Authors:	Toivonen Jarmo, Pirkola Ari, Keskustalo Heikki, Visala Kari, Järvelin Kalervo
Name of article:	Translating cross-lingual spelling variants using transformation rules
Year of publication:	2005
ISBN:	InformationProcessingManagement
Volume:	41
Number of issue:	4
Pages:	859-872
ISSN:	0306-4573
Discipline:	Natural sciences / Computer and information sciences
Language:	en
School/Other Unit:	School of Information Sciences

URN: http://urn.fi/urn:nbn:uta-3-492
DOI: http://dx.doi.org/10.1016/j.ipm.2004.02.001

Translating cross-lingual spelling variants using transformation rules ⋆

Jarmo Toivonen [a,*], Ari Pirkola [b], Heikki Keskustalo [b], Kari Visala [b], Kalervo Järvelin [b]

[a] *Institute of Signal Processing, Tampere University of Technology, P.O. Box 553, FIN-33101 Tampere, Finland*

[b] *Department of Information Studies, University of Tampere, P.O. Box 607, FIN-33101 Tampere, Finland*

Abstract

Technical terms and proper names constitute a major problem in dictionary-based cross-language information retrieval (CLIR). However, technical terms and proper names in different languages often share the same Latin or Greek origin, being thus spelling variants of each other. In this paper we present a novel two-step fuzzy translation technique for cross-lingual spelling variants. In the first step, transformation rules are applied to source words to render them more similar to their target language equivalents. The rules are generated automatically using translation dictionaries as source data. In the second step, the intermediate forms obtained in the first step are translated into a target language using fuzzy matching. The effectiveness of the technique was evaluated empirically using five source languages and English as a target language. The two-step technique performed better, in some cases considerably better, than fuzzy matching alone. Even using the first step as such showed promising results.

Key words: Cross-language retrieval, Fuzzy matching, Transliteration

⋆ A shorter version of this paper was presented at the 26^{th} Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Toronto, Canada, July 28 – August 1 2003)

* Corresponding author. Tel.: +358 3 3115 3903; fax: +358 3 3115 3087.

Email addresses: jarmo.toivonen@tut.fi (Jarmo Toivonen), ari.pirkola@uta.fi (Ari Pirkola), heikki.keskustalo@uta.fi (Heikki Keskustalo), kari.visala@uta.fi (Kari Visala), kalervo.jarvelin@uta.fi (Kalervo Järvelin).

1 Introduction

Technical terms and proper names are often central keys in requests for information. In dictionary-based cross-language information retrieval (CLIR) they constitute a major problem, since they are not found in general translation dictionaries, except for the most commonly used terms and names. In dictionary-based CLIR untranslatable query keys are typically used in target language queries in their original source language forms. Unless they are identical to the corresponding database index terms, they do not match the index terms, causing significant loss of retrieval effectiveness. However, technical terms (proper names) in different languages often share the same Latin or Greek origin, being thus *spelling variants* of each other, as German *konstruktion* and English *construction*. This allows the use of fuzzy matching (approximate string matching) techniques to find the target language correspondents of source language keys.

Approximate matching techniques involve Soundex and Phonix, which compare words on the basis of their phonetic similarity (Gadd, 1990), edit distance (Zobel and Dart, 1996), and n-gram based matching (Robertson and Willett, 1998). In n-gram matching text strings are decomposed into n-grams, i.e., substrings of length n, which usually consist of the adjacent characters of the text strings. The degree of similarity between the strings is computed on the basis of the number of similar n-grams and the total number of unique n-grams in the strings.

Transliteration refers to phonetic translation across languages with different orthographies (Knight and Graehl, 1998), such as Arabic to English (Stalls and Knight, 1998) or Japanese to English (Qu et al., 2003). In this paper we will present a novel two-step *fuzzy translation* technique for cross-lingual technical terms and proper names. It is similar to transliteration, but no phonetic elements are included. The technique bears some resemblance to query translation and transliteration research reported in Fujii and Ishikawa (2001). Fujii and Ishikawa use character-based rules to establish mapping between English characters and romanized Japanese katakana characters. They also utilize probabilistic character-based language models, which can be seen as a variation of the fuzzy-matching technique. Fujii's and Ishikawa's technique, on the other hand, is focused on languages with different orthographies and thus has a different focus from ours.

In the first step of our technique, source language words are transformed into *intermediate forms* by means of transformation rules. The intermediate forms are often correct translations or more similar word forms to their target language equivalents than the original source language words. We call this step *transformation rule based translation* (TRT). A *transformation rule* refers to an automatically extracted regular correspondence between the characters in two languages, for instance Spanish character string *ia* corresponds to English character *y*, e.g., in the term pair *somatologia – somatology*.

