DE GRUYTER OPEN

DOI 10.1515/apd-2015-0008

# Translation memory and computer assisted translation tool for medieval texts

Attila Törcsvári, Ars Ensis

**Abstract – Translation memories (TMs), as part of Computer Assisted Translation (CAT) tools, support translators reusing portions of formerly translated text. Fencing books are good candidates for using TMs due to the high number of repeated terms. Medieval texts suffer a number of drawbacks that make hard even "simple" rewording to the modern version of the same language. The analyzed difficulties are: lack of systematic spelling, unusual word orders and typos in the original. A hypothesis is made and verified that even simple modernization increases legibility and it is feasible, also it is worthwhile to apply translation memories due to the numerous and even extremely long repeated terms. Therefore, methods and algorithms are presented 1. for automated transcription of medieval texts (when a limited training set is available), and 2. collection of repeated patterns. The efficiency of the algorithms is analyzed for recall and precision.**

**Keywords – natural language processing, translation memories, computer assisted translation**

## I. INTRODUCTION

Reconstruction of the meaning of medieval texts, especially codices, treaties and *Hausbuch*s, often available as manuscripts, is always a challenging task from a number of aspects; the transliteration of handwriting, the specialties of the local dialect, yet non-standardized spelling, simple typos and colloquial style—so to say, *syntactical difficulties*—are all obvious obstacles that precede in order and relevance the final aim: the interpretation of the content itself; in the actual case, understanding, physical testing and using in trainings and in practice the actions and techniques described in *Fechtbuch*s. This interpretation is, besides considering media-rich content, such as video trainings, firstly manifested in written form: either a translation to the modern version of the same language—called *modernization* in this paper—in which the text was originally written, or in translation to another language.

However, producing this "written form", even the modernization, is not free from an interpretative attitude of the experts of the field, right because of the wish and best will of the transcriptor to provide an understandable text for the benefit of the readers, who are not expected to make all the effort of resolving certain issues in the original. An unquestionably important merit of this interpretative attitude during transcription is, indeed, a kind of translation: replacement of obsolete terms to their contemporary counterpart.

Another challenge during both modernization and translation is to achieve a certain *consistency* so that the same terms and expressions of the original would be represented in the same way in the transcripted or translated text, at least, whereas the context allows.

The efforts to analyze and propose possible solutions for supporting modernization and translation were not made without practical reasons; we have kept in mind the primary goal of translating Johannes Lecküchner's "*künst vnd zedel ym messer*" [i] *(Lecküchner [1482])*, "The Art of Messer Fencing", Cgm 582 to Hungarian. This fencing manual was completed in 1482, as a beautifully illustrated manuscript, and based on a former manuscript of the same author. The text was transcribed and published by Carsten und Julia Lorbeer, Johann Heim, Robert Brunner und Alexander Kiermayer, under *http://www.pragmatische-schriftlichkeit.de/cgm582.html* [ii] *(Lorbeer et al [2006])*, and used with the permission of the authors.[1]

In this paper I present a method of automation of modernization of German medieval texts to contemporary spelling and vocabulary, and presenting techniques to reduce translation work and achieve consistency of translation by using translation memories.

## II. ANALYSIS OF THE CORPUS, THE CURRENT TRANSLATION PRACTICE AND STATISTICS

A set of well-known *Fechtbuch*s were analyzed to see the feasibility and possible benefits of using a computer assisted approach of transcription and translation.

### 1. Analysis on effect of modernization

Our primary target was the translation of the original Early New High German [ENHG] text.

The transcription was made by a team of researchers, as mentioned above. They have, used *"a computer aided approach to find transcription errors by counting and finding all variations of all words in the text"*. The scientific version of the transcription published in [ii] *(Lorbeer et al [2006])* contains all the notations and clarifications made on the original text, with highest respect not only to the original, but also the pronunciation, usual spelling and abbreviations at that time.

A considerable part of the text was translated to modern German, published by Falko Fritz[2]:

| | Original | Modernized | Ratio |
|---|---|---|---|
| Pages | 432 | 79 complete pages | 18% |
| Paragraphs | 874 | 259 | 30% |

[1] *"after talking to all co-authors we give you permission to analyze our transcription of cgm582, quote some parts in your scientific article and to translate it to Hungarian."* (Carsten Lorbeer)

[2] *http://www.hammaborg.de/de/transkriptionen/leckuechner_cgm582/index.php*, as downloaded in September, 2012

Though the modern German translation covered about a fair 30% (the most important parts), it was still found that relevant techniques are detailed in the non-modernized part.

In order to estimate the difficulty of reading non-modernized text, a simple test was made with a native German speaker trained in proofing and checking documents under various conditions.

