with my index of nativization, as applied to Russian loanwords in Dolgan, **2.** provide some further and typologically different examples of the incompatibility of Levenshtein distance with human perception, and **3.** conclude in, hopefully, a positive way.

1. Russian loanwords in Dolgan

In Stachowski (2010), I calculated for each of the 1169 identified Russian loanwords in Dolgan, an index of nativization (= degree of adaptation). It ranges from 0 (not nativized) to 1 (fully nativized). Examples: Russ. *aèropórt* 'airport' > Dolg. *aèroport* id. (index 0), *lódka* 'boat' > *lokka* id. (0.50931), *vétka* '*Siber.* canoe' > *bǟkkä* id. (1).

The leading assumption of this method is that adaptations which are more common contribute less to the final score than those which are rarer. This entails that adaptations need to be identified ahead of time, and it is here that the first obstacle arises. Some of the adaptations require precisely the knowledge of Dolgan phonology in order to be recognized. For example, the *-dk-* (= [-tk-]) > *-kk-*change observed in *lokka* above, is not merely *an* assimilation but in fact an application of one of Dolgan phonotactic rules which are obligatory in native words across morpheme boundaries. These are also sometimes exercised for loanwords but this is a very rare case (only seven examples in the corpus of 1169 words). Hence the relatively high score of 0.5 although two out of three adaptations have not been applied here – a fully nativized shape would be **luokko* or **lōkko.*

Levenshtein distance is not bothered by the commonness of the given change. It measures the *phonetic* distance between two forms. Naturally, this requires precise phonetic transcriptions of both words in order to return valid results. This is the second obstacle. Detailed recordings are available for Dutch, English, German, Norwegian, etc. but are missing and much more difficult to obtain for lesser known and more distant languages such as Dolgan. What is more, extinct and reconstructed languages have to be automatically excluded, together with any borrowings which occurred from a dialectally mixed society, where the exact pronunciation is often impossible to establish. This happens to be the case with Russian in northeastern Siberia.

If one nevertheless decided to wade on, they would therefore find themselves forced to measure phonological rather than phonetic distance. This allows further investigation but it makes a significant difference.

One difficulty is to decide which phonemes can be treated as corresponding, and which cannot. *V* does not occur in Dolgan but in loanwords. If they were considered Fremdwörter, Dolg. *b* could correlate with both Russ. *b* and *v*. Such a solution might seem to be an exaggeration at first but its fabricated feel quickly wanes away as the number of obstacles of this type grows. This is the reason why I only provide approximate counts of examples below. An exact number would require many methodological decisions and discussing them would be beyond the scope of the current note.

On the other hand, a move to phonology brings the results closer to reality elsewhere. *K* and *ḱ* are allophonic in Russian in some positions, and so they are in Dolgan. In *vétka* '*Siber.* canoe', the *k* is not palatalized whereas in *bǟḱḱä* id. both *k*'s are, and in both cases this is not phonemic. Adopting a phonological transcription will improve the Levenshtein distance by freeing it from incorporating an irrelevant difference in its result.

I calculated the Levenshtein distance for the entire corpus of Russian loanwords in Dolgan (with *indel* = *sub* = 1) and contrasted the results with my index. The correlation turned out to be 0.43, and this hardly came as a surprise. First and foremost, the methodological approach is dramatically different. Let us consider a few cases:

- Both measures closely match
  This mostly happens when most of the possible adaptations have not been applied or when very few adaptations are applicable, and they were skipped. Both measures are 0 or draw near to it and thus they match or almost match.

  In the case of Russian loanwords in Dolgan, such examples account for less than a fourth of the total number.

  Examples: Russ. *patrón* 'cartridge' > Dolg. *patruon* id. (index 0.04667), *pártija* 'party' > *pārtija* id. (0.02864), *rabóčij* 'worker' > *rabočaj* id. (0.11023), *žurnál* 'journal' > *žurnāl* id. (0.00828); Russ. *čas* '1. hour; 2. clock', *kak* 'since', *maj* 'May', *šar* '*i.a.* balloon' > Dolg. ≡ (indices 0).

