Cognate Production using Character-based Machine Translation

Lisa Beinborn, Torsten Zesch and Iryna Gurevych

Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt

Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for International Educational Research

www.ukp.tu-darmstadt.de

Abstract

Cognates are words in different languages that are associated with each other by language learners. Thus, cognates are important indicators for the prediction of the perceived difficulty of a text. We introduce a method for automatic cognate production using character-based machine translation. We show that our approach is able to learn production patterns from noisy training data and that it works for a wide range of language pairs. It even works across different alphabets, e.g. we obtain good results on the tested language pairs English-Russian, English-Greek, and English-Farsi. Our method performs significantly better than similarity measures used in previous work on cognates.

1 Introduction

In order to improve comprehension of a text in a foreign language, learners use all possible information to make sense of an unknown word. This includes context and domain knowledge, but also knowledge from the mother tongue or any other previously acquired language. Thus, a student is more likely to understand a word if there is a similar word in a language she already knows (Ringbom, 1992). For example, consider the following German sentence:

> *Die internationale Konferenz zu kritischen Infrastrukturen im Februar ist eine Top-Adresse für Journalisten.*

Everybody who knows English might grasp the gist of the sentence with the help of associated words like *Konferenz-conference* or *Februar-February*. Such pairs of associated words are called *cognates*.

A strict definition only considers two words as cognates, if they have the same etymological origin, i.e. they are genetic cognates (Crystal, 2011). Language learners usually lack the linguistic background to make this distinction and will use all similar words to facilitate comprehension regardless of the linguistic derivation. For example, the English word *strange* has the Italian correspondent *strano*. The two words have different roots and are therefore genetically unrelated. However, for language learners the similarity is more evident than for example the English-Italian genetic cognate *father-padre*. Therefore, we aim at identifying all words that are sufficiently similar to be associated by a language learner no matter whether they are genetic cognates. As words which are borrowed from another language without any modification (such as *cappuccino*) can be easily identified by direct string comparison, we focus on word pairs that do not have identical spelling.

If the two associated words have the same or a closely related meaning, they are true cognates, while they are called false cognates or false friends in case they have a different meaning. On the one hand, true cognates are instrumental in constructing easily understandable foreign language examples, especially in early stages of language learning. On the other hand, false friends are known to be a source of errors and severe confusion for learners (Carroll, 1992) and need to be practiced more frequently. For these reasons, both types need to be considered when constructing teaching materials. However, existing lists of cognates are usually limited in size and only available for very few language pairs. In order to improve language learning support, we aim at automatically creating lists of related words between two languages, containing both, true and false cognates.

International Joint Conference on Natural Language Processing, pages 883–891,
Nagoya, Japan, 14-18 October 2013.

In order to construct such cognate lists, we need to decide whether a word in a source language has a cognate in a target language. If we already have candidate pairs, string similarity measures can be used to distinguish cognates and unrelated pairs (Montalvo et al., 2012; Sepúlveda Torres and Aluisio, 2011; Inkpen et al., 2005; Kondrak and Dorr, 2004). However, these measures do not take the regular production processes into account that can be found for most cognates, e.g. the English suffix *~tion* becomes *~ción* in Spanish like in *nation-nación* or *addition-adición*. Thus, an alternative approach is to manually extract or learn production rules that reflect the regularities (Gomes and Pereira Lopes, 2011; Schulz et al., 2004).

All these methods are based on string alignment and thus cannot be directly applied to language pairs with different alphabets. A possible workaround would be to first transliterate foreign alphabets into Latin, but unambiguous transliteration is only possible for some languages. Methods that rely on the phonetic similarity of words (Kondrak, 2000) require a phonetic transcription that is not always available. Thus, we propose a novel production approach using statistical character-based machine translation in order to directly produce cognates. We argue that this has the following advantages: (i) it captures complex patterns in the same way machine translation captures complex rephrasing of sentences, (ii) it performs better than similarity measures from previous work on cognates, and (iii) it also works for language pairs with different alphabets.

