Development and evaluation of phonological models for cognate identification

Bogdan Babych
Centre for Translation Studies
University of Leeds, UK
b.babych@leeds.ac.uk

Abstract

The paper presents a methodology for the development and task-based evaluation of phonological models, which improve the accuracy of cognate terminology identification, but may potentially be used for other applications, such as transliteration or improving character-based NMT. Terminology translation remains a bottleneck for MT, especially for under-resourced languages and domains, and automated identification of cognate terms addresses this problem. The proposed phonological models explicitly represent distinctive phonological features for each character, such as acoustic types (e.g., vowel/ consonant, voiced/ unvoiced/ sonant), place and manner of articulation (closed/open, front/back vowel; plosive, fricative, or labial, dental, glottal consonant). The advantage of such representations is that they explicate information about characters' internal structure rather than treat them as elementary atomic units of comparison, placing graphemes into a feature space that provides additional information about their articulatory (pronunciation-based) or acoustic (sound-based) distances and similarity. The article presents experimental results of using the proposed phonological models for extracting cognate terminology with the phonologically aware Levenshtein edit distance, which for Top-1 cognate ranking metric outperforms the baseline character-based Levenshtein by 16.5%. Project resources are released on:
https://github.com/bogdanbabych/cognates-phonology

© 2018 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.

1 Introduction: development of phonological models for cognate terminology identification

This paper presents a methodology for the development and automated evaluation of linguistic phonological features sets that can extend traditional methods of cognate terminology identification, such as Levenshtein edit distance.

Cognate identification is important for a range of applications. This paper evaluates its use for assisting MT developers in creating cognate term banks used in rule-based and hybrid MT, as well as in computer-assisted translation, development of dictionaries and between closely related languages (e.g., Ukrainian (Uk) and Russian (Ru), Portuguese (Pt) and Spanish (Es), Dutch (Nl) and German (De)). For many of such language pairs one of the languages can be under-resourced, therefore no electronic dictionaries are available, and only small parallel corpora with limited lexical coverage can be collected. Typically these parallel corpora can provide translations for frequently used general words, but miss the 'long tail' of less frequent, often topic-specific or terminological words. However, in closely related languages these words are often cognates, which creates a possibility to rapidly extend bilingual lexicons in semi-automated way using non-parallel, comparable corpora and automated cognate identification techniques. In this task, cognate candidates are generated from word lists created from large monolingual comparable corpora in both languages The assumption is that the developers have good linguistic intuition of both languages and work through lists of cognate candidates, checking which pairs can be added to the bilingual dictionary. Their productivity depends on whether cognates are presented high up in the list of candidates, ideally at the top of the list, or

Pérez-Ortiz, Sánchez-Martínez, Esplà-Gomis, Popović, Rico, Martins, Van den Bogaert, Forcada (eds.)
Proceedings of the 21st Annual Conference of the European Association for Machine Translation, p. 41–48
Alacant, Spain, May 2018.

at least in the top N items, where N should be relatively small, e.g., the number of lines which fit on a single screen.

Other uses of cognates for terminology identification include term extraction from parallel corpora. If multiword source terms are known, the task is to identify the boundaries of the corresponding multiword target terms in the aligned target sentences, where component words or stems of compound words within the target terms may not be necessarily cognate with the corresponding source, so correctly identified cognates can facilitate adding adjacent non-cognate words according to part-of-speech and word order patterns, e.g., En: *'**information** requirements'* ~ Uk: *'**інформаційні** потреби' ('**informatsijni** potreby')*; or splitting and extending compounds which have cognate parts, e.g., En: 'multi**national**' ~ Uk: 'багато**національний**'-('bahato**natsionalnyj**').

Yet another application of cognate identification is sentence alignment of parallel corpora, where statistical alignment methods are more accurate if cognates are used as an additional data source (Lamraoui and Langlais, 2013:2). Inaccuracies in cognate identification, which are due to orthographic differences, often create unnecessary bottlenecks for this task (Varga et al., 2015: 249). In this scenario identified cognates are not necessarily terms, but they contribute to a more accurate alignment and extraction of non-cognate terminology, produced from word alignment and monolingual terminology detection.