In the second step of fuzzy translation, the intermediate forms achieved in the first step are matched with their target language equivalents through fuzzy matching. The benefits of the combined technique are in cases where TRT does not yield correct translations but renders source words more similar to their target language equivalents. This allows n-gram matching to rank the correct equivalents high.

The transformation rules were generated automatically by extracting equivalent term pairs from translation dictionaries. The terms were then aligned pairwise and regular correspondences were identified using the edit distance measure. The rules were generated for five language pairs, with English always being a target language and Finnish, French, German, Spanish, and Swedish source languages.

The effectiveness of the two-step fuzzy translation technique was evaluated by means of test words in five different domains. The intermediate forms obtained using TRT were matched through n-gram matching against an English target word list of 189 000 words, including the correct equivalents of the source words. As an evaluation measure we used precision at the rank where all the equivalents of the source words have been retrieved. We will demonstrate that the combined fuzzy translation technique performs better, sometimes considerably better, than n-grams alone.

In Pirkola et al. (2003) we presented first results on the fuzzy translation technique. In this paper we present more detailed results and extend the first study by exploring how effective TRT is as such when used alone without fuzzy matching. This is an important question when TRT applications are considered. We evaluated the effectiveness of TRT by considering what proportion of word forms obtained through TRT are correct translations (translation precision) and what proportion of source words are translated correctly (translation recall).

The rest of this paper is organized as follows. Section 2 presents the methodology and data, and Section 3 the findings. Section 4 contains the discussion and conclusions.

2 Methods and Data

2.1 Automatic generation of rules

This section describes the automatic rule generation process. Figure 1 illustrates the process by means of examples. The process consisted of the following main steps:

- Extracting similar terms from a dictionary

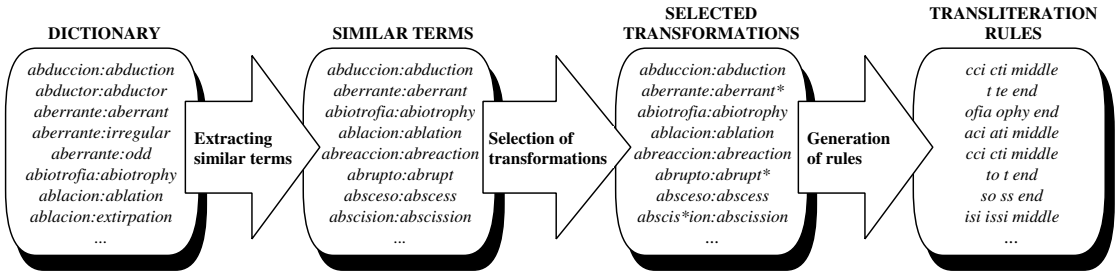

Fig. 1. The automatic rule generation process.

- Selection of transformations
- Generation of transformation rules

Extracting similar terms

The source and target language term pairs that were sufficiently similar were identified, and were extracted from a dictionary for further processing. The similarity was determined using edit distance (ED, Levenshtein distance) with a threshold value. Edit distance is a string similarity measure, which is defined as the minimum cost needed to convert one string into another. Conversion includes the operations of character substitution (sub), insertion (ins), and deletion (del). We use the term *transformation* as a general term that covers substitution, insertion, and deletion. For the strings A and B, edit distance is as follows:

$$ED(A, B) = min\{N_{sub} + N_{ins} + N_{del}\} \quad (1)$$

The equation thus gives the minimum sum of the number of operations needed to convert string A into string B.

Transformation rules were produced for the following language pairs:

- Finnish – English
- French – English
- German – English
- Spanish – English
- Swedish – English

German – English, French – English, and Spanish – English term pairs were extracted from the multilingual medical dictionary provided by André Fairchild. The number of dictionary entries varied between 11 000 - 12 000 depending on the language pair. Finnish – English and Swedish – English term pairs were obtained by translating Finnish and Swedish lists of medical terms into English using the MOT dictionaries by Kielikone Plc. The Finnish list contained 5970 terms. The Swedish list was small containing just 657 terms.

Selection of transformations

In the next step, all the transformations that produced the minimum ED were searched for each term pair using a recursive algorithm. The algorithm was based on the AllAlignment algorithm described in Charras and Lecroq (1998). From the result set of all transformations, one transformation was selected. The selection was done using the smallest sum of error values. The error values for the transformations were calculated as follows (Covington, 1996):

- 0, terms share the same character at the same position
- 1, consonant - consonant substitution, and vowel - vowel substitution
- 1, insertion or deletion of a character
- 2, consonant - vowel substitution, and vowel - consonant substitution

In Figure 1, an asterisk represents insertion when it is in a source term and deletion when it is in a target term.

The generation of rules

In the first phase of the rule generation process, the rules of double letter/single letter insertions/deletions were generated, e.g., $ss \rightarrow s$ and $s \rightarrow ss$. In the second phase the strings were studied from the start to the end, and differences were recorded.