2 Character-Based Machine Translation

Our approach relies on statistical phrase-based machine translation (MT). As we are not interested in the translation of phrases, but in the transformation of character sequences from one language into the other, we use words instead of sentences and characters instead of words, as shown in Figure 1. In the example, the English character sequence *cc* is mapped to a single *c* in Spanish and the final *e* becomes *ar*. It is important to note that these mappings only apply in certain contexts. For example, *accident* becomes *accidente* with a double *c* in Spanish and not every word-final *e* is changed into *ar*. In statistical MT, the training process generates a phrase table with transformation probabilities. This information is combined with language model probabilities and a search algorithm selects the best combination of sequences. The transformation is thus not performed on isolated characters, it also considers the surrounding sequences and can account for context-dependent phenomena. The goal of the approach is to directly produce a cognate in the target language from an input word in another language. Consequently, in the remainder of the paper, we refer to our method as COP (COgnate Production).

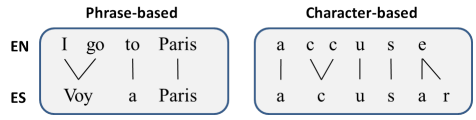

Figure 1: Character-based machine translation

Exploiting the orthographic similarity of cognates to improve the alignment of words has already been analyzed as a useful preparation for MT (Tiedemann, 2009; Koehn and Knight, 2002; Ribeiro et al., 2001). As explained above, we approach the phenomenon from the opposite direction and use statistical MT for cognate production.

Previous experiments with character-based MT have been performed for different purposes. Pennell and Liu (2011) expand text message abbreviations into proper English. In Stymne (2011), character-based MT is used for the identification of common spelling errors. Several other approaches also apply MT algorithms for transliteration of named entities to increase the vocabulary coverage (Rama and Gali, 2009; Finch and Sumita, 2008). For transliteration, characters from one alphabet are mapped onto corresponding letters in another alphabet. Cognates follow more complex production patterns. Nakov and Tiedemann (2012) aim at improving MT quality using cognates detected by character-based alignment. They focus on the language pair Macedonian-Bulgarian and use English as a bridge language. As they use cognate identification only as an intermediary step and do not provide evaluation results, we cannot directly compare with their work. To the best of our knowledge, we are the first to use statistical character-based MT for the goal of directly producing cognates.

3 Experimental Setup

Figure 2 gives an overview of the COP architecture. We use the existing statistical MT engine Moses (Koehn et al., 2007). The main difference of character-based MT to standard MT is the lim-

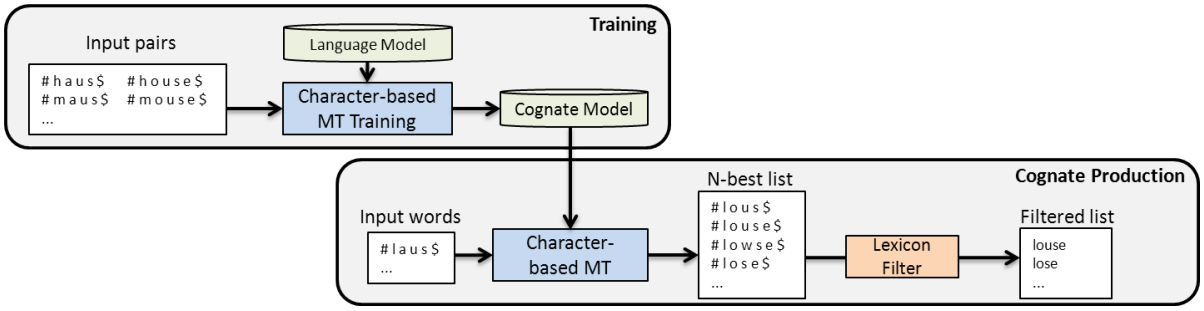

Figure 2: Architecture of our Cognate Production (COP) approach

ited lexicon. Our tokens are character n-grams instead of words, therefore we need much less training data. Additionally, distortion effects can be neglected as reordering of ngrams is not a regular morphological process for cognates.[1] Thus, we deal with less variation than standard MT.

Training As training data, we use existing lists of cognates or lists of closely related words and perform some preprocessing steps. All duplicates, multiwords, conjugated forms and all word pairs that are identical in source and target are removed. We lowercase the remaining words and introduce # as start symbol and $ as end symbol of a word. Then all characters are divided by blanks. Moses additionally requires a language model. We build an SRILM language model (Stolcke, 2002) from a list of words in the target language converted into the right format described above. On the basis of the input data, the Moses training process builds up a phrase table consisting of character sequences in our case. As a result of the training process, we receive a cognate model that can be used to produce cognates in the target language from a list of input test words.

