The Trilingual ALLEGRA Corpus: Presentation and Possible Use for Lexicon Induction

Yves Scherrer, Bruno Cartoni

Department of Linguistics, University of Geneva, Switzerland
{yves.scherrer, bruno.cartoni}@unige.ch

Abstract

In this paper, we present a trilingual parallel corpus for German, Italian and Romansh, a Swiss minority language spoken in the canton of Grisons. The corpus called ALLEGRA contains press releases automatically gathered from the website of the cantonal administration of Grisons. Texts have been preprocessed and aligned with a current state-of-the-art sentence aligner. The corpus is one of the first of its kind, and can be of great interest, particularly for the creation of natural language processing resources and tools for Romansh. We illustrate the use of such a trilingual resource for automatic induction of bilingual lexicons, which is a real challenge for under-represented languages. We induce a bilingual lexicon for German-Romansh by phrase alignment and evaluate the resulting entries with the help of a reference lexicon. We then show that the use of the third language of the corpus – Italian – as a pivot language can improve the precision of the induced lexicon, without loss in terms of quality of the extracted pairs.

Keywords: trilingual parallel corpora, lexicon induction, under-represented languages, Romansh.

1. Introduction

Under-represented languages are a real challenge for Natural Language Processing. The lack of textual and lexical resources make constructing linguistic tools even harder. In this paper, we address this challenge by presenting a trilingual parallel corpus containing an under-represented language, and by using this corpus to induce a bilingual lexicon.

The first part of this paper is dedicated to the trilingual parallel corpus for German, Italian and Romansh, a Swiss minority language. It consists of sentence-aligned press releases from the Grisons cantonal administration. It is called the ALLEGRA corpus.

In the second part, we illustrate the possible use of this corpus for lexicon induction. Indeed, bilingual lexicons are a valuable resource for many Natural Language Processing tasks, but are not readily available for under-represented languages like Romansh. We start by inducing a bilingual lexicon for German-Romansh from the ALLEGRA corpus, with the help of a phrase alignment tool. The resulting lexicon is evaluated by comparison with a reference lexicon.

We then extend the lexicon induction approach to take advantage of the third language of ALLEGRA, namely Italian. Concretely, we induce a German-Italian lexicon from the German-Italian pair of the corpus and an Italian-Romansh lexicon from the Italian-Romansh one. We combine these by transitivity to build a German-Romansh lexicon. In turn, this transitive lexicon is intersected with the directly induced lexicon to filter out some noise. The intersected lexicon is shown to obtain higher precision than the directly induced lexicon, with no loss in quality.

The paper is structured as follows: We start by giving some linguistic and sociolinguistic facts about Romansh in Section 2. In Section 3, we present the ALLEGRA corpus and explain how it has been created. The remainder of the paper deals with bilingual lexicon induction. Section 4 presents the methodology and the resources used to evaluate the induced lexicons. Section 5 details the direct lexicon induction approach and the corresponding results, while Section 6 discusses the extensions made by using Italian as a pivot language. We conclude in Section 7.

2. The Romansh Language in Switzerland

Romansh is a minority language spoken in South-Eastern Switzerland, in the canton of Grisons. The linguistic situation of this canton is very diverse, as a result of a mountainous topography that hindered communication across the valleys for a long time (Haiman and Beninca, 1992; Liver, 1999). The canton of Grisons is officially trilingual, with German-speaking, Italian-speaking and Romansh-speaking regions. Thus, the administration of the canton publishes most of its documents in all three languages, although in practice, the primary working language is German.

Romansh covers a group of Romance languages, traditionally spoken in the alpine valleys of Grisons. Since the middle of the 20th century, there are no more monolingual adult Romansh speakers. All speakers learn and use German (and Swiss German dialects) in school and in their everyday life outside of their native valley.

There are five major varieties (*idioms*) of Romansh in use today, spoken in different regions of the canton of Grisons, with a total of 32 000 native speakers: Sursilvan (17 000 speakers), Sutsilvan (1000 speakers), Surmiran (3000 speakers), Puter (5000 speakers) and Vallader (6000 speakers). Each of these idioms has its own writing conventions and its own local dialectal varieties. Besides these five "natural" idioms, a sixth language, called *Rumantsch Grischun*, has been created in the 1980s (Schmid, 1989). *Rumantsch Grischun* is intended as a supraregional written language for administrative and medial usage. It has been designed as a compromise between the five idioms. Since 2005, it is taught in some Romansh schools.