A given rule typically occurs in a certain location of a word, and prior to and after a certain character. This context information was recorded for each rule. Occurrence information of the rules was put in a hash table, and the *frequency* of the rule was computed. *Confidence factor* is defined as the frequency of a rule divided by the number of source words where the source string of the rule occurs. All these, context information, frequency, and confidence factor are utilized when the automatically generated rules are applied.

2.2 Sample rules

Table 1 shows sample German-to-English rules, sorted on the basis of the frequency of the rules. The sixth line, for example, shows that the letter k, prior to t and after e, is transformed into the letter c in the middle of words, with the confidence factor being 89.25% ($100\% * 191/214$).

2.3 Testing the effectiveness of two-step fuzzy translation

For each source language, terms in the following five domains were used as test words:

Table 1
A sample of German-to-English rules.

Source string	**Target string**	**Location of the rule**	**Frequency**	**No of words**	**Confidence factor (%)**
isc	ic	middle	549	752	73.01
ie	ia	end	474	1179	40.20
se	sis	end	352	588	59.86
n	ne	end	274	1585	17.29
ch	c	end	198	325	60.92
ekt	ect	middle	191	214	89.25
m	ma	end	169	801	21.10
akt	act	middle	163	188	86.70
ko	co	beginning	159	197	80.71
le	l	end	158	301	52.49
ka	ca	beginning	151	199	75.88
che	c	end	149	333	44.74
[*etc.*]					

- Medical, biological, and chemical terms, n=90 (number of test words) (called *Bio terms* in the tables)
- Place names, $n = 55$
- Terms in economics, $n = 31$
- Terms in technology, $n = 36$
- Miscellaneous terms, $n = 59$

The bio terms and place names were gathered by browsing the target word list from the start to the end. A set of English terms was selected as test words and was translated into the five source languages by a research assistant. The translations were checked by native speakers or advanced students of the source languages. The terms in economics, technology and the miscellaneous terms were selected from the dictionaries. In all 123 of the 126 English equivalents of the terms were found in the target word list. The remaining 3 equivalents were added into the target list so that for each source language all 126 source words were available for the tests.

As we tested 5 source languages the total number of test words used in the experiments was $5 * (90 + 55 + 31 + 36 + 59) = 1355$ words. In addition, for determining confidence factor and frequency thresholds, we used a separate Spanish training data.

The source words were translated by means of TRT to obtain the intermediate forms (Section 2.3.1). The intermediate forms were then matched using n-gram matching against the words of the target word list (section 2.3.2). The *target word list* was the index of CLEF's (Peters, 2002) LA Times collection, containing 189 000 words. Similarity between the intermediate forms and the words in the target word list was computed to obtain a ranked list of target words.

As baselines we used digrams and trigrams, i.e., n-grams containing two and three characters respectively. For baselines similarity was computed between the source words and the words in the target word list.

2.3.1 Translating source words through TRT

Two TRT translation strategies were examined in combination with n-gram matching. The first one is called a *high confidence factor* (HCF) *strategy*. Using a relatively high confidence factor as a threshold this strategy seeks to minimize the number of incorrect transformations. Based on the training results a confidence factor of 50% was used as a threshold. For each source word one intermediate form was produced by applying to a source word all the rules applicable to it (one rule, two rules etc., or no rule). A drawback associated with HCF is that the number of rules that are available is limited.

In HCF the rules were applied to source words in the following reading order: (1) the location of the rules in source words, (2) the source string length, and (3) confidence factor. In (1) the application order was as follows: end, beginning, and middle location rules. For example, for the Finnish word *konvektio* the rules of "$o \rightarrow on$ (end)", "$ko \rightarrow co$ (beginning)", and "$ekt \rightarrow ect$ (middle)" were applied in this order to yield the intermediate form *convection* (which is a correct translation). In (2) and (3) the rules were applied starting from the longest source string and the highest confidence factor value. In the case of competitive rules (i.e., the same character sequence may be transformed using more than one rule) only the first rule of the reading was applied to a word. As TRT is a new method we did not have prior knowledge which order might give the most accurate intermediate forms. The application order of the rules has effects in the case of competitive rules which, however, were not common. However, the optimization of the application order is a question that needs further investigation.

The second strategy is called a *low confidence factor* (LCF) *strategy*. For each source word all the possible intermediate forms were produced by applying to a source word all the rules applicable to it. However, a threshold confidence factor of 10% was used to filter out unreliable rules. For example, for the Finnish word *konvektio* 7 intermediate forms were obtained, including the forms *konvectio*, *convektion*, and *convection*. In LCF the application order of the rules is irrelevant, as each order yields the same intermediate forms. Each intermediate form of the source word gave one result list. These were combined on the basis of the scores of each match to yield one ranked result list.