An additional complication for the multilingual terminology extraction scenarios that rely on cognate identification is the use of different writing systems in the source and target (e.g., Cyrillic or Georgian vs. Latin script), which requires transliteration between those languages.

Transliteration is often non-trivial, because of differences in pronounciation of the same letters, the lack of direct graphemic equivalents across languages, contextual dependencies in transliteration rules, different historical conventions for different words (e.g., En/De "h" → Ru "х" (*hockey ~ хоккей,* since borrowed directly from En), or "г" (*hermeneutics ~ герменевтика*, since borrowed via Ukrainian, where En: h → Uk: г [ɣ] → Ru: г [g]). Also, even if languages use the same alphabet, pronounciation of letters and corresponding transliteration rules may differ (e.g., Cyrillic letter "и" = [i] in Ru and [y] in Uk, Latin letter "g" = [g] in En/De, and [ɣ] in Nl), so new transliteration mappings need to be created for each translation direction, each with their potential language-specific problems.

As a result, the complexity of transliteration in some cases is comparable to the complexity of MT, and it is often addressed not via simple character mappings, but via fully developed character-based MT models that require an aligned training corpus for each translation direction, and which are used in MT applications to cover out-of-vocabulary words, such as compounds, morphologically complex words, named entities and cognate terminology (Senrich et al., 2016: 1716)

Transliteration problem resembles a traditional "direct translation" bottleneck in MT: this approach cannot reuse any of the previously created mappings between languages if a new language pair or translation direction need to be covered. A more principled approach to the transliteration problem in the context of automated cognate identification, developed in this paper, is mapping characters for each language into a language-independent ("interlingual") phonological feature space.

2 Related work

The use of phonological features for cognate identification has been initially proposed in the context of dialectological studies (Nerbonne & Heeringa, 1997) and diachronic phonology (Kondrak, 2000: 288), (Kondrak, 2009). Some limitations of these approaches for MT-related tasks have been discussed in (Babych, 2016), such as the need for phonological transcription of orthographic words and the absence of reliable evaluation for different ways of organising the complex phonological feature space and computing similarity between phonological segments. For instance (Kondrak, 2000: 290-293) acknowledges that different phonological features make unequal contribution in computing similarity between segments. To address this problem, in the ALINE phonetic aligner an introspective set of weights for each of the features is adopted from (Ladefoged, 1995). Machine-learning algorithms based on learning phonetic mappings from bilingual texts (Kondrak, and Sherif, 2006) outperform the introspective linguistic model based on weighted phonological features.

However, the most important difference between identification of cognates for dialectological or historical studies of language vs. for MT-

oriented tasks of cognate term identification is the range of the compared candidate cognates and therefore the need of the metric to be optimised for both recall and precision on the large dictionary data sets. Addition of phonological features on such tasks often results in overgeneration, so additional features have to be used, such as semantic similarity of terms, WordNet-based and semantic features, clustering (Kondrak, 2009, St Arnaud et al., 2017).

On the large scale for cognate identification for MT, where datasets are not limited only to candidate cognate pairs, a character-based Levenshtein edit distance (Levenshtein, 1966) is typically used, without additional linguistic features. Levenshtein metric calculates the number of insertions, deletions and substitutions between compared word pairs from different languages and determines if they pass a threshold to be considered cognate candidates. For example, if cognate candidates are extracted from a non-aligned or non-parallel corpus, the Levenshtein distance is computed for every pair of words in the two word lists created for each language (the Cartesian product of the lists), the search space may be restricted to comparing words with the same part-of-speech (PoS) codes, if PoS annotation is available for the corpus.

However the problem with the character-based Levenshtein metric is that all characters in comparison are treated as atomic units that do not have any internal structure and therefore, can be substituted only as a whole character. Because of this the Levenshtein metric does not distinguish between the substitutions of characters that correspond to acoustically/articulatory similar sounds vs. the substitution of phonologically distant letters. As a result, words that are intuitively close may receive a large distance score, e.g.,

Uk "*жовтий*" (*zhovtyj*)='yellow'
Ru "*жёлтый*" (*zheltyj*) = 'yellow'
(Lev distance = 3),

where, for historical reasons, articulatory similar sounds are represented by different characters: the sound [o] – by 'о' in Uk and 'ё' in Ru, the sound [y] – by 'и' in Uk and 'ы' in Ru. On the other hand, words that are not cognates and are phonologically and intuitively far apart, still receive the same distance scores, such as:

Uk "*жовтий*" (*zhovtyj*) = 'yellow' and
Ru "*жуткий*" (*zhutkij*) = 'dismal' (Lev = 3).