Similar size sections were selected from the translated and non-translated part (with notations removed, but additions provided by the translators kept). To measure the effect of getting used to the spelling and learning the vocabulary of the text, in both cases a training page was given to the reader. As a third test, a piece of text was manually modernized. The time required for simple reading was measured.

| | Translated | Original | Manually modernized |
|---|---|---|---|
| Training page | 1'20" | 2'03" (~150%) | 1'44 (~125%) |
| Test page | 1'23" | 1'46" (~125%) | 1'20" (≈) |

From the tests we have concluded, that, it causes, as expected, measurable difficulties (+25%) for the reader to interpret the spelling and vocabulary of the original 15th century text, even after training.

A more interesting test pointed out that the manually modernized text required about the same speed as reading the translated text, at least after the training.

## 2. Translation

### 2.1 Modernization and translation issues detected in well-known texts are given below as examples

1. Peter von Danzig, Longsword[3]

| | … mit dem rechten fuess… | …mit dem rechten Fuß… | 5 |
|---|---|---|---|
| 20 v | …vnd spring **mit dem rechten fuess** hinder seinen lincken füeß… | …Spring **mit deinem rechten** hinter seinen linken Fuß… | 1 |

In the above case, the original text seems containing an overbroad word—but *may be considered*[4] more accurate than the modernized version.

2. Joachim Meyer, Longsword, „*Gründtliche Beschreibung*…", ed. 1570[5]

| VIrv | **Ochs** | **Ox** |
|---|---|---|

[3] http://www.hammaborg.de/en/transkriptionen/peter_von_danzig/02_langes_schwert.php, as downloaded in September, 2012

[4] Personal interpretation of the author of this article, marked with Italics in this article.

[5] http://wiktenauer.com/wiki/Joachim_Me%C3%BFer/Longsword, as downloaded in September, 2012

| | | |
|---|---|---|
| | …Zum Lincken Ochsen schick dich disem zugegen / nemlich **trit mit dem Rechten Fuß vor…** | …For the Left Ox reverse this, namely **stand with your Right Foot forward…** |
| XVIv | **Vom versetzen ein nützliche vermanung** Schick dich in die Zornhut / wirt denn auff dich von Oben her gehauwen / so **trit mit dem Rechten fuß** | **Of Displacing, a useful concept** Place yourself into the Wrathful Guard, if you are then struck from above, then **step with the right foot forward…** |
| XXrv | **Zirckel** …wischt er als dann mit den Armen undersich dem Schwerdt nach / so **trit mit dem Rechten fuß** wol beseits auff sein Rechte seiten… | **Circle** …the sword thus clips him with your arms under yourself, then **step with the right foot** to take on his right side… |

It is clear, that the translator used different translations of the word "*trit*" (*step*) for a good reason: the first case the translator took a static concept (*stand*), since speaking about a stance, while in the second used a motion verb (*step*), expressing the movement of the foot.

Therefore, it is obviously not a translation mistake, to use "stand" instead of "step".

However, even for a stance, to reach the proper position from the previously described *Right Ox*, one must, indeed, make a step. Taking in consideration the *training concept*, that seemed the original intention of the esteemed Author according to the Introduction[6], it *may be* a more appropriate translation to take a "step" rather than "standing" with right foot forward.

These cases are not at all translation mistakes, but *can be considered* as immediate interpretations during translation, either "undertranslations" or "overtranslations".

Judging all such particular cases, if discovered at all, takes some time for the reader. Naturally, this time cannot be measured in any way to the time spared by the translators providing us an already digested content. However, it may be more faithful to the original providing a translated version that is consistent or inconsistent to the extent of the original—or, at least, applying necessary marks in the translated version.

## 3. Statistics

A simple statistic was made for estimating the possible reuse ratio of terms in the translated part (until page XL) of Meyer's *Fechtbuch*. The terms were selected by removing a minimal number of German stopwords (just definite and indefinite articles).

The file size was about 55 kB, there were about 350 independent terms found occurring more than once—the maximal occurrence was 19 for "*gegen seiner Lincken*", and,

6 *"… from your clarity attain and exude the proper judgement in Stance and Strikes so that Youth will not have to learn this art unguided because of your unspoken word…" /*
*"…wie sie soll auß den erklerten häuen und Legern ins werck gericht werden / auff das nit allein die Jugend so sich auff solche kunst zubegeben willens / durch solche inen unbekandte wort…"* (translated by Mike Rasmusson)

intrestingly[7], 10 times only "*gegen seiner Rechten*"—resulting in about 8 kB sparing when translating them only once, which is about **14% of the original**. In the subject corpus, there were also some extremely long repeated n-grams detected, composed of 12 consecutive words.

This was according to my assumptions—due to the narrow scope of the *Fechtbuch*s and the disciplined wording of the author.

### 4. Conclusion

From the statistics of modernization one can deduct, that providing a modernized version decreases reading time. In the subsequent chapter it is presented, that *automation of modernization* is feasible.

The above drawn (minor) translation inconsistencies and also considerable translation, or at least, typing work can be supported by *translation memories*.