Cognate Production Using the learned cognate model, Moses returns a ranked n-best list containing the n most probable transformations of each input word. In order to eliminate non-words, we check the n-best list against a lexicon list of the target language. The filtered list then represents our set of produced cognates. Note that, as discussed in Section 1, the list will contain true and false cognates. The distinction can be performed using a bilingual dictionary (if available) or with statistical and semantic measures for the identification of false friends (Mitkov et al., 2008; Nakov et al., 2007). For language learning, we need both types of cognates as foreign words also trigger wrong associations in learners (see Section 5.4).

Evaluation Metrics In order to estimate the cognate production quality without having to rely on repeated human judgment, we evaluate COP against a list of known cognates. Existing cognate lists only contain pairs of true cognates, but a word might have several true cognates. For example, the Spanish word *música* has at least three English cognates: *music*, *musical*, and *musician*. Therefore, not even a perfect cognate production process will be able to always rank the right true cognate on the top position. In order to account for the issue, we evaluate the coverage using a relaxed metric that counts a positive match if the gold standard cognate is found in the n-best list of cognate productions. We determined $n = 5$ to provide a reasonable approximation of the overall coverage.

We additionally calculate the mean reciprocal rank (MRR) as

$$MRR = \frac{1}{|C|} \sum_{i=1}^{|C|} \frac{1}{rank_i}$$

where C is the set of input words and $rank_i$ the rank of the correct cognate production. For example, if the target cognate is always ranked second-best, then the MRR would be 0.5.[2]

Note that in our language learning scenario, we are also interested in words that might be associated with the foreign word by learners, but are actually not true cognates (e.g. the English word *muse* might also be mistakenly associated with *música* by language learners). Unfortunately, an evaluation of the false cognates produced by COP is not covered by those metrics and thus left to a qualitative analysis as performed in section 5.4.

[1] We use these parameters: `-weight-l 1 -weight-d 0 -weight-w -1 -dl 0 -weight-t 0.2 0.2 0.2 0.2 0.2`

[2] BLEU (Papineni et al., 2002) is the common evaluation metric for MT, but would be misleading in our setting.

4 Experiments & Results

We conducted a set of experiments that cover different aspects of the cognate production process. First, we test whether the approach is able to learn simple production rules. We select optimal parameters and test the influence of the size and quality of the available training data. We then compare our best model to previous work. For these experiments, we use the language pair English-Spanish, as a large manually collected list of cognates is available for training and evaluation.

4.1 Ability to Learn Production Rules

We train COP on a list of just ten cognates all following the same production process in order to test whether COP can generally learn cognate production rules. We test two different processes: i) the pattern (*~tion→~ción*), as in *tradition-tradición* ii) the pattern (*~ance→~ancia*) as in *elegance-elegancia*. The experiment shows that COP correctly produces the respective target cognates for new input words with the same pattern. We can conclude that COP succeeds in learning the necessary patterns for cognate production. In the following, we investigate whether our approach can also be applied to noisy training data containing a mixture of many different production processes.

4.2 Parameter Selection

We vary the following COP parameters: the character n-gram size used for tokenization, the order of the language model, the lexicon used for filtering, and tuning of Moses parameters. We collected a list of 3,403 English-Spanish cognates and split it into training set (2,403), development set (673), and test set (327).[3] Table 1 shows the coverage in the 5 best productions and the MRR for each parameter.

N-gram Size We start with the n-gram size parameter that determines the tokenization of the input, the respective format for unigrams, bigrams, and trigrams for the word *banc* looks as follows:
 # b a n c $ / #b ba an nc c$ / #ba ban anc nc$
Higher order n-grams in general increase the vocabulary and thus lead to better alignment. However, they also require a larger amount of training data, otherwise the number of unseen instances is too high. We find that bigrams produce slightly better results than unigrams and trigrams, this is in line with findings by Nakov and Tiedemann (2012). Thus, in the following experiments, we use character bigrams.

[3]The cognates have been retrieved from several web resources and merged with the set used by Montalvo et al. (2012). All test cognate list can be found at:
`http://www.ukp.tu-darmstadt.de/data`

		Cov. (n=5)	MRR
1)	Unigram	.63	.43
	Bigram	.65	.49
	Trigram	.51	.40
2)	LM-order 5	.68	.48
	LM-order 10	.65	.49
3)	Web1T-Filter	.68	.52
	Wordlist-Filter	.65	.54
4)	Moses Tuning	.66	.54

Table 1: Parameter selection for COP. The settings in bold are used for the subsequent experiments.