For example, here no distinction is made between, on the one hand, the substitution "*о*" (o) → "*ё*" ([i]o) of phonologically similar sounds (which differ only in a peripheral feature – triggering palatalization of the preceding consonant (Uk: -- Ru: +; in addition, this feature is neutralised after the sibilant "*ж*" (zh)), and on the other hand – the substitution "*о*" (o) → "*у*" (u), where sounds differ in core articulatory features of the place of vowel articulation (Uk: middle; Ru: close/high).

Some existing modifications and extensions of the Levenshtein metric introduce weightings for different character mapping, but these weights need to be set or empirically determined for each specific mapping: compared characters still do not have internal structure and there is no way to predict the weights in advance for any possible pair in a principled way.

This paper presents an automated task-based evaluation framework for an extension to the Levenshtein edit distance metric, which explicitly represents linguistic phonological features of compared characters, so the metric can use information about characters' internal feature structure rather than treat them as elementary atomic units of comparison. Similar sets of distinctive features have been used for comparing transcriptions of spoken words in modeling dialectological variation and historical changes in languages (Nerbonne and Heeringa, 1997). In the proposed approach, phonological feature representations are applied to cognate identification and terminology extraction tasks, transliteration, and as well as modeling morphological variation. Previously it has been shown that there are multiple ways of identifying, representing, structurally arranging and comparing these features in a phonological feature space (Babych, 2016), so there is a need for a methodology for evaluating alternative feature configurations. The results of the previously reported pilot experiment, using a small-scale manual evaluation, indicated the need to use hierarchical phonological feature structures for consonants rather than flat feature vectors previously used in dialectological research.

Manual evaluation methods in previous pilot experiment cannot be used for systematically testing and optimising weights or alternative phonological feature representations used in the Levenshtein phonological metric.

For instance, a serious problem for the proposed phonologically aware metric has been overestimation of its insertion and deletion costs, which is mainly due to the relatively smaller average substitution cost, and no corresponding reduction in the average insertion or deletion costs. E.g., for non-cognates a replacement of a

consonant with another phonologically unrelated consonant produces a substitution distance of 0.8, because one feature – "*type:consonant*" does not have to be rewritten (phonological structure of consonants in the proposed models has 5 features). If insertion and deletion costs remain =1, this leads to disproportional under-generation of cognates that contain inserted or deleted characters. Even though the need of adjusting insertion/deletion distances has been highlighted in the pilot stage, manual evaluation methods used then did not allow us to test and optimise multiple parameter settings for the phonological metric, such as a range of different insertion and deletion costs. Their values have to be determined experimentally using an automated evaluation methodology.

This paper develops an automated framework for evaluating different arrangements of phonological features and parameters using the task of cognate identification, which enables us to experimentally find optimal setup of a metric for a given task. Apart from practical applications mentioned above, this methodology creates a framework for feature engineering for phonologically aware character-based models for a wider range of machine translation and machine learning methods and tools, to design and calibrate phonological feature structures in a systematic way tuned for optimal the performance on specific tasks.

The proposed automated evaluation framework uses standard automatically computed evaluation metrics, such as number of cognates in top-N candidates and an average rank of a correct cognate in an ordered candidate list. Evaluation is performed on a larger data set of candidate cognate lists generated from large Ukrainian and Russian corpora on a high-performance computing cluster. The evaluation results show the settings where phonological Levenshtein metrics achieves best performance on the cognate identification task and allow us to rule out some unproductive modifications.