## III. AUTOMATING MODERNIZATION

The German original, given in early new high German, looks somewhat unfamiliar to contemporary readers—and also for computers. As an average, about 50% (19k words from 38k words) are reported as spell errors.[8]

The baseline translation from the original to English, using Google Trans, resulted, as expected, a poor translation: about 32% (279 of 860, using the chapter about *Zornhau*) of the words were not found. As a comparison, the manually modernized version contains about 3% of the words that could not be translated to English by Google Trans.

The above are not surprising, because of basic and also less obvious spelling issues.

At the first sight, many of the problems can be resolved by simple word-by-word, or, at most, some pattern based replacement:

| vechten | fechten | vnd | und |
|---|---|---|---|
| yn | ihn | seytten | seiten |

As seen above, a simple word-to-word rewriting procedure, when a dictionary is available, will provide a simpler to understand text, assuming, that such a dictionary can be either obtained, created from scratch or built.

Applying manual translation caused nearly 5 minutes per page when producing the test samples. Automation seemed therefore necessary.

---

[7] Besides the mere statistical fact, the latter finding is very important for a fencer, since it points out a main characteristics of Meyer's school of fence. However, it may worth a detailed study to compare the various *Fechtbuch*s from this aspect.

[8] Using Microsoft Word German (Germany) spell checker.

## 1. Use of an existing ENHG dictionary

I have learnt, that our effort to create such a transcription dictionary is not unique, as an extensive work is presented in [iii] *(West [2008])*. and also available as online application[9]. Unfortunately no downloadable or reusable dictionaries found for translating ENHG to modern German.

## 2. Creating a dictionary: the manual way

Due to the number of the unique words and also internal inconsistencies of the author, the *manual approach* of constructing a dictionary looks infeasible. Also, applying just a sequence of ad-hoc rewrite rules (global search-replaces) may not achieve a minimum quality, because the replacements may be interfering. Therefore, an automated solution is demanded.

Due to the difficulty of manual POS tagging the complete text, having e.g. poems, short instructions etc., we have checked *statistical methods* for the creation of the dictionary.

## 3. Studies of best practice

The research on automated modernization reported 60-80% precision on general corpus in Early New High German, as discussed in [iv] *(Bollmann et al [2012])*, using statistical methods. (Precision is measured on correct modernization vs. total number of tokens.)

For our limited corpus, we expect higher precision, applying the techniques as described below.

## 4. Method of constructing a dictionary

Therefore, the only way to solve the limited dictionary can be performed by creation of the dictionary ourselves.

The outline of building such a dictionary is as follows:

1. preparating *bitexts* from available corpus for the purpose of training

2. constructing a set of *rewrite rules* applicable to both the original ENHG and modern German texts, to find "cognate" (differently spelled) couples,

3. using statistical dictionary building algorithms (building a dictionary form training documents), taking also benefit of the cognate computation,

4. proposing translations for words that are not in the training set, based on the rewrite rules and a valid list of modern German words.

### 4.1 Preparation of bitexts

I have considered using *Fechtbuch*s for training the dictionary is the obvious choice, since many of them are already completely modernized.

[9] http://www.woerterbuch-portal.de/woebus_alle/Woebu21

However, since the changes in spelling were considerable during the period from which there are various *Fechtbuch*s are available, I have chosen to limit the dictionary building to the actual text only. The partially modernized version seemed sufficient, covering about 30% of the text.

The bitexts were built by manual alignment of sentences and sub-sentences, providing punctuation characters in the original, based on the modernized text. (Punctuation is given in curly braces.)

| | |
|---|---|
| So der meister das vechten des messers yn dy stuck geteylt hat vnd eyn ytlichs mit namen genent**{,}** nw hebt er an ze sagen von dem ersten glid der tailung**{,}** als von dem zorenhaw**{.}** vnd ist zw wissen**{,}** das der zorenhaw mit dem ortt bricht all öberhaw**{,}** vnd ist doch eyn schlechter pawren schlag | Nachdem der Meister das Messerfechten in die Stücke eingeteilt hat und ein jedes mit Namen benannt, beginnt nun die Rede von dem ersten Punkt der Aufzählung, und das ist der Zornhau. Davon ist zu wissen, dass der Zornhau mit dem Ort alle Haue von oben bricht, und sei er auch ein einfacher Bauernschlag. |

The translation followed rather faithfully the original; the alignment was checked and found less than 1% of paragraphs with misalignments.

### 4.2 Constructing rewrite rules

A series of rule-sets were created, with decreasing reliability. The series of rule-sets were applied in the order given below, to achieve the highest possible accuracy.

Each rule set contained separate rules for the ENHG and modern German words. A single rule is composed of either a

- hierarchic branch of further rules (so that interference of rules could be minimized), or
- an atomic rule, in a form of a regular expression, that actually describes the rewriting.

The rewrite rules, for the purpose of performance, can be specified as conditional rules. Each rewrite rule can be either

- a *populating rule*, creating a new version but leaving the original word as alternative, or an
- *overwrite* rule*,* changing the original word.