The rationale behind LCF is that it is likely that the set of intermediate forms obtained through TRT includes the correct equivalent of the source word, provided that the rules are good (the original source word was included in the set of intermediate forms). A drawback associated with LCF is that it may give many

incorrect transformations.

Both in HCF and LCF the (bad) rules whose frequency was less than 50 were removed.

2.3.2 N-gram matching

In n-gram matching words are decomposed into n-grams, i.e., into substrings of length n (Keskustalo et al., 2003; Pfeifer et al., 1996; Pirkola et al., 2002; Robertson and Willett, 1998; Salton, 1989). N-gram matching has been reported to be an effective technique among various approximate matching techniques in name searching (Pfeifer et al., 1996; Zobel and Dart, 1995) and cross-lingual spelling variant matching (Keskustalo et al., 2003) and is an appropriate fuzzy matching technique for use with TRT.

In this study, digrams and trigrams were used as baselines, and in combination with TRT. Both in digrams and trigrams the start and end white spaces were used as constituent characters of n-grams.

The degree of similarity between the source words (intermediate forms) and target words, w_1 and w_2, was computed on the basis of the number of n-grams that the words have in common and the total number of unique n-grams in the words, as follows (Pfeifer et al., 1996):

$$SIM(w_1, w_2) = \frac{\mid N_1 \cap N_2 \mid}{\mid N_1 \cup N_2 \mid} \tag{2}$$

where N_i refers to the set of n-grams derived from the word w_i, with $i = 1, 2$.

For each test word *precision* was calculated. In this context precision is defined as the proportion of correct equivalents at the last correct equivalent in the ranked result list of n-gram matching. However, each source word possessed one correct equivalent in the target word list, and the calculation of precision was reduced to $1/pce$, where *pce* stands for the position of the correct equivalent in the result list of n-gram matching. Finally, *average precision* over all test words was computed. Sometimes two or more words share the same SIM-value. Therefore two variants of precision were computed: in the first one, the correct equivalent was assumed to be the last word among the words with equal SIM-value; in the second one, the correct equivalent was assumed to be in the middle of the set of the words with equal SIM-value. Due to the small number of matching words having the same SIM-value the two variant measures of precision gave almost the same results. We therefore show average position precision results only in Tables 2-9 in the Findings section.

2.4 Testing the effectiveness of TRT

The effectiveness of TRT was evaluated by calculating *translation recall*, i.e., the proportion of source words for which TRT yields correct equivalents among all source words, and *translation precision*, i.e., the proportion of correct equivalents among all word forms yielded by TRT. Both translation recall and precision were computed at four confidence factor (CF) and frequency (Fr.) levels. The following combinations were tested:

- CF=50%, Fr.=50
- CF=10%, Fr.=50
- CF=10%, Fr.=10
- CF=2.0%, Fr.=4

In this test the test words were the same as in the fuzzy translation test (Section 2.3).

3 Findings

3.1 Two-step fuzzy translation

For Swedish, transformation rules were produced by using only 657 term pairs. The combined TRT and fuzzy matching technique was not useful, but it performed as well or slightly worse than fuzzy matching alone. The Swedish results suggest that the rules should be formed on the basis of thousands rather than hundreds of term pairs.

The results of fuzzy translation tests are presented in Tables 2-5 (HCF strategy) and Tables 6-9 (LCF strategy). There are several clear trends in Tables 2-9:

- The combined TRT and fuzzy matching technique performs well, but its effectiveness depends on the source language. For Finnish performance improvements are considerable (Tables 2 and 6). For German and Spanish performance improvements are smaller but the combined technique performs clearly better than digrams and trigrams alone. For French precision is changed only slightly (Tables 3 and 7). In most cases it is improved but sometimes slightly decreased.
- LCF yields better results than HCF in particular for French and Spanish. However, no major differences are found between the effectiveness of the strategies.
- In HCF, TRT with digrams performs better than TRT with trigrams in most cases. In LCF, TRT with trigrams performs roughly as well as TRT with digrams.
- The combined technique is useful for all term types. Thus, the results clearly

answer the question whether the rules generated by medical dictionaries, i.e., one specific domain, are suited for all term types (or just medical terms). In fact, the best improvements are achieved in the domain of technology.

In part, the results can be explained on the basis of the number of identical terms shared by a source language and English. For the set of technical terms ($n = 216$) the percentages of identical terms are as follows: Finnish (0.0%), French (48.8%), German (21.7%), and Spanish (11.1%). Thus, French and English terms often are identical, and the effects of TRT are minor, while Finnish very often uses its own spelling, and TRT is very effective in Finnish-to-English translation. Also the choice of using the same thresholds (confidence factor and frequency) for all test languages probably has clear effects on the results. It is possible that tuning the confidence factor for each language would have given even better results for some source languages. This issue needs further investigation.