From a Natural Language Processing perspective, little work has been done on Romansh. An ongoing project aims to make the *Crestomazia Retorumantscha*, a large

Year	DE	IT	RM	Common
1997	182	11	12	9
1998	169	150	150	145
1999	157	131	130	129
2000	208	167	171	167
2001	234	158	171	155
2002	238	162	168	141
2003	169	112	113	97
2004	137	100	97	85
2005	161	136	135	126
2006	211	171	173	164
2007	201	146	144	138
2008	197	167	168	163
2009	175	175	175	174
2010	184	182	184	182
	2623	1968	1991	1875

Table 1: Number of press releases per language (DE = German, IT = Italian, RM = Rumantsch Grischun). The last column shows the number of press releases available in all three languages.

		Words	Sentences
DE-RM	DE	799 576	49 964
	RM	1 012 111	
DE-IT	DE	651 726	39 812
	IT	796 380	
IT-RM	IT	786 404	39 337
	RM	818 351	

Table 2: Numbers of sentences and words in the sentence-aligned files per language pair.

collection of historic texts, digitally available.[1] The language conservation agency *Lia Rumantscha* publishes a reference German-Romansh dictionary (*Pledari Grond*, see Section 4.2) and also leads efforts to localize commonly used software. This lack of linguistic tools and resources for contemporary Romansh was one of the main motivations for building the ALLEGRA corpus presented in the following section.

3. The ALLEGRA Corpus: a Trilingual Corpus with an Under-Represented Language

The ALLEGRA corpus is a new language resource for *Rumantsch Grischun*. It takes the form of a sentence-aligned trilingual corpus consisting of press releases in the three official languages of the canton of Grisons (its name is the acronym for "ALigned press reLEases of the GRisons Administration"; *allegra* also means 'hello' in Romansh). The web site of the Grisons administration[2] gives access to all press releases since 1997. Most of these releases have been written in German and translated to *Rumantsch Grischun* and Italian. They are intended for a large audience and do not contain much specialized language.

DE	*Die Bündner Regierung hat in einem Schreiben an den Dachverband Swiss Olympic ihr Interesse bekundet, eine Kandidatur für Olympische Winterspiele voranzutreiben.*
IT	*In una lettera a destinazione dell'associazione mantello Swiss Olympic, il Governo grigionese ha manifestato il proprio interesse nel portare avanti una candidatura per i Giochi Olimpici invernali.*
RM	*En ina brev a la federaziun da tetg "associaziun olimpica svizra" ha la regenza grischuna fatg manifestà ses interess da preparar ina candidatura per gieus olimpics d'enviern.*

Figure 1: Example of an aligned sentence in all three languages.

ALLEGRA resembles the CLE corpus (Streiter et al., 2004) for German, Italian and Ladin. Ladin is a group of idioms closely related to Romansh and spoken in adjacent Northern Italy. The "Monitor" subcorpus of CLE also contains documents produced by local and regional administrations. The ALLEGRA corpus was prepared as follows:

- All documents were downloaded, cleaned and converted to plain text format.
- Only documents available in the three languages were kept.
- Until 2009, the different language versions of a document were not linked. This linkage was added manually on the basis of the release date and the title of the document.
- All documents were sentence-aligned with a standard alignment tool (Gale and Church, 1993).

Table 1 shows the number of original documents. The statistics of the sentence-aligned corpus are given in Table 2. An example of an aligned sentence in the three languages is given in Figure 1. The corpus is available for download in raw and sentence-aligned formats.[3]

Possible uses of trilingual corpora are numerous in NLP-related tasks, particularly when they include an under-represented language. In the following, we show how such resources can be exploited to improve the automated constitution of bilingual lexicons.

4. Bilingual Lexicon Induction by Phrase Alignment

4.1. Related Work

In statistical machine translation, bilingual word correspondences are commonly induced by finding word alignments in parallel corpora. While the initial models aligned single

[1] See `http://www.crestomazia.ch`. The *Crestomazia* project is led by Prof. Jürgen Rolshoven and Prof. Wolfgang Schmitz (University of Cologne).