All matching rules are applied in sequence, thus producing from one word multiple alternatives, when populating rules were applied. Each alternative is marked with the minimal number of rules applied to reach the alternative from the original word.

The basic function of the atomic rules were *not* to construct the modern spelling of the given word; instead, the same kind of rules were applied on the words of both the source and target sentences, resulting in a set of possible rewrites for each word, and leaving finding the closest couples for the statistical dictionary generation algorithm using the number of

rewrites as a proximity measure[10]. The atomic rules, indeed, provided *hypothetic and merely artificial modern pronunciation alternatives*, based on well-known phonological changes[11], and discovered inconsistencies in spelling.

It is also important to mention, that the produced word forms were *not tested against valid words in a general German dictionary* when there were any bitexts in which the word appeared, but against the words appearing in the coupled modern German translation.

The rewrite rule sets were applied in the following sequence, computing the word couples for each.

*1. No rewriting and case insensitive rewriting rule*

This *id* rewrite rule supports finding equal words.

A case insensitive rewrite rule is also added, that couples words if their lower case version is the same.

*2. Common consonant rewriting rules, e.g.*

| **Populating rules** | **Overwriting rules** | |
|---|---|---|
| z? tzt,tzd,zt,tz,z ➔ C | ß ➔ s | v ➔ f, |
| k? ck ➔ K | ss ➔ s, tt ➔ t, nn➔n, | p➔b |
| m? mb,mp,mm ➔ m | rr➔ r, | [td]+ ➔ t |
| | ll➔l, ff➔f | |

For example, producing a couple with distance **1**:

| *bekandt* | *bekannt* |
|---|---|
| 0: bekandt | **0: bekant** |
| **1: bekannt** | |

*3. Pronunciation rewrite rules*

The original text represented the manuscript with proper accents (macrons), i.e. contained accented vowels and consonants. They were rewritten to their non-accented version—a questionable technique, but the German language, and the limited purpose, allows this.

| **Populating rules** | **Overwriting rules** | |
|---|---|---|
| ū ➔ O, ū ➔ U | ü ➔ U | ō ➔ O |
| ä ➔ a, ä ➔ e | v̈ ➔ U | ẅ ➔ w |
| | ü ➔ U | ö ➔ O |

10 Using the Levenshtein distance of the two words was found less efficient than using the sum of the applied rewrite rules, since this sum is comparable to the number of changes in coding the phonemes.

11 e.g. http://en.wikipedia.org/wiki/Old_High_German#Consonants

A set of further rules were applied to allow coupling words containing v instead of u (e.g. vnd), when v is a vowel position, i.e. not between vowels or at start.

| | |
|---|---|
| ^v([^aeiouy]) ➔ U$1 | ([^aeiouy])v([^aeiouy]) ➔ $1U$2 |

A set of rules were applied for affricate and fricative coding:

| **Populating rules** | | |
|---|---|---|
| ck ➔ CH | ph ➔ F | pf ➔ F |
| | uu ➔ UF | |
| | ([ieoauAEIOU])h ➔ $1 (unsouded h) | |

Due to the great variability in the use of y and unsounded h for denoting various diphthongs or long counterparts of vowels, that are differently spelled in modern German, a set of complex rules were applied:

| **Overwriting rules** | **Populating rules** |
|---|---|
| ye ➔ ie | ey ➔ EI, ey ➔ AI |
| ay ➔ ai | y ➔ IE, y ➔ I |
| ([ieoauAEIOU])h ➔ $1 | |
| rh ➔ r, hr ➔ r | |

All common rules were also applied.

A pair of words coupled by these more complex rule:

| | | | |
|---|---|---|---|
| scharphen | scharfen | gest | gehst |
| gefahren | Nimm | laufft | Läuft |

*4. Stemming rules*

Unfortunately the modern German version sometimes presented the words in different case or otherwise inflected differently.

Therefore, an obvious stemming was implemented as rewrite rules, changing usual affixes to the expectedly simpler form.

For brevity, only a few samples are given, all as populating rules:

| | |
|---|---|
| ([^ ][^ ])e[nsm]$ ➔ $1 | removal of en,es,em at end |
| ([^ ][^ ])e[nsm]$ ➔ $1e | adding e instead of en,es,em |
| ([^ ][^ ])es$ | replacing es by s |

Further rules were applied for prefixes, e.g.

| | |
|---|---|
| ^ver(.....) ➔ ge$1 | ^be(.....) ➔ ge$1 |

*5. Spell mistake rewrite rules*

If all the former failed to find a couple for a word, a set of exceptional rules were given for cases not found previously, all as populating rules, e.g.:

| | |
|---|---|
| h ➔ | total removal of unsound h |
| zwiua ➔ CWEIFA | |
| sw ➔ SCHW | s+consonant case |
| a➔e, u➔l, u➔o | reasonable in various context of consonants |

The above rules were defined in configurable XML files for the software.

*6. Exception dictionary*

After checking the output, a few manual exceptional translations are defined that were not coupled by any of the preceding rewrite rules.

