Later , many researchers extend and improve the SOUNDEX algorithm . For example [39] combine the letter - grouping idea of SOUNDEX with the mini - mum edit distance (MED) measure and describe algorithm called EDITEX which assigns smaller weight for replacing letters belonging to the same group .

In our modified minimum edit distance ratio algorithm (MMEDR) [30] we also assign smaller weights to the transformations that replace phonetically s imilar letters .

Guy [1 1] describes an algorithm for identification of cognates in bilingual wordlists based on the recurrent sound correspondences . It est imates the pro - babilities of phoneme correspondences by using a variant of chi - square stat ist ic on a contingency table , which holds how often given two phonemes co - occur in words of the same meaning . The algorithm used only s imple on - to - one phoneme correspondences .

The ALINE algorithm ([1 7]) is capable to identify phonetic s imilarity between a pair of phonetically transcribed words . It assigns a s imilarity score to pairs of transcribed words by decomposing phonemes into e lementary phonetic features , such as place of articulation , manner of art iculat ion , voice , etc . Features are assigned a weight based on their relat ive importance . Feature values are encoded as numbers between 0 and 1 . The s imilarity score i s then computed by a dynamic - programming algorithm that finds the optimal sequence of operations insert / delete , substitute , and expand / compress .

Following these ideas of Kondrak [1 9] also used sound correspondence to identify cognates between languages . His algorithm was init ially designed for extracting non - compositional compounds from bitexts but it is also able to find complex sound correspondences in bilingual wordlists , not j ust s imple on - to - one phoneme correspondences .

Phonetic s imilarity can be measured on the basis of the phonetic tran - scription of the words : first the words are transcribed as a sequence of sounds represented by characters and then the orthographic s imilarity between these sequences is measured . Transcription allows measuring phonetic s imilarity between languages using different alphabets . For example in our modified minimum edit distance rat io algorithm (MMEDR) [30] we perform phonetic transcription to replace Russian letters with their Bulgarian equivalents .

Kondrak and Dorr [1 6] combine several phonetic and orthographic approa - ches in their study of the identification of confusable drug names and report high accuracy . They conclude that a s imple average of several orthographic s imilarity measures outperforms all individual measures on the task of the identification of confusable drug names .

Manual $T-r$ ansformation Rules . Rather than applying directly some string s imilarity measure l ike MEDR or LCSR some studies first apply a set of transformation rules that reflect some typical cross - l ingual transformation pat - terns observed for given pair of languages . This is absolutely necessary when the languages do not use exactly the same alphabet which requires some letters from the first language to be replaced with letters from the second . This idea can be further developed to replace not j ust s ingle letters but also syllables , endings and prefixes .

For example in [30] we apply a set of manually constructed transformation rules for replacing Bulgarian with Russian endings , replace double consonants with s ingle and replace Russian - specific letters with their Bulgarian equivalents . After that we use a modification of MEDR algorithm that assigns weights for the replace operations reflecting some regular phonetic changes between Bulgarian and Russian .

Manually constructed transformation rules between English and German words (l ike replacing the letters k and z by c and changing the ending - t $\ddot{a}$ t by - ty)
are exploited also by [1 5] for expanding a l ist of cognates .

Learning $T-r$ ansformation Rules . The idea of learning automatically cross - l ingual transformation rules that reflect the regular phonetic changes between a pair of languages has been exploited by number of researchers . Such techniques follow naturally the idea of using manually constructed transformation rules .

Tiedemann [36] used various measures to learn the regular spelling trans - formations between English and Swedish from a set of known cognate pairs . His best performing string s imilarity measure algorithm NMmap uses LCSR algorithm to identify the non - matching parts of two strings and statist ically assigns weights corresponding to the probability for transforming between them .

The algorithm proposed by Mulloni and Pekar [2 7] extracts automatically from a l i st of known cognates a set of rules that capture regularities in the orthographic transformations between given two languages . These transformations are substitutions of a sequence of letters from the first language with a sequence of letters in the second language identified through the minimum edit distance algorithm . Special characters are added at the word boundaries to allow capturing of rules that transform the start , the middle and the end of the words . For each rule chi - square statistics is calculated and most regular rules are truncated and used while the others are ignored . Finally the transformation rules are applied as a preprocessing step and after that the normalized minimum edit distance is calculated as a s imilarity measure .

Mitkov et al . [26] use very s imilar methodology . They collect and score the transformation rules the same way l ike Mulloni and Pekar [2 7] but do not account word boundaries as special case . Once the rules are collected and scored by chi - square statistics they apply the rules on candidate pair of words and use LCSR to calculate their s imilarity .

All of the above techniques use positive examples of cognate pairs to learn regular transformation rules . Unlike them Bergsma and Kondrak [2] use positive and negative examples of cognate pairs to learn positive or negative weights on substring pairings in order to better identify related substring transformations . Starting from minimum edit distance they obtain an alignment of the letters in the given strings and extract corresponding substrings consistent with the alignment . Finally a support vector machine (SVM) i s trained by using set s of positive and negative cognate examples and the SVM is used to discriminatively classify given two words as cognates or not .

2 . 2 . Statistical Approach for False $F - r$ iends Identification . There is no much research concerning extracting false fr iends directly from text corpora .

Most methods (l ike [26] and [32]) first extract cognates and false fr iends candidates using some measure of orthographical or phonetic s imilarity and later try to dist inguish between true cognates and false fr iends .