3.2 TRT effectiveness

Tables 10-13 report translation recall and precision results by term domain. The number of rules (NoR) available for each level is also reported. As can be seen, translation recall is increased as confidence factor and frequency are decreased. For French, German and Spanish at CF=2.0% and Fr.=4 recall is between 77.8% and 97.8% in the domains of bio terms and technology. For Finnish the percentages are generally lower than for the other languages. This can be accounted for the low starting level, i.e., the percentage of identical terms between Finnish and English was 0.0% (the original source words were included in the set of word forms yielded by TRT). On the other hand, the automatic rule generation process was not able to capture some important Finnish-to-English rules (some vowel deletion rules) for reasons which have to be analyzed. This may in part explain lower recall for Finnish.

We analyzed the transformation errors in Finnish-to-English and French-to-English TRT in the cases of CF=2.0% and Fr.=4, and terms in economics. The results of the analysis are presented in Table 14. There were three types of errors. In the error type 1, the source and target words are similar but the transformation is rare or irregular so that no general rule can be derived easily. As an example of such an error, Finnish-to-English transformation *tariffi* → *tariff* would need a rule "*ffi* → *ff* (end)". This transformation is very rare, and that rule did not occur in our rule collection. In the error type 2, the source and target words differ much from each other. In these cases the application of the rules may be very complicated, and rules may be available only for some parts of a word, e.g., *kansleri* → *chancellor* from Finnish to English. In the error type 3, transformation occurs frequently in a language pair but our method was not able to capture the rule. An example of such a case is the word pair *koodi* → *code*, where a common deletion of a vowel occurs

when translating from Finnish to English. Probably, this rule was not found in our rule collection due to a relatively small number of Finnish-English word pairs available for rule generation.

As shown in Table 14, the most common error types are 1 and 3. It might be possible to address them by increasing the size of a word pair set used in rule generation. TRT through intermediate languages seems a promising method to improve TRT effectiveness. It could give correct translations in cases where direct rules are not available for source and target words.

As shown in Tables 10-13, TRT translates most terms correctly at low confidence factor and frequency levels. The high recall achieved at CF=2.0% and Fr.=4 is associated with low precision. For French bio terms the recall of 92.2% is associated with the precision of 9.3% (83/894). For German the corresponding percentages are 97.8% and 12.8% (88/685), respectively. These numbers suggest that TRT as such without fuzzy matching is insufficient in machine translation, in CLIR, or other applications, and must be supplemented with a disambiguation method capable of filtering out the incorrect word forms.

In the domain of place names recall is generally lower than in the other domains for all test languages. Cross-lingual variation in geographical names cannot be captured by means of transformation rules to the same extent as variation in technical terms, since cross-lingual geographical names have often diverged much from their original forms.

4 Discussion and Conclusions

Technical terms and proper names often are untranslatable due to limited coverage of translation dictionaries. This has a depressing effect on CLIR performance, as such expressions often are central keys in queries. In this study we presented a novel fuzzy translation technique based on automatically generated transformation rules and fuzzy matching. Two translation strategies were tested. In the high confidence factor strategy the aim was to minimize the number of incorrect transformations by using a relatively high confidence factor. In the low confidence factor translation strategy the rules were applied extensively, with a source word often yielding several intermediate forms. Digram and trigram matching were tested in combination with TRT. The results were encouraging as both strategies and combination methods performed better than digrams and trigrams alone. The results also showed that the effectiveness of HCF and LCF translation strategies, as well as digrams and trigrams, depends on a source language.

Digrams and trigrams alone often failed to give precise translations for terms which differed in more than two letters, viz., the extent of variation in the spelling variants

was relatively high. For example, the correct equivalent *allergy* of the Spanish term *alergia* was found at the 27^{th} position in the digram result list, whereas in the combined TRT and digram list it was at the first position, since TRT gave a correct translation. The strengths of the combined technique are marked particularly in cases where the extent of variation is very high, e.g., *Chechnya – Tchetchenie*. In cases like this fuzzy matching alone is powerless.

The figures below show the percentage of correct equivalents in four position classes in the ranked result list of Fin-Eng/TRT and digram matching (avg. precision 72.0%, Table 2). The distribution statistics is typical of all cases of this level precision.

Ranked position class	Percentage of correct equivalents belonging to class
1–2	72.2
3–4	7.8
5–6	6.7
> 10	13.3

As shown, 80% of the correct equivalents are within the set of four highest ranked words. Manual analysis of the n-gram result lists showed that TRT often raises the correct equivalents to the positions 1-2. In CLIR it is reasonable to select for the final query just a few highest ranked keys from the n-gram result list. The distribution figures above suggest that the TRT based fuzzy translation technique is viable in operational CLIR systems, the noise being acceptable. Moreover, it should be noted that there are several ways to improve this novel technique (see below).