Language Model The next parameter is the language model which determines the probability of a sequence in the target language, e.g. a model of order 5 considers sequences of character n-grams up to a maximum length of 5. Order 5 seems to be already sufficient for capturing the regular character sequences in a language. However, the ranks for the order-10 model are slightly better and as our "vocabulary" is very limited, we can savely decide for the language model of order 10.

Lexicon Filter For filtering the n-best cognate productions, we tried two different lexicon filter lists. A relatively broad one extracted from the English Web1T (Brants and Franz, 2006) word counts, and a more restrictive corpus-based list. The more restrictive filter decreases the coverage as it also eliminates some correct solutions, but it improves the MRR as non-words are deleted from the n-best list and the ranking is adjusted accordingly. The choice of the filter adjusts the trade-off between cognate coverage and the quality of the n-best list. For our language learning scenario, we decide to use the more restrictive filter in order to assure high quality results.

Moses Parameters Finally, we tune the Moses parameter weights by applying minimum error rate training (Och and Ney, 2003) using the development set, but it makes almost no difference in this setting. Tuning optimizes the model with respect to the BLEU score. For our data, the BLEU score is quite high for all produced cognate candidates, but it is not indicative of the usefulness of the transformation. A word containing one wrong character is not necessarily better than a word con-

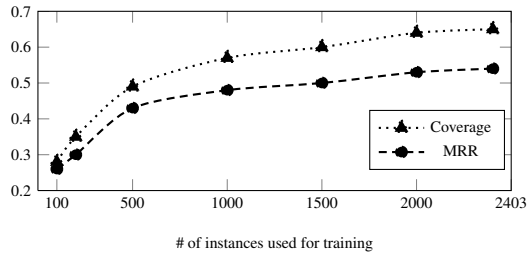

Figure 3: COP learning curve

taining two wrong characters. This explains why tuning has little effect.

Generally, COP reaches a coverage of about 65%. If we consider an n-best list with the 100 best translations (instead of only 5), the coverage increases only by less than 1% on average, i.e. the majority of the correct cognates can be found in the top 5. This is also reflected by the high MRR. In the following experiments, we use the optimal parameter setting (highlighted in Table 1).

4.3 Training Data Size & Quality

As we have seen in the experiments in Section 4.1, COP is able to learn a production rule from only few training instances. However, the test dataset contains a variety of cognates following many different production processes. Thus, we evaluate the effect of the size of the training data on COP. The learning curve in Figure 3 shows the results. As expected, both coverage and MRR improve with increasing size of the training data, but we do not see much improvement after about 1,000 training instances. Thus, COP is able to learn stable patterns from relatively few training instances.

However, even a list of 1,000 cognates is a hard constraint for some language pairs. Thus, we test if we can also produce satisfactory results with lower quality sets of training pairs that might be easier to obtain than a list of cognates.

We use word pairs extracted from the freely available multilingual resources UBY (Gurevych et al., 2012) and Universal WordNet (UWN) (de Melo and Weikum, 2009). UBY combines several lexical-semantic resources, we use translations which were extracted from Wiktionary. UWN is based on WordNet and Wikipedia and provides automatically extracted translations for over 200 languages that are a bit noisier compared to UBY translations. Additionally, we queried the Microsoft Bing translation API using all words from an English word list as query words.[4] We also test a knowledge-free approach by pairing all words from the English and Spanish Web1T corpus.[5] While the translation pairs always share the same meaning, this is not the case for the Web1T pairs, where the majority of pairs will be unrelated.

		Training Size	Cov. (n=5)	MRR
Cognates		1,000 / 2,403	.57/ .65	.48 /.54
Transl.	UBY	1,000 / 6,048	.53 /.69	.47 /.56
	UWN	1,000 /10,531	.50 /.69	.43 /.54
	Bing	1,000 / 5,567	.51 /.64	.44 /.54
Knowledge-free		1,000 /34,019	.21 /.47	.18 /.33

Table 2: Influence of data size and quality

In order to increase the ratio of possible cognates in the training data, we apply a string similarity filter using the XDICE-measure with a threshold of 0.425[6] on the translation pairs, For the knowledge-free pairs, we use a stricter threshold of 0.6 in order to account for the lower quality.