3 Phonological distinctive features and their application for cognate identification

A theory of phonological distinctive features, which was first proposed by Roman Jacobson (Jakobson and Halle, 1956: 46; Anderson, 1995: 116), associates each phoneme (an elementary segmental unit of speech that distinguishes meanings and is intentionally produced by speakers) with its unique set of values for categories, which apply to classes of sounds. For example, the phoneme [t] has the following values for its associated phonological categories:

'type': consonant
'voice': unvoiced
'maner of articulation': plosive;
'active articulation organ': front of the tongue
'passive articulation organ': alveolar

Phoneme [d] has the same set of articulatory features apart from one: it is pronounced with vocal cords vibrating, while organs and manner of articulation remain the same, so it differs only in the value of one distinctive feature,

'voice': voiced.

In historical development of languages and in morphological variation within a language the phonological changes more often apply only to values of certain distinctive features within characters, but much less often extend to the whole category-value system, e.g.: De: "*Tag*" = Nl "*dag*" ('day); De: "*machen*" = Nl "*maken*" ('make'). Therefore, for languages where the writing system is at least partially motivated by pronunciation, for certain character based models, e.g., modelling morphological variation or cognates in different languages, it would be useful to represent phonological distinctive features of characters, in order to differentiate between varying degrees of closeness for their different classes, e.g., vowels, sonants and consonants, or sounds with identical or similar articulation. Greater closeness between characters in terms of their phonological features has important linguistic and technical applications, such as modelling dialectal variation, historical change, morphological and derivational changes in words, e.g., stem alternations in inflected forms.

(1) In past research (Babych, 2016: 123) phonological distinctive features have been integrated into the Levenshtein distance metric in the following way: e.g., to substitute [t] with [d] in Nl: "tag" → De: "dag" there is a need to re-write only one feature out of 5, so the distance is 0.2 rather than 1. However, in the general case different classes of characters use different numbers of features, so substitution distance *Subst* is calculated as:

$$Subst = 1 - F\text{-}score,$$

where F-measure is the harmonic mean of Precision and Recall of the overlap between sets of their phonological features. This allows the metric to calculate the distance for characters with different numbers of features remaining symmetric.

(2) The order of matching the distinctive features was found to be important. The experiment described in Section 4 compares two different arrangements of features: as flat feature vectors and as feature hierarchies. In the hierarchies the higher level features need to be matched as a pre-condition for attempting to match lower level features. Hierarchical organization consistently achieves better performance compared to flat feature vectors. Intuitively this means that not all feature categories should be treated equally; some are more central, have higher priority, and license comparison of lower level features on the periphery of the phonological feature system.

(3) Insertion and deletion costs have been calibrated for the range between 0.2 and 1 using the proposed evaluation framework, described in this paper in Section 4. Optimal performance on cognate identification was achieved for cost of insertion = deletion = 0.8.

For the task of cognate identification, the introduction of these features distinguishes different types of character substitutions and gives more accurate prediction of the degree of closeness between compared characters and words, e.g., for the word pairs discussed above, where the baseline Levenshtein distance =3 for both (matching features, which do not need to be rewritten, are highlighted in bold):

Graphemic-Phonological (graphonological) feature Uk: "*жовтий*" (*zhovtyj*) = 'yellow'

ж (zh) 'type:consonant', 'voice:ff-voiced', 'maner:ff-fricative', 'active:ff-fronttongue', 'passive:ff-palatal'
о (o) **'type:vowel'**, **'backness:back',** **'height:mid'**, **'roundedness:rounded'**, 'palate:nonpalatalizing'
в (v) **'type:consonant'**, 'voice:fl-voiced', 'maner:fl-fricative', 'active:fl-labial', 'passive:fl-bilabial'
т (t) 'type:consonant', 'voice:pf-unvoiced', 'maner:pf-plosive', 'active:pf-fronttongue', 'passive:pf-alveolar'
и (y) **'type:vowel'**, 'backness:front', **'height:closemid'**,**'roundedness:unrounded'**, **'palate:nonpalatalizing'**
й (j) 'type:consonant', 'voice:xm-sonorant', 'maner:xm-approximant','active:xm-midtongue', 'passive:am-palatal'

Feature representations for corresponding characters in Ru: "*жёлтый*" (*zheltyj*) = 'yellow'.