In the second test we investigated how effective TRT is as such. Recall improvements were remarkable when confidence factor and frequency were decreased. We regard this as a promising result suggesting that TRT may be applied without fuzzy matching. However, at low confidence factor and frequency levels precision was low. In most applications precision close to 100% is required. Therefore an effective disambiguation method is needed to filter out incorrect word forms and to leave just the correct equivalents of the source words translated by means of TRT. One possible disambiguation method might be the use of word collocation statistics.

Our future research also involves the identification of language pairs for which fuzzy translation is effective, the improvement of the rules (for example, analysis of technical deficiencies involved in rules at present, and utilizing rule co-occurrence information), testing the effects of tuning a confidence factor by a specific language pair, selecting the best TRT and fuzzy matching combination, and testing how to apply fuzzy translation in actual CLIR research. Regarding the best combination we will explore other fuzzy matching techniques than those tested in this study together with TRT. One promising method is LCS (longest common subsequence) and another *skipgrams* described in Keskustalo et al. (2003) and Pirkola et al.

(2002). The actual CLIR research seeks to answer the question how fuzzy translation should be applied in an automatic CLIR query formulation and interactive CLIR to achieve the best possible retrieval performance.

Acknowledgments

Multilingual Medical Technical Dictionary (http://www.interfold.com/translator/) was provided by André Fairchild, of Denver, Colorado, USA. We would like to thank André Fairchild for permission to use the dictionary.

ENGTWOL morphological analyzer was used for the morphological analysis of the English data. ENGTWOL (Morphological Transducer Lexicon Description of English): Copyright (c) 1989-1992 Atro Voutilainen and Juha Heikkilä.
TWOL-R (Run-Time Two-Level Program): Copyright (c) Kimmo Koskenniemi and Lingsoft Ltd. 1983-1992.

This work was partly financed by the Clarity – Information Society Technologies (IST) Programme, Proposal/Contract no: IST-2000-25310.

References

Charras, C., & Lecroq, T. (1998). Sequence comparison. Available from: http://www-igm.univ-mlv.fr/~lecroq/seqcomp/.

Covington, M. A. (1996). An Algorithm to Align Words for Historical Comparison. *Computational Linguistics* 22 (4), 481–496.

Fujii, A., & Ishikawa, T. (2001). Japanese/English Cross-Language Information Retrieval: Exploration of Query Translation and Transliteration. *Computers and the Humanities* 35 (4), 389–420.

Gadd, T. (1990). Phonix: the algorithm. *Program* 24 (4), 363–369.

Keskustalo, H., Pirkola, A., Visala, K., Leppänen, E., & Järvelin, K. (2003). Non-adjacent Digrams Improve Matching of Cross-Lingual Spelling Variants. In: Nascimento, M. A., de Moura, E. S., & Oliveira, A. L. (Eds.), *Proceedings of the 10th International Symposium on String Processing and Information Retrieval (SPIRE 2003)* (pp. 252–265). No. 2857 in Lecture Notes in Computer Science. Heidelberg: Springer–Verlag.

Knight, K., & Graehl, J. (1998). Machine Transliteration. *Computational Linguistics* 24 (4), 599–612.

Peters, C. (2002). Cross-Language Evaluation Forum (CLEF). Available from: http://clef.iei.pi.cnr.it:2002/.

Pfeifer, U., Poersch, T., & Fuhr, N. (1996). Retrieval Effectiveness of Proper Name Search Methods. *Information Processing & Management* 32 (6), 667–679.

Pirkola, A., Keskustalo, H., Leppänen, E., Känsälä, A.-P., & Järvelin, K. (2002). Targeted s-gram matching: a novel n-gram matching technique for cross- and monolingual word form variants. *Information Research* 7 (2), available from http://InformationR.net/ir/7-2/paper126.html.

Pirkola, A., Toivonen, J., Keskustalo, H., Visala, K., & Järvelin, K. (2003). Fuzzy Translation of Cross-Lingual Spelling Variants. In: *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 345–352). New York: ACM Press.

Qu, Y., Grefenstette, G., & Evans, D. A. (2003). Automatic Transliteration for Japanese-to-English Text Retrieval. In: *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 353–360). New York: ACM Press.

Robertson, A. M., & Willett, P. (1998). Applications of n-grams in textual information systems. *Journal of Documentation* 54 (1), 48–69.

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Reading, Mass.: Addison-Wesley.

Stalls, B. G., & Knight, K. (1998). Translating Names and Technical Terms in Arabic Text. In: *Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages* (pp. 34–41), Montreal: ACL.

Zobel, J., & Dart, P. (1995). Finding Approximate Matches in Large Lexicons. *Software — Practice and Experience* 25 (3), 331–345.