For a fair quality comparison, we first limit the number of training instances to 1,000, where (as shown above) the performance increases leveled off. The left columns for coverage and MRR in Table 2 show the results. It can be seen, that the results for the translation pairs extracted from UBY, UWN and Bing are only slightly inferior to the use of manually collected cognates for training. The small differences between the resources mirror the different level of linguistic control that has been applied in their creation. The knowledge-free pairs from Web1T yield drastically inferior results. We can conclude that training data consisting of selected cognates is beneficial, but that a high quality list of translations in combination with a string similarity filter can also be sufficient and is usually easier to obtain.

In a follow-up experiment, we use the full size of each training set. As expected, coverage and MRR both increase in all settings. Even with the knowledge-free training set that introduces many noisy pairs, satisfactory results can be obtained. This shows that COP can be used for the production of cognates, even if no language-specific information beyond a lexicon list is available.

4.4 Comparison to Previous Work

Previous work (Kondrak and Dorr, 2004; Inkpen et al., 2005; Sepúlveda Torres and Aluisio, 2011;

[4] http://www.bing.com/translator

[5] We only use every 5th word in order to limit the number of results to a manageable size.

[6] The threshold was selected to cover ~80% of the test set.

	Cov. (n=5)	MRR
DICE	.46	.21
XDICE	.52	.25
LCSR	.51	.24
SpSim	.52	.22
COP	**.65**	**.54**

Table 3: Comparison of different approaches for cognate production.

	Language Pair	Cov. (n=5)	MRR
Same alphabet	en-es	.65	.54
	es-en	.68	.48
	en-de	.55	.46
Cross-alphabet	en-ru	.59	.47
	en-el	.61	.37
	en-fa	.71	.54

Table 4: COP results for other languages

Montalvo et al., 2012) is based on similarity measures that are used to decide whether a candidate word pair is a cognate pair, while COP directly produces a target cognate from the source word. In order to compare those approaches to COP, we pair the English input words from the previous experiments with all words from a list of Spanish words[7] and consider all resulting pairs as candidate pairs. For each pair, we then calculate the similarity score and rank the pairs accordingly. As the similarity measures often assign the same value to several candidate pairs, we get many pairs with tied ranks, which is problematic for computing coverage and MRR. Thus, we randomize pairs within one rank and report averaged results over 10 randomization runs.[8]

We compare COP to three frequently used string similarity measures (LCSR, DICE, and XDICE), which performed well in (Inkpen et al., 2005; Montalvo et al., 2012), and to SpSim which is based on learning production rules. The longest common subsequence ratio (LCSR) calculates the ratio of the length of the longest (not necessarily contiguous) common subsequence and the length of the longer word (Melamed, 1999). DICE (Adamson and Boreham, 1974) measures the shared character bigrams, while the variant XDICE (Brew and McKelvie, 1996) uses extended bigrams, i.e. trigrams without the middle letter. SpSim (Gomes and Pereira Lopes, 2011) is based on string alignment of identical characters for the extraction and generalization of the most frequent cognate patterns. Word pairs that follow these extracted cognate patterns are considered equally similar as pairs with identical spelling.

Table 3 shows the results. The differences between the individual similarity measures are very small, string similarity performs on par with SpSim. The low MRR indicates that the four measures are not strict enough and consider too many candidate pairs as sufficiently similar. COP performs significantly better than all other measures for both, coverage and MRR. The results for the similarity measures are comparable to the knowledge-free variant of COP (Cov = .47 and MRR = .33, compare Table 2). Obviously, COP better captures the relevant cognate patterns and thus is able to provide a better ranking of the production list. Another advantage of COP is its applicability to language pairs with different alphabets (see Section 5.2), while the similarity measures can only operate within one alphabet.

5 Multilinguality

The previous experiments showed that COP works well for the production of Spanish cognates from English source words. However, in language learning, we need to consider all languages previously acquired by a learner, which leads to a large set of language combinations. Imagine, for example, an American physician who wants to learn German. She has studied Spanish in school and the terminology in her professional field has accustomed her to Greek and Latin roots. When facing a foreign text, she might unconsciously activate cues from any of these languages. Thus, if we want to select suitable text for her, we need to consider cognates from many different languages.

In the following experiments, we test how COP performs for other languages with the same alphabet and across alphabets. In addition, we evaluate how well the cognates produced by COP correlate with human judgments.