*7. Separation or melding words*

There were typical cases found, when words are melt in the original corpus or separated by space, for example *vor rede* vs. *vorrede*, *wiltu* vs. *willst du*, i.e. there are found one-to-two or two-to-one cases.

Dictionary builder algorithms often use the one word-to-one word. assumption (see, for example, [v] *(Melamed [1996])*, extensively discussed below). This approach can be transformed easily to one token to one token, whereas a token can be provided as merging two (or even more) words).

In order to accommodate our dictionary builder to come over this shortcoming, and at the same time give extra recognition capability for the above cases, a set of multi-word rewrite rules were applied, applicable only between word boundaries, e.g.

| | |
|---|---|
| (..)est[ ]➔ $1st\x20 | est replaced to st on word boundary only |

**4.3 A statistical dictionary building algorithm**

The primary task of the dictionary building algorithm to produce couples of words that are translations of each other, basically based on their cooccurrence in bitexts.

Though it is usually assumed by dictionary builders, it cannot be expected that one word has exactly and only one translation and vice versa. It is also possible, that a sequence of words form a token that is translated to a single word or also a sequence of words, as described above. Therefore I will discuss below coupling tokens, as sequence of words.

Besides the cooccurrence of token couples, the distance of the physical appearance of the tokens, i.e. their rewritten form, may be used to increase the likelihood for those couples, which really match.

I propose below, and also implemented an algorithm that was tested in building a dictionary for our text.

*1. Cooccurrence vs. likelihood of matching*

A naive approach in dictionary building is based on mere cooccurrence statistics of words in bitexts. The basic idea behind: the more times a word cooccurs with a

translation, the higher the likelihood that they are translations of each other. Finding most probable couples and removing from the sentences will retrieve the most probable couples with decreasing accuracy.

This method is partially described in [v] *(Melamed [1996])*.

In case of Indo-European languages, especially in modernization, a much finer alignment of the possible couple can be provided, since the order of the words in sentences is rather the same. (This is the linearity assumption described in [vi] *(Melamed [1997] p. 306)*). This holds especially for texts with short, declarative sentences (e.g. technical texts, description of constituents in a recipe, etc.), noun forms, or even poems given in the *Fechtbuch*s.

The basic scheme for such a coupling is, computing, instead of cooccurrence, a probability of matching based on the hashed distance of the center of the token in a sequence to the target sentence:

| So | haw | Im | von | deyner | rechten | achsel | von | oben | lanck | eyn |
|---|---|---|---|---|---|---|---|---|---|---|
| 4.5 | 13.6 | 22.7 | 31.8 | 40.9 | 50 | 59.1 | 68.2 | 77.3 | 86.4 | 95.5 |

likelihood[12] of matching ($\pi_k$)

| dann | schlag | lang | von | diener | rechten | Schulter | auf | ihn | ein |
|---|---|---|---|---|---|---|---|---|---|
| 5 | 15 | 25 | 35 | 45 | 55 | 65 | 75 | 85 | 95 |

(The second line shows in % the position of the middle of the word in the list of words.)

In the above case, for achsel: deiner, rechten, Schulter and auf is selected, giving highest rank to rechten. However, having more sentences for the cooccurrence of achsel and Schulter, at any position near to each other, the dictionary builder will sum the likelihood values and Schulter will be the winner.

The function that describes the likelihood of matching, i.e. the proximity of a certain couple, the following function was used:

source sentence $s = \underline{v}$, $m = |v|$, and
target sentence $t = \underline{w}$, $n = |w|$

For the $i$th source word in $s$, $i \in [0..m)$

$$c(i) := \frac{(i+0.5)}{m} * n$$

(the expected centroid in the target sentence).

12 The figure is somewhat simplified, presenting a linear function; however, we have chosen a non-linear function during the evaluation as given below.

For a given expected couple *j* in *t, i* ∈ [0..m),

$$d(i,j) := j - c(i)$$

(the distance of the centroid from the expected couple.)

Then the proximity of matching *i* and *j,* and an environment threshold *k:*

$$\pi_k(i,j) := \pi_k'(d(i,j))$$

where

$$\pi_k'(d) = \begin{cases} 0 & |d| > k \\ (1 - \frac{|d|}{k}) & |d| <= k \end{cases}$$

This function represents a similar approach—to increase the probability of matching proper couples—to the linearity assumption in [vi] (Melamed [1997], p. 306), with significantly less computation complexity, and less parameterization than the least-square method computation.

A linear $\pi'_k$ is also tested but found producing less precise results.