  A close match can also happen in other situations, in particular when the adaptations applied exhaust all the possibilities as completely as much they change the phonetic shape.

  However, such examples only account for less than an eleventh of the total number.

  Examples: Russ. *blagosloví* 'may he bless' > Dolg. *lastabi* id. (index 0.78626), *Fëdor* (given name) > *Pädär* id. (0.3666), *vóvse* 'completely' > *buosa* id. (0.44574), *zdoróvьe* 'health' > *dorōbuja* id. (0.55297).

- The two measures are almost opposite
  This mostly happens when there are very few adaptations possible and they have all been applied but without changing the word's phonetic shape much.

Such cases are very rare and only account for less than a twenty-seventh of the total number.

Examples: Russ. *Ánna* (given name) > Dolg. *Ānna* id., *barán* 'fur jacket' > *barān* id., *pop* 'Orthodox priest' > *puop* id., *ukázka* 'pointer' > *ukāska* id. (all indices 1).

- The two measures diverge randomly

  This happens when the degree in which the applied adaptations exhaust all the possibilities, does not coincide with their shape-changing power.

  One borderline case of this has already been mentioned above. *-dk-* > *-kk-* in *lódka* 'boat' is phonetically a minor assimilation but in Dolgan, it is a sign of far-reaching nativization. *Óčeredь* 'order, sequence' > *uočarat* id. (index 0.12155), on the other hand, is phonetically a considerable change but for Dolgan phonology, it is merely a combination of a fairly common substitution of a diphthong for a Russian accented vowel (unsurprisingly, especially often with *ó*), an even more common repair of vowel harmony, and an equally common removal of palatalization from *t́* since such sound does not exist in Dolgan, and *t* does.

  This is by far the most common case and it accounts for about two thirds of the total number.

  Examples: Russ. *krovátь* 'bed' > Dolg. *kyrbat* id. (index 0.99227), *Oksínьja* (given name) > *Oksiäńńä* id. (0.02864), *séjanka* (a kind of meat dish) > *hiäŋki* id. (0.2576), *Vasílьevič* (patronym) > *Bahylajbys* (0.99676).

To conclude, Levenshtein distance applied to Russian loanwords in Dolgan will return a valid and true measurement of phonological difference between the etymon and the loanword – but it will be a purely surface measure which may or may not correlate with actual human perception. More often the latter. The Levenshtein algorithm is quite flexible and probably can be refined so as to take note of those adaptations which are phonotactically trivial but phonetically devastating to the shape of the etymon, such as vowel harmony. Should this prove impossible, another algorithm can be used or a new one can be invented to perhaps achieve a full correlation with human perception. However, the crux is that it will always have to be based on the knowledge of the languages in question. This knowledge can be obscured by using a universal algorithm which itself learns from training data (e.g. Dunning 1994, Sanders, Chin 2009) but this does not change the essential need for such knowledge in general.

2. Other examples

Russian and Dolgan are most definitely not the only pair of languages where phonetic dissimilarity does not necessarily coincide with perception as distant forms.

When a word is borrowed into a language with a phonology considerably more restrictive than that of the donor language, it is prone to be heavily altered, and at the same time, the “new owners” are likely to not even realize that any change has taken place. The Japanese have rendered Engl. *drama* in two ways: [dorama] and [ʒ́urama] (Polivanov 1968: 237f.). A Polish asked to name a Hungarian author replied that some say [pätɔ·f́i] and others say [pätä·f́i], without at all being aware of missing the actual pronunciation of [pä̬·tȫf́i] by quite a distance. A German, after hearing the names [anja] and [ańa] repeated many times side by side, was only able to admit that there might perhaps indeed be something like a very slight difference between the two, at the very edge of recognition available to humans.

Stories of this sort will pop up every now and then at any party attended by linguists and they will gather the more applause the more exotic the phonologies in question are when compared to those of the listeners. This is to say, the audience will act much as the Levenshtein algorithm would. However, if they were told to a Japanese, a Polish or a German, chances are that the public would miss the punchline entirely. Knowledge of the relevant phonologies is key here to obtain the desired effect.