5.1 Same Alphabet

We first analyze whether the cognate production also works in the reverse direction and test the production of English cognates from Spanish source words. The results in Table 4 (upper part) show that COP works bi-directionally, as the scores for Spanish to English are comparable to those for English to Spanish. In addition, we train a model for another Western European language pair, namely English-German. The results show that COP also works well for other language pairs.

[7] In order to ensure a fair comparison, we use the Spanish word list that is also used as lexicon filter in COP.

[8] The average standard deviation is 0.01.

English	Spanish	German	Russian	Greek	Farsi
alcohol	alcohol, alcoholar	alkohol, alkoholisch	алкоголь, алкогольный	αλκοολικό, αλκοολικά	الکی, الکل
coffee	café	-	кофей, кофе	-	قهوه
director	director, directora	direktor, direkt	директор	-	غیر, دیر
machine	machina	maschine, machen	махина, машина	μηχανή, μαχίν	ماشینی, ماشین
music	músico, música	musik, musisch	-	μουσικής, μουσικές	موسی, موسیقی
optimal	óptimo	optimal, optimiert	оптимальный	-	مطلوب
popular	popular	populär	популярный	-	محبور, محبوب
theory	teoría	theorie	теория	θεωρία, θεωρίας	نظری, تئوری
tradition	tradición	tradition	традиция, традиционный	-	سنتی, سنت

Table 5: Multilingual cognates for English source words produced by COP

5.2 Cross-Alphabet

Previous approaches to cognate identification only operate on languages using the same alphabet. As COP is able to learn correspondences between arbitrary symbols, it can easily be applied on cross-alphabet language pairs. In the previous experiments, we had excluded cognate pairs that have exactly the same string representation. For cross-alphabet pairs, this is not possible. Thus, the task is to tackle both, standard transliteration (as in the English-Greek pair *atlas*-άτλας)[9] and cognate production (as in *archangel*-αρχάγγελος)[10].

We evaluate COP for Russian (ru), Greek (el), and Farsi (fa). For Russian, we use a list of UBY-pairs as training data. Unfortunately, UWN and UBY contain only few examples for Greek and Farsi, so we use Bing translations of English source words. In order to filter the resulting list of words, we transliterate Russian and Greek into the Latin alphabet[11] and apply a string similarity filter. We do not filter the training data for Farsi, as the transliteration is insufficient.

The lower part of Table 4 lists the results. Given that those language pairs are considered to be less related than English-Spanish or English-German, the results are surprisingly good. Especially the production of Farsi cognates works very well, although the training data has not been filtered. The low MRR for Greek indicates that our lexicon filter is not restrictive enough. COP often produces Greek words in several declinations (e.g. nouns in genitive case) which are not eliminated and lead to a worse rank of the correct target. We conclude that COP also works well across alphabets.

5.3 Multilingual Cognates

In order to provide the reader with some examples of cognates produced by COP, we compiled a short list of international words that are likely to occur in all languages under study. In Table 5, we give the two top-ranked productions. It can be seen that COP produces both, true and false cognates (e.g. *direkt* for *director*), which is useful for language learning scenarios. Of course, some produced forms are questionable, e.g. the second Farsi match for *music* means *Moses*. Note that the gaps in the table are often cases where the absence of a cognate production is an indicator of COP's quality. For example, the Greek words for *director*, *popular*, and *tradition* are not cognates of the English word but have a very different form.

5.4 Human Associations

The examples in Table 5 showed that COP produces not only the correct cognate, but all target words that can be created from the input word based on the learned production processes. In order to assess how well these additional productions of COP correlate with human associations, we conducted a user study. We presented Czech words with German origin to 15 native German speakers that did not have any experience with Eastern-European languages. The participants were asked to name up to 3 guesses for the German translation of the Czech source word. Table 6 gives an overview of the Czech source words together with the German associations named by more than one person (number of mentions in brackets). The table shows that some Czech words are strongly associated with their correct German translations (e.g. *nudle-Nudel*), while other words trigger false friend associations (e.g. *talíř-Taler*).

Another interesting aspect is the influence of languages besides the L1. For example, the German association *himmel* for the Czech word *cíl* is very likely rooted in the Czech-French association

[9]The transliteration of άτλας is *átlas*.

[10]The transliteration of αρχάγγελο is *ark'aggelos*.

[11]Using ICU: `http://site.icu-project.org/`

[14]Note that forms like *stak* also pass the lexicon filter, as this is an infrequent, but nevertheless valid German word. Other words like *san* are part of the German lexicon from city names like *San Francisco*.