Zobel, J., & Dart, P. (1996). Phonetic String Matching: Lessons from Information Retrieval. In: *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 166–172). New York: ACM Press.

Table 2
Precision (%) of the combined TRT and fuzzy *Finnish*-to-English matching.
High Confidence Factor.

Term type	**Digrams**	**TRT & digrams**	**change (%)**	**Trigrams**	**TRT & trigrams**	**change (%)**
Bio terms, n=90	61.4	72.0	+17.3	49.2	65.0	+32.1
Place names, n=55	30.0	35.9	+19.7	29.3	33.6	+14.7
Economics, n=31	32.2	38.0	+18.0	30.7	41.0	+33.6
Technology, n=36	31.6	53.7	+69.9	21.2	50.2	+136.8
Miscellaneous, n=59	33.8	40.6	+20.1	28.9	36.1	+24.9

Table 3
Precision (%) of the combined TRT and fuzzy *French*-to-English matching.
High Confidence Factor.

Term type	**Digrams**	**TRT & digrams**	**change (%)**	**Trigrams**	**TRT & trigrams**	**change (%)**
Bio terms, n=90	88.3	89.7	+1.6	87.7	88.9	+1.4
Place names, n=55	52.5	51.7	-1.5	53.3	52.4	-1.9
Economics, n=31	80.1	77.0	-3.9	78.1	75.0	-4.0
Technology, n=36	78.4	83.3	+6.3	78.8	83.7	+6.2
Miscellaneous, n=59	64.6	66.1	+2.3	64.7	66.3	+2.5

Table 4
Precision (%) of the combined TRT and fuzzy *German*-to-English matching.
High Confidence Factor.

Term type	**Digrams**	**TRT & digrams**	**change (%)**	**Trigrams**	**TRT & trigrams**	**change (%)**
Bio terms, n=90	77.8	86.9	+11.7	75.7	86.8	+14.7
Place names, n=55	41.4	48.8	+17.9	42.7	49.2	+15.2
Economics, n=31	52.6	60.0	+14.1	52.3	60.8	+16.3
Technology, n=36	60.5	72.3	+19.5	58.3	68.1	+16.8
Miscellaneous, n=59	54.0	56.2	+4.1	52.3	56.8	+8.6

Table 5
Precision (%) of the combined TRT and fuzzy *Spanish*-to-English matching.
High Confidence Factor.

Term type	**Digrams**	**TRT & digrams**	**change (%)**	**Trigrams**	**TRT & trigrams**	**change (%)**
Bio terms, n=90	67.6	73.7	+9.0	63.2	69.4	+9.8
Place names, n=55	55.5	55.5	+-0.0	54.9	54.9	+-0.0
Economics, n=31	45.5	49.2	+8.1	45.9	45.6	-0.7
Technology, n=36	57.5	61.7	+7.3	57.9	61.9	+6.9
Miscellaneous, n=59	41.1	42.1	+2.4	41.9	41.6	-0.7

Table 6
Precision (%) of the combined TRT and fuzzy *Finnish*-to-English matching.
Low Confidence Factor.

Term type	**Digrams**	**TRT & digrams**	**change (%)**	**Trigrams**	**TRT & trigrams**	**change (%)**
Bio terms, n=90	61.4	75.6	+23.1	49.2	68.3	+38.8
Place names, n=55	30.0	34.5	+15.0	29.3	34.1	+16.4
Economics, n=31	32.2	35.3	+9.6	30.7	38.5	+25.4
Technology, n=36	31.6	53.2	+68.4	21.2	51.0	+140.6
Miscellaneous, n=59	33.8	38.2	+13.0	28.9	33.1	+14.5

Table 7
Precision (%) of the combined TRT and fuzzy *French*-to-English matching.
Low Confidence Factor.

Term type	**Digrams**	**TRT & digrams**	**change (%)**	**Trigrams**	**TRT & trigrams**	**change (%)**
Bio terms, n=90	88.3	94.2	+6.7	87.7	94.1	+7.3
Place names, n=55	52.5	58.5	+11.4	53.3	57.6	+8.1
Economics, n=31	80.1	79.5	-0.7	78.1	77.6	-0.6
Technology, n=36	78.4	85.4	+8.9	78.8	85.5	+8.5
Miscellaneous, n=59	64.6	65.9	+2.0	64.7	66.6	+2.9

Table 8
Precision (%) of the combined TRT and fuzzy *German*-to-English matching.
Low Confidence Factor.

Term type	**Digrams**	**TRT & digrams**	**change (%)**	**Trigrams**	**TRT & trigrams**	**change (%)**
Bio terms, n=90	77.8	92.3	+18.6	75.7	92.2	+21.8
Place names, n=55	41.4	51.7	+24.9	42.7	51.3	+20.1
Economics, n=31	52.6	57.5	+9.3	52.3	58.9	+12.6
Technology, n=36	60.5	72.1	+19.2	58.3	70.1	+20.2
Miscellaneous, n=59	54.0	52.9	-2.0	52.3	52.5	+0.3

Table 9
Precision (%) of the combined TRT and fuzzy *Spanish*-to-English matching.
Low Confidence Factor.