There is also the opposite case, when a loanword has not been significantly changed phonetically but in such a way that it raises associations with some other word in the borrowing language. Turkish *okul* ‘school’ is in fact a neologism and for every Turkish speaker its link to *oku-* ‘to learn’ is apparent. Polish dialectal *smentarz* ‘boneyard’ is a result of folk etymology by the dialectal shape *smętek* [-änt-] ‘sorrow’. (The literary form is *cmentarz* ≪ Lat. *cœmētērium* id. ≫ Engl. *cemetery*.) French *choucroute* ‘sauerkraut’ is a loan from Middle German *sūrkrūt* id. despite its shape which resembles more the French words *chou* ‘cabbage’ and *croûte* ‘crust’.

In these cases, the Levenshtein distance is invariably low but actual native speakers’ perception is often complicated. They will frequently admit a considerable phonetic similarity but at the same time refuse to connect the two words because lexical and semantic associations with other words they know in their own languages, are too strong.

3. Conclusion

Levenshtein distance, in its basic form and the more so in its highly refined versions, can be a good measure of *phonetic* distance. However, it needs to be remembered that phonetic distance is not necessarily equal to phonological distance, and that neither of them has to be equal to the perceived distance. The results will depart the further, the more and deeper differences there are between the phonological systems from which the compared words are taken. Levenshtein’s method can serve as a (very) good approximation for what has been its main domain of implementation so far, i.e. comparison of relatively similar dialects, but it fails when phonologically more distant languages are attempted to be analyzed. Its independence of the knowledge of language is sometimes raised as one of its most important strengths. In many cases, however, this will be the sole feature responsible for its failure.

Leonard Bloomfield is usually credited for the witty saying that *if you want to compare two languages, it helps to know one of them*. He died in 1949 and could not witness or perhaps even foresee the computer revolution of the last decades. But the technological advance does not obsolete his observation. It still helps, and it appears that it always will, if at least one actor of the comparison – be it the algorithm only – knows at least one of the languages being compared.

Abbreviations

Dolg. = Dolgan | **Engl.** = English | **Lat.** = Latin | **Russ.** = Russian

References

van der Ark R., Mennecier P., Nerbonne J., Manni F. 2007. Preliminary identification of language groups and loan words in Central Asia. – Osenova P. et al. (eds.) *Proceedings of the RANLP workshop on computational phonology workshop at the conference Recent Advances in Natural Language Processing*. Borovets: 13–20. [www.let.rug.nl/nerbonne/paper.html, accessed 2010.12.17].

Dunning T. 1994. *Statistical identification of language*. – *Technical Report CRL MCCS* 94-273. New Mexico State University. [ucrel.lancs.ac.uk/papers, accessed 2010.12.18].

Heeringa W., Kleiweg P., Gooskens Ch., Nerbonne J. 2006. Evaluation of string distance algorithms for dialectology. – Nerbonne J., Hinrichs E. (eds.) *Linguistic distances workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics*. Sydney: 51–62. [www.let.rug.nl/nerbonne/paper.html, accessed 2010.12.17].

Heggarty P. 2006. Interdisciplinary indiscipline? Can phylogenetic methods meaningfully be applied to language data — and to dating language? – Renfrew C., Forster P. (eds.) *Phylogenetic methods and the prehistory of languages*. Cambridge: 183–94.

Nerbonne J., Heeringa W. 2009. Measuring dialect differences. – Schmidt J.E., Auer P. (eds.) *Language and space: theories and methods* [= *Handbücher zur Sprach- und Kommunikationswissenschaft* 30.1]. Berlin: 550–67.

Polivanov E.D. 1968. *Statьi po obščemu jazykoznaniju*. Moskva.

Sanders N.C., Chin S.B. 2009. Phonological distance measures. – *Journal of Quantitative Linguistics* 16.1: 96–114. [citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.95.2447, accessed 2010.12.18].

Stachowski K. 2010. Quantifying phonetic adaptations of Russian loanwords in Dolgan. – *Studia Linguistica Universitatis Iagellonicae Cracoviensis* 127: 101–77.