ё (io) **'type:vowel', 'backness:back', 'height:mid', 'roundedness:rounded'**, 'palate:palatalizing'
л (l) **'type:consonant'**, 'voice:lf-sonorant', ' maner:lf-lateral', 'active:lf-fronttongue', 'passive:lf-alveolar'
…
ы (y) **'type:vowel'**, 'backness:central', **'height:closemid','roundedness:unrounded', 'palate:nonpalatalizing'**

It can be seen from the examples above, why for the task of cognate identification it is important that character substitution in the graphonological Levenshtein metric only touches some distinctive feature in a characters' feature sets. Such feature substitution at the sub-character level still unambiguously changes one character into another, since there is a one-to-one correspondence between a new set of phonological features and the corresponding sound or character: according to Jacobson's distinctive features model (implemented in the proposed phonological representations), there cannot be two sounds in a language that share exactly the same set of values for their phonological categories.

If only some sub-character features are changed, the substitution cost is < 1, and normally reflects the proportion of phonological features which need to be rewritten.

Calculation of the Graphonological Levenshtein metric for Uk "*жовтий*" (*zhovtyj*) = 'yellow' and (Ru) "жёлтый" (zheltyi) = 'yellow':

0.0	1.0	2.0	3.0	4.0	5.0	6.0
1.0	**0.0**	1.0	2.0	3.0	4.0	5.0
2.0	1.0	**0.2**	1.2	2.2	3.2	4.2
3.0	2.0	1.2	**1.0**	2.0	3.0	4.0
4.0	3.0	2.2	2.0	**1.0**	2.0	3.0
5.0	4.0	3.2	3.0	2.0	**1.2**	2.2
6.0	5.0	4.2	4.0	3.0	2.2	**1.2**

cf.: Metric calculated for Uk "*жовтий*" (*zhovtyj*) = 'yellow' with Ru "*жутκий*" (*zhutkij*) 'dismal':

0.0	1.0	2.0	3.0	4.0	5.0	6.0
1.0	**0.0**	1.0	2.0	3.0	4.0	5.0
2.0	1.0	**0.2**	1.2	2.2	3.2	4.2
3.0	2.0	1.2	**1.0**	1.2	2.2	3.2
4.0	3.0	2.2	2.0	**1.8**	2.2	3.0
5.0	4.0	3.2	3.0	2.8	**2.0**	3.0
6.0	5.0	4.2	4.0	3.8	3.0	**2.0**

While the baseline Levenshtein distance Lev=2 for both pairs shown above, the phonolog-

ically-aware distance, GLev = 2.0 for non-cognates, which is > 1.2 for cognates.

An additional advantage of using of phonological feature representations for graphemes is a more natural "interlingual" transliteration between different scripts and languages. The phonological models, presented in this paper, map characters from any given language into a universal space of acoustic and articulatory phonological features, which is independent of any specific writing system or a language-pair. This space can be seen as a phonological "interlingua", which shares some advantages with the idea of interlingual MT: graphonological mappings enable implicit cross-lingual transliteration, where mappings from individual languages into the common phonological feature space can be reused when new translation directions are added.

4 Set-up and results of the evaluation experiment

This section presents a methodology for automated performance-based evaluation that is used in testing different settings of phonological categories and values for the extended Levenshtein metric. The experiment is set up in the following way:

(1) Small freely available electronic dictionaries for Ukrainian–Russian and Russian–Ukrainian directions were used to develop a gold-standard translation glossary of 11000 Ukrainian words, each having one or more Russian translation equivalents. All source words and their translation equivalents were used as they appear in the dictionaries (for the Russian–Ukrainian dictionary the translation direction was reversed and the translation equivalents missing from the original Ukrainian–Russian list were added to it. Cognates were not specifically selected or annotated in any way, so the gold standard evaluation set represented a standard introductory size bilingual glossary, such that similar resources could be found or compiled for many other language pairs.

(2) For identification of cognates two large monolingual corpora of Ukrainian and Russian news were used (250 million words each) with a standard morphological annotation of parts-of-speech (PoS) and lemmas. For each language frequency lists of lemmas and PoS codes were generated from these morphologically annotated corpora. After this the source and target words from the Ukrainian—Russian glossary have been intersected with the Russian and Ukrainian word lists compiled from PoS-tagged corpora for corresponding languages. The resulting Ukrainian evaluation set with corresponding gold-standard Russian dictionary equivalents included only those entries that were found both on the source and target sides in the glossary and both in the Ukrainian and Russian monolingual word lists. As a result, the evaluation set contained only the entries that could in principle be found by the cognate identification tool in the word lists and evaluated using the glossary.