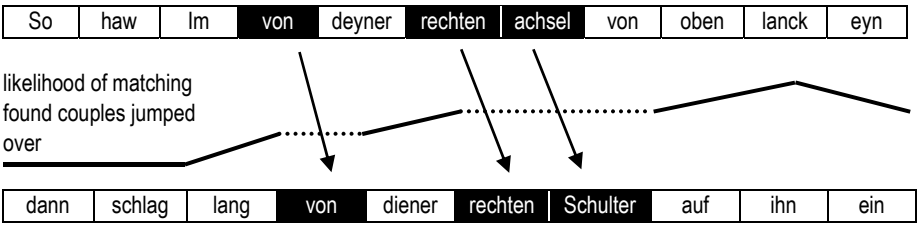
*2. Using the phonological rules for words*

Since our task is to find translations that are often just cognates, certain couples can be excluded automatically, i.e. those where there seems no common rewritten form as discussed in point *4.2* above. However, in order to further increase accuracy, the above proximity can be simply divided by the proximity of the word forms in the source and target.

*3. A greedy method vs. least-square optimization method*

Due to the very limited number of samples, use of a sophisticated optimization. as discussed in [vi] (Melamed [1997]) and single-pass evaluation of matching couples was not found necessary.

Instead, after finding some couples, we have recomputed the proximity, using the found couples as anchors; for example, using the "*id*" rewrite rule first, the coupling the sentences is significantly easier:




Naturally, the centroid is computed for the actual segment, allowing crossing of segment boundaries. With an appropriate selection of the parameter *k* for maximum

distance (and subtracting the number of matching couples in between), in the above example, lanck and lang still could be coupled.

Therefore, we have applied the following method:

*for each* cognate-matching method (i.e. rule-set) in decreasing accuracy sequence as given in point *4.2*,

> *repeat* finding non-interfering couples of tokens whose for a summed proximity is above a threshold *until* any found
>
> > use the already found couples as anchors for segment boundaries
> >
> > recompute sum of proximity values

The presented algorithm not only makes more precise the subsequent calculations, but also allows finding n:m couples, i.e. when the same word of the original is coupled to many target words and vice versa.

It is obvious also, that further duplication of the coupling step first with a small environment threshold (*k*) and with a larger one, will further decrease the noise. Not providing an extremely large environment threshold, the accuracy, in comparison to the sentence-wise cooccurrence method, remains this way controllable.

*4. expressions: dealing with separation and melding of words*

The *Fechtbuch*s, as pointed out earlier, contain a number of repeated expressions, and also a number of situations, when in the ENHG spelling of a term the author separated two words with a space whereas in modern German the term is written in one word, or when words (or their cognates) are simply swapped.

It would be advantageous for disambiguation and even for finding terms to use a Hidden Markov Model (HMM), as proposed in [vii] (Vogel et al [1996])—however, it will not solve the spacing problems. It was also obvious, that we cannot expect the HMM efficiently working on a few hundred matched expressions only.

Therefore, since having a statistical aligner algorithm anyway, we have chosen a four-step method; instead of finding the couple of a single word only:

1. composing double-word tokens from couples of sequential words in both the source and target sentences (sorting alphabetically consecutive words and applying the rewrite rules)

2. coupling, in separate steps double-word tokens to double-word tokens, single-word tokens to double-word tokens, double-word tokens to single-word tokens and single word tokens to single-word tokens.

This approach is also more permissive than the one-to-one assumption given in point 3 of [v] (Melamed [1996]).

However, the approach could be further refined collecting n-grams from the source and target texts, as described in the subsequent point.

### 4.4 Fallback scenario for words not appearing in bitexts

However, there are some words in the non-translated sentences, that are not valid modern German words, and also not found in any bitexts.

The only choice we had, to apply the rewrite rules in decreasing reliability and finding in an external German dictionary[13] a word that seems a cognate. From the possible alternatives, the same minimum distance meth

Not surprisingly, this method may lead to a number of inaccurate couples.

## 5. Evaluation of the results

After creation of the possible couples, a "best" translation was computed, giving some additional preference to those translations that were found more than once.

| Word couple vs. single word | | 2 to 2 | 2 to 1 | 1 to 2 | 1 to 1 | Ratio |
|---|---|---|---|---|---|---|
| **Where bitext was available…** | | | | | | |
| | **1261** | | | | | **46%** |
| Equivalent words | 159 | | | | | 6% |
| Lower-case equivalence | 49 | | 7 | | 42 | 2% |
| Simple rewrite rules | 138 | 28 | 10 | | 100 | 5% |
| Complex rewrite rules | 286 | 145 | 12 | | 129 | 10% |
| Stemming rules | 235 | 77 | 17 | 17 | 124 | 8.5% |
| Spelling | 39 | 7 | 8 | 1 | 23 | 1.5% |
| No proximity thresold | 267 | | | | 267 | 9.5% |
| Non-cognate sagt➔spricht | 29 | | | | | 1% |
| Self-dictionary | 59 | | 2 | | 57 | 2% |
| **Where translation was selected from word list…** | | | | | | |
| | **1485** | not applicable | | | | **54%** |
| Equivalent | 262 | | | | | 9.5% |
| Lower-case equivalence | 73 | | | | | 2.7% |
| Simple rewrite rules | 216 | | | | | 7.9% |
| Complex rewrite rules | 322 | | | | | 11.7% |
| Spelling | 343 | | | | | 12.5% |
| Stemming rules | 8 | | | | | 0.3% |
| None | 261 | | | | | 9.5% |
| **Total** | **2746** | | | | | **100%** |

[13] *http://sourceforge.net/projects/germandict/files/*

## 6. Precision and recall

### 6.1 Estimated precision values

Since the various rewrite rule sets have a certain reliability, I have assigned to each rewrite rule an estimated precision value, in range 0.3-1.