Fung [10] proposes methods for creating bilingual lexicons from parallel corpora and comparable corpora . His method for extracting semantically related words from sentence level aligned parallel corpus works as follows : for each word pair two binary occurrence vectors are constructed . The first vector maps the occurrences of the first word in the sentences at the left s ide of the parallel text . The second vector maps the occurrences of the second word in the sentences at the r ight s ide of the parallel text . Finally the correlation between these vectors is calculated and used as measure for semantic relatedness .

Brew and McKelvie [5] use sentence alignment to extract cognates and false friends directly from parallel bilingual corpora . The semantic relatedness is identified by statistical method based on collocation analysis in the aligned sentences . The orthographic s imilarity i s measured by various string s imilarity algorithms . As a result the extracted candidate pairs are classified as cognates , translations , false friends , or unrelated . Their experiments are l imited to verbs in English and French but their approach i s capable to be applied for other languages as well .

Nakov and Pacovski [29] extract false friends directly from a parallel corpus . Their idea follows the intuit ion that false fr iends are unlikely to co - occur

Kondrak [20] extended his algorithm for measuring semantic s imilarity based on WordNet and used e ight semantic s imilarity levels as binary features : gloss identity , keyword identity , gloss synonymy , keyword synonymy , gloss hy - pernymy , keyword hypernymy , gloss meronymy and keyword meronymy . These features are combined with a feature based on phonetic s imilarity and naive Bayes classifier i s used to distinguish between cognates and non - cognates .

Mitkov et al . [26] proposed few methods for measuring semantic s imilarity between orthographically s imilar pairs of words used to dist inguish between cog - nates and false friends on the basis of s imilarity threshold estimated on a training data set . Their first method uses comparable corpora and relies on the distri - butional s imilarity . For given pair of words a set of N most s imilar words are collected using skew divergence [2 1] as s imilarity function . The s imilarity between the words i s calculated as Dice coefficient between the obtained set s . A bilingual glossary i s used to check if two words can be translations of each other . Their second method extracts co - occurrence statistics for each word of interest from the respective monolingual corpus using a dependency parser . Thus verbs are used as distributional features of the nouns . Semantic vectors are created for the two sets of verbs (using skew divergence again) and s imilarity between them i s measured by Dice coefficient and using a bilingual glossary . The first method requires a glossary of equivalent nouns while the second requires a glossary of equivalent verbs . In the same study the first method is further extended to use taxonomy data from EuroWordNet (when available) . The proposed methods are shown to have different performance on different language pairs and none of them was superior to the others .

The idea of using the Web as a corpus has been exploited by many scientists working on different problems (see [14] for an overview) . Some of them use Web search engines for finding how many t imes a word or phrase i s met on the Web and extracting pointwise mutual information ([13]) , whereas others directly retrieve context from the text snippets returned by the Web search engines ([3 1]) .

The idea of retrieving information from the text snippets returned by Web search engines is used in [6] . The model they introduce i s based on the idea that if two words X and Y are semantically bound , then searching for X should cause Y to appear often in the results , and vice versa : searching for Y should cause X to appear often in the results . As it is later discovered by Bollegala et al . [4] , this produces incorrect zero semantic s imilarity for most of the processed pairs .

Bollegala et al . [4] combine retrieval of information about the number of occurrences of two words (both together and individually) from a Web search

Lemmatization Lexicons . We used two large monolingual morphologi - cal lexicons for lemmatization for Bulgarian and Russian .

The Bulgarian morphological lexicon [33] is created at the Linguistic Modeling Department of the Inst itute for Parallel Processing in the Bulgarian Academy of Sciences (BAS) and contains about 1 0 0 wordforms and 70 0 lemmata . Each lexicon entry consists of a wordform , a corresponding lemma , followed by morphological and grammatical information . There can be multiple entries for the same wordform , in case of multiple homographs .

The Russian morphological lexicon [33] is also created at the Linguistic Modeling Department of the Inst itute for Parallel Processing in the Bulgarian Academy of Sciences (BAS) . It is in the same format l ike the Bulgarian and contains about 1 500 0 wordforms and 1 0 0 lemmata . It s core content is based on the grammatical dictionary of [38] .

Bilingual Glossary . We used a large Bulgarian - Russian e lectronic glos - sary consisting of 59 582 pairs of words which are translations of each other . The glossary was adopted by scanning , parsing and processing the Bulgarian - Russian dict ionary of [3] and the Russian - Bulgarian dictionary of [7] . We use the word - word translations from these dict ionaries ignoring the phrase - word and phrase - phrase translations . Most of the words have multiple translations so we have a set of Russian translation words for each Bulgarian word and vice versa . This i s taken into account during the comparison of the Bulgarian and Russian contextual semantic vectors as described in Section 3 . 4 .

Se $a - r$ ches in Google . During our experiments we performed searches in Google for 55 7 Bulgarian and 550 Russian wordforms and collected as many as possible (up to 1 0) page t it les and text snippets from the search results . We used this text information to extract the local contexts of these words and build their contextual semantic vectors as described in Section 3 . 4 .

4 . 2 . Experiments . This sect ion describes the experiments performed with the stat ist ical , semantic and combined algorithms for identification of false friends .

Baseline . As baseline we took the following algorithm :

- ASC – words pairs sorted in ascending order (first by the Bulgarian word and second by the Russian word) . It behaves nearly l ike a random function .

Statistical Algorithms . We performed the following experiments based on the stat ist ical approach for identifying false fr iends in a parallel text :