(3) An additional requirement has been introduced that in both word lists the cognates should be tagged with the same part-of-speech. This reduces the search space for cognates and computing time needed to calculate phonological Levenshtein distances.

(4) Candidate cognate lists were generated for 809 randomly selected entries from the Ukrainian evaluation set in the following way. For each Ukrainian word in the evaluation set different variants of the Levenshtein edit distances were calculated to each word in the large Russian monolingual word list from the news corpus (around 106.000 unique lemmas, further filtered by their of speech codes). This process is computationally intensive and required parallel processing of the Ukrainian test entries on a high-performance computing cluster. Even though calculation of the baseline traditional Levenshtein distance is relatively fast, calculation of the phonological variant of this metric is much more computationally demanding, as it requires generating and comparing phonological feature sets for each of the compared characters in a large number of strings. For the current implementation, sequential generation of the phonological Levenshtein edit distances between a test Ukrainian entry and each of the 106.000 entries in the Russian monolingual word list takes about 4 minutes of computing time (54 hours of sequential computation for the whole evaluation set of 809 Ukrainian words).

In future, for the task of a large-scale induction of cognates between languages phonological feature representations will be optimised for speed and other techniques such as hashing of phonological features for the searched target entries will be implemented, which is expected to make the developed metric more usable for generation of wide-coverage translation resources.

(5) Candidate cognate lists were ranked according to distance scores produced by the fol-

lowing edit distance metrics: the Baseline Levenshtein edit distance, the phonological Levenshtein distance that used flat feature vectors, and by five variants of the phonological Levenshtein distance metric that used hierarchical phonological feature representations and one of the five possible weights for insertions/deletions: 0.2, 0.4, 0.6, 0.8 and 1.

For each Ukrainian word from the evaluation set, its Russian translation equivalents from the gold standard dictionary translations were automatically searched in the ranked cognate lists generated for that word by different variants of the Levenshtein metric. The position of the top dictionary translation equivalent was recorded in each of the ranked cognate lists.

(6) Even though dictionary equivalents were not necessarily cognates in the evaluation set, the experiment produced meaningful results, because non-cognate equivalents were simply not found and disregarded for the consideration. In this way the experimental set-up automatically focussed on the quality of cognate identification. Importantly, this allows us to avoid expensive manual selection or annotation of cognates: as the evolution methodology is automatic, all translation equivalents available in the gold standard are treated equally: in this stage no distinction is made between cognates and non-cognate equivalents. This removes the need for the manual filtering of the gold standard and also naturally covers 'near-cognates' or words with cognate morphemes where only parts of words match. Since the baseline and the modified Levenshtein metric are evaluated on the same gold standard, performance figures are relative and show the difference in finding translation equivalents for any degree of 'cognateness'.

(7) Different variants of the metric are compared by the following parameters: Median top-N number for the metric; In top-1, top-5, top-10 and top-25.

(8) The following settings were compared:

(a) Baseline Levenshtein edit distance;

(b) Levenshtein distance extended with phonological features with flat feature vectors;

(c) Levenshtein distance extended with hierarchical phonological features (where manner and active place of articulation are treated as top-level features, which need to be matched in order for other features to match;

(d) Variants of the (b) and (c) metric with different insertion / deletion values – between 0.2 and 0.8.

The results of the evaluation experiment are presented in Table 1, where:

BaseL Lev = baseline Levenshtein metric

Phon Lev H = Phonological extension to Levenshtein metric with feature hierarchy

Phon Lev V = Phonological extension to Levenshtein metric with flat feature vectors

PhonLevi=0.X = Phonological extension to Levenshtein metric with modified insertion / deletion cost: i0.2 = the cost of insertion deletion is set to 0.2, i0.8 = is set to 0.8 (it is set to 1 in the Phon Lev metrics.

5 Discussion of the results, conclusion

It can be seen from Table 1 that:

(1) Hierarchical phonological Levenshtein metric outperforms the baseline on the Top 1 and Top 2 measures, the median rank improvements is +5%

(2) Flat phonological feature vector metric on all measures performs worse than the baseline. This can be interpreted as the need to take into account the order of matching higher-level features. Match of low-level features is not meaningful if higher-level features are not matched.