[2] `http://www.gr.ch`

[3] `http://www.latl.unige.ch/allegra`

source words with single target words (Brown et al., 1993; Och and Ney, 2003), they have been subsequently extended to allow longer segments of words on the source and target sides. These aligners output *phrase tables*, which associate source phrases with target phrases and compute their respective translation probabilities. These phrase tables can then be used as translation lexicons.

Automatically inducing bilingual lexicons is challenging because the quality and the coverage of the resulting lexicon depend on the quality of the alignment and the size of the corpus. For under-represented languages, large parallel corpora are not often available, resulting in poor alignment performance.

For language pairs with scarce parallel resources, a recent trend in machine translation is to use a third (pivot) language (Utiyama and Isahara, 2007; Wu and Wang, 2007). This idea has also been pursued for lexicon creation, where two existing translation lexicons can be joined on the basis of a common pivot language (we also refer to this approach as *transitive*). In most of these studies, the source-pivot lexicons have been created manually by lexicographers (Ahn and Frampton, 2006; Nerima and Wehrli, 2008).

In all experiments reported here, we use the phrase aligner Anymalign (Lardilleux and Lepage, 2009), which has been shown to perform better on the lexicon induction task (Lardilleux et al., 2010) than the well-known GIZA++/Moses pipeline (Koehn et al., 2007). Anymalign is a sampling-based aligner that can be stopped at any time. In all experiments, we ran Anymalign for 15 minutes.

4.2. The Reference Lexicon *Pledari Grond*

The phrase pairs induced by alignment from the ALLEGRA corpus can be viewed as bilingual lexicons. As explained in Lardilleux et al. (2010), bilingual lexicon induction methods can be evaluated by comparing the induced phrase pairs with an existing reference lexicon.

We use *Pledari Grond*, the German-Romansh dictionary conceived as a reference for *Rumantsch Grischun*. It is available online[4] and is continually updated. Our version was obtained and pre-processed on October 15th, 2011 and contains 210 015 entries. Each entry consists of one lexical entry in the source language (German) that is matched with another lexical entry in the target language (Rumantsch Grischun). Lexical entries may consist of one or several words.

It is unlikely that all words of the dictionary appear in the ALLEGRA corpus. Therefore, recall rates of the lexicon induction approach are expected to be very low. In order to define a more sensible recall measure, we follow Lardilleux et al. (2010) and filter the reference lexicon by retaining only word pairs that occur in at least one sentence pair of the corpus. The idea is that words that do not occur in the same sentence pair are very unlikely to be phrase-aligned. As a result of filtering, the number of dictionary entries drops from 210 015 to 26 627.[5]

[4] `http://www.pledarigrond.ch/`

[5] The pivot language approach presented in Section 6 is able to recover correspondences that do not occur in aligned sentence pairs. This means that in theory, this approach may yield coverage values of over 100%. In practice however, the coverage of all lexicon induction approaches, direct or transitive, is much lower.

4.3. Stemming

The main issue for evaluation is that the reference lexicon contains lemmatized forms, while the induced phrase tables contain inflected word forms. We resolved this issue by accepting all entries with matching stems.

In practice, for the German part of the bilingual lexical pairs, we used the Snowball stemmer included in NLTK[6] to reduce the German words to their stems. We modified the stemmer so that derivational suffixes are not removed. The same stemmer is applied on the German side of the reference lexicon. For Romansh, we created a similar stemmer on the basis of a *Rumantsch Grischun* grammar available on the *PledariGrond* website.[7] It covers regular inflection patterns of verbs, adjectives and nouns.

Even if the stemming algorithm does not provide 100% accuracy, the few errors do not degrade the comparison results. An evaluation performed manually on 500 word pairs showed that only 3 (0.6%) of them were incorrectly classified because of stemming errors.