Czech	Human associations (German)	COP productions (German)
nudle	**Nudel** (15)	**nudel**, nadel, ode
švagr	**Schwager** (13)	sauger, **schwager**, berg
šlak	**Schlag** (12), Schlagsahne (3), schlagen (2)	stak
brýle	**Brille** (12), brüllen (4)	**brille**, brie
cíl	**Ziel** (9), Himmel (2)	set, zelle, teller
žold	**Sold**(9), Zoll (5), Gold (2), verkauft (2), Schuld (2)	**sold**, gold, geld
sál	Salz (13) , **Saal** (8)	set, san, all, **saal**
taška	**Tasche** (8), Aufgabe (4), Tasse (4), Taste (2)	task, as, tick
skříň	**Schrein** (5), Bildschirm/Screen (3), schreien (2)	-
flétna	**Flöte** (4), Flotte (4), Pfannkuchen (2), fliehen (2)	flut, filet
muset	Museum (11), **müssen** (3), Musik (3), Muse (2), Mus (2)	mus, most, muse, mit
valčík	Walze (4), **Walzer** (3), falsch (2)	-
talíř	Taler (5), **Teller** (2), zahlen (3), teilen (2)	**teller**, taler, ader
šunka	schunkeln (2), Sonne (2), **Schinken** (1),	sun
knoflík	Knoblauch (11), knifflig (4), **Knopf** (1)	-

Table 6: Human associations and cognate productions from Czech to German
Correct translations are in bold, underlined words are COP productions that match human associations.[14]

cíl-ciel.[15] A similar process applies for the association *aufgabe*, which is *task* in English and therefore close to *taška*. These cross-linguistic cognitive processes highlight the importance of considering cognates from all languages a learner knows.

In order to examine how well COP reflects the human associations, we train it on manually collected Czech-German cognates and translation pairs from UBY. The number of training instances is rather small, as a language reform in the 19th century eliminated many Czech words with Austrian or German roots. Consequently, the model does not generalize as well as for other language pairs (see the column "COP Productions" in Table 6).[16] However, it correctly identifies cognates like *nudel*, *brille*, and *sold* which are ranked first by the human participants. As we argued above, COP also correctly produces some of the 'wrong' associations, e.g. *gold* or *taler*. Thus, COP is to a certain extent able to mimic the association process that humans apply when identifying cognates.

6 Conclusions

We introduced COP, a novel method for cognate production using character-based MT. We have shown that COP succeeds in learning the necessary patterns for producing cognates in different languages and alphabets. COP performs significantly better than similarity measures used in previous work on cognates. COP relies on training data, but we have shown that it can be applied even if no language-specific information beyond a word list is available. A user study on German-Czech cognates supports our assumption that COP productions are comparable to human associations and can be applied for language learning.

In future work, we will focus on the application of cognates in language learning. True cognates are easier to understand for learners and thus can be an important factor for readability assessment and the selection of language learning examples. False cognates, on the other hand, can be confusing and need to be practiced more frequently. They could also be used as good distractors for multiple choice questions. In addition, COP productions that do not pass the lexical filter might serve as pseudo-words in psycholinguistic experiments as they contain very probable character sequences.

[15] Both words, *himmel* and *ciel* mean *heaven* in English.

[16] Coverage (0.4) and MRR (0.32) are not representative as the test set is too small.

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806, and by the Klaus Tschira Foundation under project No. 00.133.2008.

References

George W Adamson and Jillian Boreham. 1974. The Use of an Association Measure based on Character Structure to Identify Semantically Related Pairs of Words and Document Titles. *Information Storage and Retrieval*, 10(7):253–260.

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram corpus version 1. *Linguistic Data Consortium*.

Chris Brew and David McKelvie. 1996. Word-Pair Extraction for Lexicography. *Proc. of the 2nd international conference on new methods in language processing*, pages 45–55.

Susan E. Carroll. 1992. On Cognates. *Second Language Research*, 8(2):93–119, June.

David Crystal. 2011. *Dictionary of linguistics and phonetics*, volume 30. Wiley-Blackwell.

Gerard de Melo and Gerhard Weikum. 2009. Towards a Universal Wordnet by Learning from Combined Evidence. *Proc. of the 18th ACM conference*, pages 513–522.

Andrew Finch and Eiichiro Sumita. 2008. Phrase-based machine transliteration. In *Proc. of the Workshop on Technologies and Corpora for Asia-Pacific Speech Translation (TCAST)*, pages 13–18.