(3) The Hierarchical metric with insertion / deletion cost set to 0.8 outperforms both the baseline and the Levenshtein metric with the insertion/deletion cost = 1, especially on the Median

Experiment	**Median topN**	**Top 1**	**Top 5**	**Top 10**	**Top 25**
BaseL Lev	50	206	328	360	382
Phon Lev V	87.5	**215**	289	319	349
DiffBase L	-75%	+4.4%	-10%	-11%	-9%
PhLev Hierarchy:					
PhLev i=0.2	125.5	216	291	315	342
PhLev i=0.4	54.5	230	307	334	367
PhLev i=0.6	*48*	235	328	354	385
PhLev i=0.8	**40**	**240**	**337**	**359**	**391**
Ph Lev i=1.0	47.5	**240**	334	359	385
Best BaseL Improv	**+20%**	**+16.5%**	**+3%**	**0%**	**2%**

Table 1: Automated evaluation of metric settings.

Top N, Top 1 and Top5 measures. This can be interpreted as the need to scale down insertion cost moderately, since the average substitution cost is down.

The results show that phonological extension to the Levenshtein edit distance metric on the task of cognate identification outperforms the character-based baseline. The proposed frame-

work also allows accurate calibration of the feature arrangement and other parameter settings of the metric.

The modified Levenshtein metrics, phonological features sets for several alphabets and sample input files are released as an open-source software on the github repository (Babych, 2018).

Future work will include systematic evaluation of different possible feature hierarchies and costs, and metrics application to other tasks, such as transliteration.

References

Anderson, Stephen R. 1985. *Phonology in the twentieth century: Theories of rules and theories of representations*. University of Chicago Press.

Babych, Bogdan. 2016. Graphonological Levenshtein Edit Distance: Application for Automated Cognate Identification. Baltic Journal of Modern Computing 4.2 (2016): 115.

Babych, Bogdan. 2018. Phonological models for cognate terminology identification. GitHub repository, https://github.com/bogdanbabych/cognates-phonology

Jakobson, Roman, and Moris Halle. 1956. *Fundamentals of language.* Vol. 1. Walter de Gruyter. URL: http://pubman.mpdl.mpg.de/pubman/item/escidoc:2350620/component/escidoc:2350619/Jakobson_Halle_1956_fundamentals.pdf

Kondrak, Grzegorz. 2000. A new algorithm for the alignment of phonetic sequences. In Proceedings of NAACL 2000, pages 288–295.

Kondrak, Grzegorz 2009. Identification of cognates and recurrent sound correspondences in word lists. Traitement automatique des langues et langues anciennes, 50(2):201–235.

Kondrak, Grzegorz, and Tarek Sherif. 2006. Evaluation of several phonetic similarity algorithms on the task of cognate identification. In Proceedings of the Workshop on Linguistic Distances. Association for Computational Linguistics

Ladefoged, Peter. 1995. A Course in Phonetics. New York: Harcourt Brace Jovanovich

Lamraoui, Fethi, and Philippe Langlais. 2013. Yet another fast, robust and open source sentence aligner. time to reconsider sentence alignment. *XIV Machine Translation Summit*. URL: http://rali.iro.umontreal.ca/rali/sites/default/files/publis/MTSummit-2013-Fethi.pdf

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady.* Vol. 10. No. 8.

Nerbonne, John, and Wilbert Heeringa. 1997. Measuring dialect distance phonetically. *Computational Phonology: Third Meeting of the ACL Special Interest Group in Computational Phonology.*

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers). Vol. 1. URL: http://www.aclweb.org/anthology/P16-1162

St Arnaud, Adam, David Beck, and Grzegorz Kondrak. 2017. Identifying Cognate Sets Across Dictionaries of Related Languages. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Varga, Dániel, Peter Hal´acsy, Andras Kornai, Viktor Nagy, Laszl´o N´emeth and Viktor Tron. 2007. Parallel corpora for medium density languages. *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4 292*: pp. 247-253. URL: http://eprints.sztaki.hu/7902/1/Kornai_1762382_ny.pdf