5. Direct Lexicon Induction

In this first experiment, we use the German-Romansh part of the ALLEGRA corpus to induce a bilingual lexicon. We ran Anymalign for 15 minutes, which resulted in a phrase table consisting of 186 159 phrase pairs. Like in all following experiments, phrases containing punctuation signs and/or numbers were removed, since they are not relevant for bilingual lexicon induction. The phrase table can be characterized by the following figures (see Table 3, second-last column):

- More than three quarters of all phrases consist of a single word (76.95% of German phrases, 79.43% of Romansh phrases). As a result, the mean length of all phrases is close to 1 in both languages.
- There are a lot of entries with low probabilities. The average probability of all phrases is 0.25, while the median probability only lies at 0.09.
- About half of the unstemmed phrases are unambiguous, i.e., occur in a single phrase pair (51.83% of German phrases, 49.17% of Romansh phrases). However, the remaining phrases can be highly ambiguous: there are only 52 636 unique German entries and 39 915 unique Romansh entries.
- Stemming increases the ambiguity: after applying the stemming, there remain 45 357 unique German entries and 34 707 unique Romansh entries (numbers not reported in Table 3).

[6] The Snowball stemmer is an extended version of the Porter stemmer (Porter, 1980), available online at `http://snowball.tartarus.org/`. NLTK is described in Bird (2006). The NLTK implementation of the Snowball stemmer has been contributed by P.M. Stahl.

[7] `http://www.pledarigrond.ch/grammatica.pdf`

	Direct				Intersection
	Correct	Wrong	Unknown	All phrases	All phrases
Phrase pairs	10 243	95 338	80 578	186 159	120 251
Single-word DE phrases	97.98%	98.95%	64.56%	76.95%	83.06%
Single-word RM phrases	91.44%	97.85%	69.55%	79.43%	84.31%
Mean length (DE)	1.02	1.01	1.55	1.35	1.24
Mean length (RM)	1.10	1.02	1.58	1.38	1.27
Mean probability	0.47	0.06	0.32	0.25	n/a
Median probability	0.47	0.01	0.25	0.09	n/a
Unique DE entries	7666	9572	47 832	52 636	29 750
Unique RM entries	6507	6601	37 635	39 915	22 872
Unambiguous DE entries	78.72%	42.57%	56.98%	51.83%	49.98%
Unambiguous RM entries	67.97%	29.63%	52.28%	49.17%	43.65%

Table 3: General characteristics of the directly induced phrase table and the intersected phrase table.

These figures show that phrase alignment produces a lot of low-probability entries with high ambiguity rates. This is where the difficulties lie because lexicon induction should ideally provide high-probability and low-ambiguity entries to be included in dictionaries.
In the following subsection, we evaluate the induced phrase pairs by comparison with the reference lexicon.

5.1. Evaluation of the Phrase Quality

When comparing the phrases obtained by alignment with the entries of the reference lexicon, three possibilities arise:

1. The phrase pair is exactly matched by a reference lexicon entry. In this case, the phrase pair is considered **correct**.

2. The phrase pair does not constitute an entry in the reference lexicon, but both phrases figure somewhere else in the lexicon. Such phrase pairs are considered **wrong**.

3. One or both phrases do not appear in the reference lexicon. This may mean that the phrase pair is wrong, but it can also simply be a consequence of limited lexicon coverage. We label these cases as **unknown**.

These three cases allow us to perform two types of evaluation. The ratio of cases 1 and 2 measures the quality of the alignment. In contrast, the ratio of case 1 and case 3 rather gives an indication about the quality and the coverage of the reference lexicon.
Table 3 (first line) sums up the results. 10 243 phrase pairs (5.50%) are correct, 95 338 (51.21%) are wrong, and 80 578 (43.28%) could not be evaluated accurately. The results show a low ratio of correctly induced word pairs. However, we consider these results satisfactory given the small size of the corpus.
The following observations may shed some light on the characteristics of the induced phrase pairs.

- In the correct and wrong categories, we find an overwhelming majority of single-word entries. This contrasts with the unknown category, where the proportion of single-word phrases is much lower, suggesting that single-word entries are more readily found in the reference lexicon. This is partially true: only 16.9% of all German entries in the *Pledari Grond* are multi-word expressions, but 60.4% of Romansh entries are. This divergence is probably caused by a large number of compound nouns which are single words in German but consist of several words in Romansh.

  Furthermore, the phrase aligner yields some phrases that do not form constituents; such phrases would require post-processing to be included in a dictionary:

  (1) *aus Gemeinden und — da vischnancas e*
  'from municipalities and'

- The translation probability[8] seems to be a rather good indicator of the correctness of a phrase pair: on average, correct phrases have a higher translation probability than wrong ones (0.47 vs. 0.06). The probability of the unknown pairs lies in the middle (0.32). This large difference in probabilities provides further justification to distinguish the unknown word pairs from the plain wrong ones.