Luís Gomes and José Gabriel Pereira Lopes. 2011. Measuring Spelling Similarity for Cognate Identification. *Progress in Artificial Intelligence*, pages 624–633.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M Meyer, and Christian Wirth. 2012. A Large-Scale Unified Lexical-Semantic Resource Based on LMF. *Proc. of the 13th Conference of the EACL*, pages 580–590.

Diana Inkpen, Oana Frunza, and Grzegorz Kondrak. 2005. Automatic Identification of Cognates and False Friends in French and English. In *Proc. of the International Conference Recent Advances in NLP*, pages 251–257.

Philipp Koehn and Kevin Knight. 2002. Learning a Translation Lexicon from Monolingual Corpora. In *Proceedings of the ACL workshop on Unsupervised lexical acquisition*, pages 9–16, July.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-burch, Richard Zens, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Chris Dyer, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. *Annual Meeting of the Association for Computational Linguistics*.

Grzegorz Kondrak and Bonnie Dorr. 2004. Identification of Confusable Drug Names: A New Approach and Evaluation Methodology. In *Proc. of the 20th international conference on Computational Linguistics*, pages 952–958.

Grzegorz Kondrak. 2000. A New Algorithm for the Alignment of Phonetic Sequences. In *Proceedings of the 1st NAACL*, pages 288–295.

I. Dan Melamed. 1999. Bitext Maps and Alignment via Pattern Recognition. *Computational Linguistics*, 25(1):107–130.

Ruslan Mitkov, Viktor Pekar, Dimitar Blagoev, and Andrea Mulloni. 2008. Methods for extracting and classifying pairs of cognates and false friends. *Machine Translation*, 21(1):29–53, May.

Soto Montalvo, Eduardo G. Pardo, Raquel Martinez, and Victor Fresno. 2012. Automatic Cognate Identification based on a Fuzzy Combination of String Similarity Measures. *IEEE International Conference on Fuzzy Systems*, pages 1–8, June.

Preslav Nakov and Jörg Tiedemann. 2012. Combining Word-Level and Character-Level Models for Machine Translation Between Closely-Related Languages. In *Proceedings of the ACL*, pages 301–305.

Svetlin Nakov, Preslav Nakov, and Elena Paskaleva. 2007. Cognate or False Friend? Ask the Web! In *Proc. of the RANLP workshop: Acquisition and management of multilingual lexicons*, pages 55–62.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu Ibm. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proc. of the 40th annual meeting of the ACL*, pages 311–318, July.

Deana L Pennell and Yang Liu. 2011. A Character-Level Machine Translation Approach for Normalization of SMS Abbreviations. pages 974–982.

Taraka Rama and Karthik Gali. 2009. Modeling Machine Transliteration as a Phrase Based Statistical Machine Translation Problem. *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*, (August):124–127.

António Ribeiro, Gaël Dias, Gabriel Lopes, and João Mexia. 2001. Cognates Alignment. *Proc. of the Machine Translation Summit 2001*.

Hakan Ringbom. 1992. On L1 Transfer in L2 Comprehension and L2 Production. *Language Learning*, 42(1):85–112.

Stefan Schulz, Eduardo Sbrissia, Percy Nohama, and Udo Hahn. 2004. Cognate Mapping - A Heuristic Strategy for the Semi-Supervised Acquisition of a Spanish Lexicon from a Portuguese Seed Lexicon. In *Proc. of the 20th international conference on Computational Linguistics*.

Lianet Sepúlveda Torres and Sandra Maria Aluisio. 2011. Using Machine Learning Methods to avoid the Pitfall of Cognates and False Friends in Spanish-Portuguese Word Pairs. In *Proc. of the 8th Brazilian Symposium in Information and Human Language Technology*, pages 67–76.

Andreas Stolcke. 2002. SRILM-an Extensible Language Modeling Toolkit. In *Proc. of the international conference on spoken language processing*, volume 2, pages 901–904.

Sara Stymne. 2011. Spell Checking Techniques for Replacement of Unknown Words and Data Cleaning for Haitian Creole SMS Translation. *Proc. of the 6th Workshop on SMT*, pages 470–477.

Jörg Tiedemann. 2009. Character-based PSMT for Closely Related Languages. In *Proc. of 13th Annual Conference of the European Association for Machine Translation*, volume 9, pages 12–19.