For a couple, the highest rank precision number was first assigned, and slightly increased in case more algorithms found the same couple, to the maximum extent of the next highest rank in order.

For dictionary matches, where there was no bitext to train the matcher, lower precision numbers were used.

The associated precisions are as follows:

| Bitext sentences match | Estimated precision | Dictionary match |
|---|---|---|
| "id" | 1 | id |
| case-insensitive | 0.95 | case-insensitive |
| dictionary rules | 0.9 | |
| common rewriting rules | 0.9 | |
| pronunciation rewrite rules | 0.8 | |
| stemming rules | 0.7 | |
| spelling rules | 0.6 | common rewriting rules |
| spell rewriting rules, any distance | 0.5 | pronunciation rewrite rules |
| proximity rules | 0.4 | stemming rewrite rules |
| | 0.3 | spelling rewrite rules |
| | 0 | no dictionary match. |

### 6.2 Precision evaluation

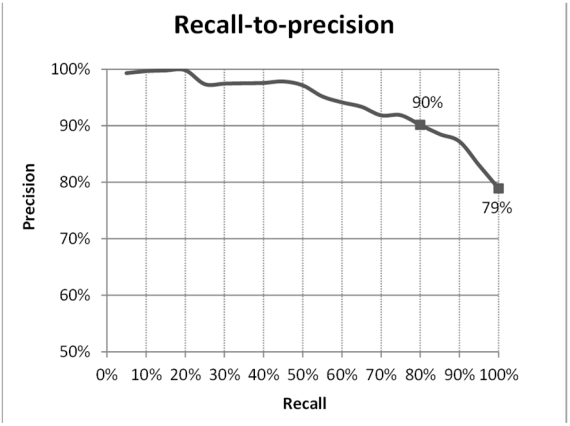
The precision of the ENHG vs. modern German terms was evaluated by a native German translator.

Those words were not marked as mistranslations where the stem of the found word was equal to the original ENHG term, knowing, that this way the automatic translation, at least at a few points, may look pidginized.

The manual verification of the nearly 3000 terms took about one hour only.

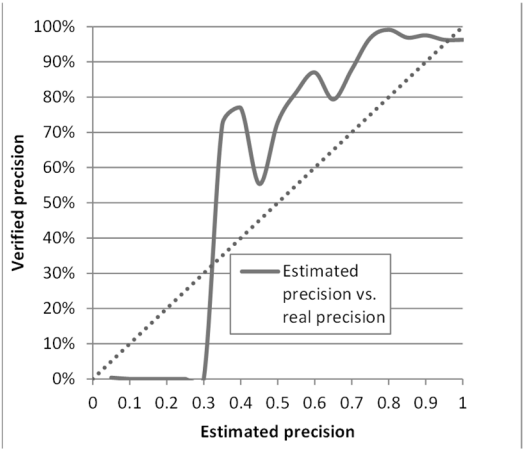
### 6.3 Precision vs. recall

The recall / precision graph is as follows:




At 100% recall of the words and 2-word terms, the precision is at 79%, reaching nearly the top of the similar task in [iv] *(Bollmann [2012])*.

It is also relevant if the estimated precision values correlate to the precision determined by the proofreader:




It can be seen that the estimated precision was pessimistic at realistic values (above 0.3 estimated precision) and became near to 0.98 estimated precision.

## IV. COLLECTING REPEATED "EXPRESSIONS"

The translation of basic expressions and terms already decreases significantly the work of the translator, especially when the feature terms (field-specific terms) are translated and revised by a specialist. However, further reduction of the translation work can be achieved by collecting repeated multi-word patterns.

It is obvious that the found word sequences cannot be considered as valid terms in linguistic sense, however, using POS tagging and building syntax graphs would require overwhelming manual work.

The multi-word patterns—repeated n-grams collected from terms—can be collected using either with *apriori algorithm*, as first introduced in [viii] (Agrawal et al [1996]), or with growing n-gram trees, especially *fp-growth algorithm*, as introduced in [ix] (Han et al [2004]).

## 1. The apriori algorithm

We have deliberately chosen the *apriori algorithm*, that was used successfully previously for building translation memories.

The basic algorithm, briefly, is

1. segmenting the input paragraphs to sentences and sub-expressions whereas possible

2. collecting set of frequent words, ($F_1$)

3. repeatedly scanning input segments,
*for-each n*
collect and count each *c* candidate *n*-gram into $C_n$
using $F_{n-1}$, so that the first *n-1* words and last *n-1* words of *c* must be in $F_{n-1}$.
and collect into $F_n$ only the frequent word sequences in $C_n$
*until* $F_n$ is not empty.