- Wrong pairs are more ambiguous on average than the correct ones (see last four lines of Table 3). This presumably correlates with translation probability: the more ambiguous a source word is, the lower on average its different translation probabilities are. The unknown pairs are more ambiguous than the correct pairs, but less ambiguous than the wrong pairs.

These observations suggest that word length, translation probability and/or ambiguity may be correlated with the correctness of the results. We computed point-biserial correlations by leaving out unknown phrases. For length and ambiguity, we obtained very low but significant Pearson correlation coefficients (between 0.04 and 0.27, all at

[8]For each phrase pair, the alignment tool outputs two probabilities: $p(DE|RM)$ and $p(RM|DE)$. We computed the average of both probabilities for each phrase pair. The mean and median probabilities reported in Table 3 are computed on the basis of these average probabilities.

$p < 0.001$). Translation probability correlates with correctness at 0.61 ($p < 0.001$). These features could be used as filters to improve lexicon induction in future work. But in this study, we continue by considering all proposed phrases, regardless of their length, probability and ambiguity.

To conclude this sub-section, it is useful to draw a general picture of the performance of the direct induction approach in terms of precision and recall. For precision, we propose two modes of evaluation: a "strict" mode that considers all unknown phrase pairs as wrong, and a "large" mode that discards the unknown pairs from the evaluation altogether. In other words, strict precision is the ratio between correct pairs and correct + wrong + unknown pairs. "Large" precision is the ratio between correct pairs and correct + wrong pairs. In both strict and large evaluation modes, recall is defined as the ratio between correct pairs and lexicon entries (26 627, see Section 4.2). Table 4 summarizes the results.

5.2. Manual Evaluation

The automatic matching between phrase pairs and reference lexicon entries may be subject to errors. In order to assess the quality of these matches, we inspected them manually. We distinguish four cases: (i) genuinely correct pairs; (ii) partially correct pairs; (iii) named entities and (iv) genuinely wrong pairs. Partially correct pairs are pairs that contain more lexical material on one side, but still have a common lexical unit. They include cases of partially translated German compound nouns (see example (2)), spurious determiners and prepositions, as well as part-of-speech mismatch (e.g. a verb instead of a noun derived from that verb, as in (3)).

(2) *Strassenprojekte* 'road projects' — *vias* 'roads'

(3) *Erhöhung* 'increase' — *augmentar* 'to increase'

First, we evaluated the 100 most frequent phrase pairs in each of the three categories (correct, wrong, unknown). All phrase pairs annotated as correct were correct indeed. Rather surprisingly, among the pairs annotated as wrong, 35% can be considered correct, and 31% are partially correct. The unknown pairs turned out to be correct in 44% of the cases, and 38% of them were named entities that typically do not figure in a lexicon.

Second, in order to counter the effects of Zipf's law, we carried out the same evaluation on a more representative subset of 100 randomly selected phrase pairs for each category. The phrase pairs annotated as correct were mostly correct. The randomly selected wrong pairs show a different behavior than the most frequent wrong pairs: 93% of wrong pairs were indeed wrong. Among the unknown pairs, 51% were genuinely wrong, 28% were partially correct, 10% of the pairs were correct, and 9% were named entities.

The evaluation on the random pairs is summed up in Table 5 and compared with a similar evaluation of the other investigated lexicon induction approach (see Section 6).

With this closer look at the aligned phrases, it appears that the low performance of the lexicon induction approach is not only due to the method itself, but also to the quality of the reference lexicon on which the method is assessed. Nevertheless, the use of a reference is still useful to evaluate possible improvements obtained by other methods, such as the one presented in Section 6.

		Direct	**Intersection**
	Induced phrase pairs	186 159	120 251
	Correct	10 243	8 018
	Wrong	95 338	68 267
	Unknown	80 578	43 966
Strict	Precision	5.50%	6.67%
	Recall	25.01%	19.58%
	F-Measure	9.02%	9.95%
Large	Precision	9.70%	10.51%
	Recall	25.01%	19.58%
	F-Measure	13.98%	13.68%

Table 4: Performances of the direct and the intersection lexicon induction approaches.