Term type	**Digrams**	**TRT & digrams**	**change (%)**	**Trigrams**	**TRT & trigrams**	**change (%)**
Bio terms, n=90	67.6	81.1	+20.0	63.2	80.9	+28.0
Place names, n=55	55.5	55.6	+0.2	54.9	54.4	-0.9
Economics, n=31	45.5	49.2	+8.1	45.9	49.5	+7.8
Technology, n=36	57.5	63.9	+11.1	57.9	64.0	+10.5
Miscellaneous, n=59	41.1	46.5	+13.1	41.9	48.4	+15.5

Table 10
Translation recall (%) and precision (%) for *Finnish*-to-English TRT.

	CF=50% Fr.=50 NoR=18	**CF=10% Fr.=50 NoR=25**	**CF=10% Fr.=10 NoR=169**	**CF=2.0% Fr.=4 NoR=331**
Recall				
Bio terms	8.9	14.4	46.7	62.3
Place names	16.4	18.2	23.6	23.6
Economics	12.9	12.9	29.0	39.7
Technology	11.1	11.1	30.6	33.3
Miscellaneous	8.5	8.5	18.6	23.7
Precision				
Bio terms	5.1	3.9	5.7	3.5
Place names	13.0	11.4	10.7	7.0
Economics	7.5	4.1	4.0	2.8
Technology	5.6	2.8	3.8	1.6
Miscellaneous	6.4	3.4	4.3	2.7

Table 11
Translation recall (%) and precision (%) for *French*-to-English TRT.

	CF=50% Fr.=50 NoR=8	**CF=10% Fr.=50 NoR=26**	**CF=10% Fr.=10 NoR=151**	**CF=2.0% Fr.=4 NoR=525**
Recall				
Bio terms	46.7	76.7	84.4	92.2
Place names	29.1	38.2	40.0	41.8
Economics	41.9	58.1	64.5	64.5
Technology	77.7	83.3	83.3	83.3
Miscellaneous	37.3	50.8	55.9	62.7
Precision				
Bio terms	44.2	30.1	25.8	9.3
Place names	26.2	20.6	18.0	6.1
Economics	40.6	40.0	32.8	6.6
Technology	70.0	50.8	38.0	9.9
Miscellaneous	35.5	24.6	20.5	6.0

Table 12
Translation recall (%) and precision (%) for *German*-to-English TRT.

	CF=50% Fr.=50 NoR=21	**CF=10% Fr.=50 NoR=33**	**CF=10% Fr.=10 NoR=162**	**CF=2.0% Fr.=4 NoR=472**
Recall				
Bio terms	37.7	72.2	85.6	97.8
Place names	27.3	34.5	34.5	38.2
Economics	32.2	35.5	58.1	58.1
Technology	44.4	47.2	80.6	88.9
Miscellaneous	35.6	39.0	49.2	54.2
Precision				
Bio terms	23.9	25.8	23.3	12.8
Place names	20.8	12.4	10.4	5.0
Economics	20.4	14.5	16.7	8.7
Technology	32.0	22.7	23.4	11.5
Miscellaneous	29.6	20.0	18.6	9.8

Table 13
Translation recall (%) and precision (%) for *Spanish*-to-English TRT.

	CF=50% Fr.=50 NoR=18	**CF=10% Fr.=50 NoR=49**	**CF=10% Fr.=10 NoR=226**	**CF=2.0% Fr.=4 NoR=692**
Recall				
Bio terms	22.2	42.2	72.2	94.4
Place names	29.1	29.1	32.7	34.5
Economics	25.8	41.9	48.4	51.6
Technology	36.1	44.4	63.9	77.8
Miscellaneous	8.1	30.5	37.3	49.2
Precision				
Bio terms	16.0	8.1	11.1	2.5
Place names	27.1	20.3	13.2	2.4
Economics	18.6	16.9	14.4	2.4
Technology	29.5	19.5	14.7	3.0
Miscellaneous	11.0	14.3	10.9	2.0

Table 14
Error analysis of Finnish-to-English and French-to-English economy word pairs ($n = 31$). CF=2.0%, Fr.=4.

Finnish-to-English			**French-to-English**		
Error type	**Example**	**% of all errors**	**Error type**	**Example**	**% of all errors**
1	tariffi → tarif	52	1	azerbaidjan → azerbaijan	50
2	kansleri → chancellor	11	2	poivre → pepper	10
3	koodi → code	37	3	bureaucratie → bureaucracy	40