## 2. Refinements of the apriori algorithm

In order to improve the quality of n-grams, in order to detect more natural terms without deeper grammatical analysis, and improve recall and also performance, the general *apriori algorithm* was modified the following way:

1. some words (e.g. articles) are not allowed as first or as last words (e.g. prepositions) of a sequence *(both higher recall and precision)*

2. some words are disregarded inside the sequence (articles) *(higher recall)*

3. segments are split at rare words or rare sequences *(performance improvement)*

4. a tree of included terms is computed, so shorter repeated terms can be reused *(decreased translation time)*

While the first two improvements requires language-specific setup of stopwords, the second two are language-insensitive.

Though the expressions created in the above way are, indeed, not all valid terms in the source language, with the above processing the results are more acceptable.

## 3. Benefit of repeated terms in the corpus

For our text, the *apriori algorithm* found some extremely long repeated terms (12 words):

| | | |
|---|---|---|
| 30r | *alzo* **[stee mit deynem lincken fuß fur vnd halt deyn messer auff deyner rechten]** *achsl* | Steh mit deinem linken Fuß vor und halt dein Messer auf deiner rechten |
| 189v | *also* **[stee mit deynem lincken fuß fur vnd halt deyn messer auff deyner rechten]** *seytten* | (not translated) |

From this example it can be also seen that if a translation is already given for a term, it can be simply reused.

An example when a term appears within another term:

| | |
|---|---|
| 101 r, 153 v | achsel vnd **schreytt mit deynem rechten fuß hinter seynen rechten** |
| 75 v, 88 v, **101 r, 153 v,** 157 r | schreytt mit deynem rechten fuß hinter seynen rechten |

The reuse ratio is surprisingly high:

| | **Size** | **To be translated** | **Spared** |
|---|---|---|---|
| Total size | 221 kB | | |
| After removal of duplicated terms | 131 kB | 59% | 41% |
| Size of terms: | 33 kB, 1900 terms | | |
| Term dictionary translation after removing internal terms | 23 kB | | -10% |
| **Total** | | **69%** | **31%** |

## V. FURTHER WORK

Learning from the example at 4.3, it may worth to revisit finding couples for long expressions; the generic algorithm can be used, however, a new metrics is to be defined that allows coupling of frequently occurred terms with similar word numbers.

A user interface will be developed to support the workflow of the translator.

Once a draft translation is available, we expect that the research group for *Messer* will further refine and annotate the translation. This activity is actually the physical implementation of the techniques described in the corpus, and a reward of the work invested into the translation.

We foresee the reuse of this work, at least the repeated expression collection and translation memory, for translating works of other Masters—not necessarily given in German.

## VI. LITERATURE


**[i]** Hannsen Lecküchner von Nurenberg [1482*] "künst vnd zedel ym messer",* manuscript

**[ii]** Lorbeer , C., Lorbeer, J. – Heim, J., Brunner, R. – Kiermayer, A. [2006] Das ist Herr hannsen Lecküchner von Nurenberg künst vnd zedel ym messer, Wissenschaftliche Fassung mit Kennzeichnung der aufgelösten Abbreviaturen, Transkription der Fechthandschrift cgm582; revised edition January 2006 http://www.pragmatische-schriftlichkeit.de/transkription/trans_cgm582_w_d.pdf

**[iii]** West, J. [2008*]. "Early New High German - English Dictionary"* Electronic edition according to TEI P5, *available at:* http://www.germanstudies.org.uk/enhg_dic/enhg_dic_intro.htm

**[iv]** Bollmann, M. – Dipper, S. – Krasselt, J. – Petran, F. [2012], *"Manual and Semi-automatic Normalization of Historical Spelling — Case Studies from Early New High German"*, 2012, in Proceedings of KONVENS 2012, p. 342-350

**[v]** Melamed, I. D., [1996] *"Automatic Construction of Clean Broad-Coverage Translation Lexicons"*, *in* 2nd Conference of the Association for Machine Translation in the Americas (AMTA), Montreal, PQ

**[vi]** Melamed, I. D. [1997] *"A Portable Algorithm for Mapping Bitext Correspondence"*, *in* 35th Annual Meeting of the Association for Computational Linguistics, p. 305-312

**[vii]** Vogel, S. – Ney, H. – Tillmann. C, [1996] *"HMM-based word alignment in statistical translation"*, in: Proceedings of COLING, pages 836–841.

**[viii]** Agrawal, R. – Mannila, H. –Srikant, R. – Toivonen, H, – Verkamo, A. I. [1996*]"Fast discovery of association rules,"* in Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, pp. 307–328.

**[ix]** Han, J. – Pei, J. – Yin, Y. – Mao, R. [2004] *"Mining frequent patterns without candidate generation: A frequent-pattern tree approach,"* in Data Min. Knowl. Discov., vol. 8, no. 1, pp. 53–87