6. Improving Phrase Alignment by Transitivity

When working with under-represented languages, it often occurs that machine learning algorithms perform badly due to a lack of training data. While small in size, the great asset of the ALLEGRA corpus is that is trilingual, and so the third language (Italian) can be exploited to improve bilingual lexicon induction. In this section, we test the hypothesis of improved lexicon induction by using all language pairs available in the corpus.

In a first step, we create a German-Romansh phrase table by transitivity, using Italian as a pivot language. However, such a phrase table is likely to contain a lot of noise due to the pivot approach. Consequently, in a second step, we intersect the "transitive" phrase table with the "direct" phrase table presented in the previous section, in order to increase the precision. The creation of the phrase table by transitivity is presented in Section 6.1 and the intersection of that phrase table with the direct one is presented in Section 6.2.

6.1. Joining Phrase Tables by Transitivity

Lexicon induction by transitivity consists of three steps: (i) inducing a lexicon for the source-pivot pair; (ii) inducing a lexicon for the pivot-target pair; and (iii) combining the two induced lexicons.

We ran Anymalign for 15 minutes on the DE-IT section of ALLEGRA, which resulted in a phrase table of 212 496 phrase pairs. Analogously, the alignment of the IT-RM section yielded 221 742 phrase pairs. The two phrase tables were merged following the algorithm proposed by Wu and Wang (2007). The probability $\phi(f \mid e)$ of translating a source phrase f into a target phrase e is obtained by summing the probabilities of all paths leading from f to e through any pivot phrase p:

$$\phi(f \mid e) = \sum_{p} \phi(f \mid p)\phi(p \mid e)$$

As expected, the resulting transitive lexicon exploses in size: it contains 5 842 211 phrase pairs – 31 times as many

	Direct			Intersection		
	Correct	Wrong	Unknown	Correct	Wrong	Unknown
Correct	97%	1%	9%	97%	3%	16%
Partially correct	2%	6%	31%	2%	8%	24%
Named entity	1%	0%	9%	0%	0%	4%
Wrong	0%	93%	51%	1%	89%	56%

Table 5: Manual evaluation of the error categories obtained by automatic matching of the phrase table with the reference lexicon. Automatically assigned error categories are in columns, manually annotated error categories on the rows.

as the directly induced lexicon of Section 5. Recall is increased (14 843 correct phrase pairs vs. 10 243 with the direct approach), but the pivot approach introduces so many wrong low-probability entries that precision is virtually non-existent (0.25% in the strict evaluation mode, 0.47% in the large mode).

The inferior quality of the transitive lexicon can also be seen in the ambiguity scores: only 12.04% of the German entries and 13.73% of the Romansh entries are unambiguous. This contrasts with 51.83% and 49.17% in the direct approach (see Table 3). In other words, a German phrase has 139 different Romansh correspondences on average, and a Romansh phrase has 177 German correspondences on average.

As such, the transitive lexicon cannot be of any interest for a lexicon induction task, but it is of great value for the creation of the intersected lexicon presented below. Therefore, we refrain from an extensive evaluation of the transitive lexicon.

6.2. Intersecting the Direct and Transitive Phrase Tables

The transitive phrase table improves recall, but at the price of much lower precision. However, it can be viewed as a filter on the directly induced phrase table: if a lexical correspondence is found through two paths (the direct and the transitive one), it is more likely to be correct than if it is found through only one path. Following this idea, we intersect the direct phrase table with the transitive phrase table. Only phrase pairs that appear in both tables are retained; their probabilities are multiplied.

Table 3 (rightmost column) reports the main figures of the intersected lexicon. The total number of induced phrase pairs diminishes in comparison to the direct phrase table. This is expected, as not all directly induced phrases are validated by the transitive table.

6.3. Alignment Results

Two properties of the intersected phrase table are worth reporting (see Table 3, last column).

First, the tendency towards single words is even greater than with the direct approach: 83.06% of all German phrases and 84.31% of all Romansh phrases consist of a single word. The intersection approach successfully filters out multiword phrases.

Second, the number of unique entries decreases by half (29 750 German entries before stemming, 22 872 Romansh entries before stemming), which means that there is quite a lot more ambiguity than in the direct phrase table. However, the proportion of unambiguous words is only slightly lower than in the direct approach (49.98% of German entries, 43.65% of Romansh entries).

The translation probabilities of the intersected lexicon were computed differently than the ones of the direct approach. A comparison in terms of mean/median translation probabilities is therefore not meaningful.

6.4. Evaluation of the Phrase Quality

In Section 5.1, we have established a three-fold distinction between correct pairs, wrong pairs, and unknown pairs (pairs which could not be validated reliably because the reference lexicon was incomplete). Here, we proceed to the same type of evaluation with the intersected phrase table. The results are reported on the right column of Table 4.

8018 phrase pairs are found correct (6.67%), 68 267 pairs are found wrong (56.77%), and 43 966 pairs could not be evaluated (36.56%). With respect to the direct approach, the proportion of correct pairs increases as well as the proportion of wrong pairs, while the proportion of unknown pairs decreases.

Precision and recall figures are again reported according to the "strict" and "large" modes introduced at the end of Section 5.1. With respect to the direct approach, precision rises by about 1 percentage point in both modes, while recall diminishes by about 5.5 percentage points. The F-measure of the intersected lexicon is slightly higher in the strict mode, but slightly lower in the large evaluation mode.

6.5. Manual Evaluation

Again, we randomly chose 100 phrase pairs of each type (correct, wrong and unknown) and analyzed them manually. Results are shown on the right half of Table 5 and compared with the evaluation results from the direct induction approach. The two approaches do not differ much in terms of phrase quality.

Thus, the quality of the intersected lexicon entry does not significantly improve, nor is there any remarkable loss. This is a positive result, as it allows precision to increase.

7. Discussion and Concluding Remarks

We have presented a trilingual sentence-aligned corpus for German, Italian and Romansh. We have also discussed an example of an application of this corpus, namely bilingual lexicon induction. The results show that a small corpus can be better exploited by taking several paths. In our case, we combine the direct German-Romansh path with a transitive German-Italian-Romansh path to improve the precision of

the induced phrase pairs. In general, the intersected phrase table has a slighlty better precision, which is rather encouraging. Moreover, it proves to be of equal quality, with more single word pairs.

However, the obtained results remain quite low: this suggests (i) that lexicon induction is a difficult task, and (ii) that there is room for improvement in future research. We address these two points in turn.

It should be noted that completely automatic methods of bilingual lexicon induction never achieve optimal quality. For instance, Lardilleux et al. (2010) never obtain F-measures higher than 50%. The induced lexicons should therefore be viewed as a starting point for semi-automatic lexical acquisition. In this case, their quality should be measured in terms of manual post-processing.

Another crucial point in our approach is that we never filtered the phrase pairs by translation probability. There were mainly technical obstacles behind this choice. Since the probabilities in the three lexicons are computed differently, they cannot be compared directly, and different thresholds would have to be defined. Still, in Section 5.1, we reported a fair correlation between translation probability and correctness of the phrase pair. There is thus reason to believe that translation probability can act as a sensible filter to remove some wrongly induced word pairs and increase precision.

The intersection approach only yields slight improvements. This is mainly due to the fact that intersection removes some word pairs that appear with high probability in the direct table, but are absent from the transitive table. Future work will assess a more sophisticated intersection approach. Likewise, the phrase tables created by transitivity can be improved. For example, the German-Italian phrase table created with ALLEGRA could be extended with the help of a larger training corpus such as Europarl (Koehn, 2005).

Finally, we have not taken advantage of the close linguistic relationship between Italian and Romansh. For instance, phrase pairs that are not cognates could be penalized, and similar linguistics-based filtering methods could be envisaged.

Acknowledgments

This work is partially funded by the COMTIS project (see `http://www.idiap.ch/comtis`) under the Sinergia program of the Swiss National Science Fundation (CRSI22 127510).

8. References

Kisuh Ahn and Matthew Frampton. 2006. Automatic generation of translation dictionaries using intermediate languages. In *Proceedings of the EACL 06 Workshop on Cross-Language Knowledge Induction*, pages 41–44, Trento, Italy.

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, Sydney, Australia.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

John Haiman and Paola Beninca. 1992. *The Rhaetoromance Languages*. Routledge, London, New York.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Prague, Czech Republic.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit 2005*.

Adrien Lardilleux and Yves Lepage. 2009. Sampling-based multilingual alignment. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 09)*, Borovets, Bulgaria.

Adrien Lardilleux, Julien Gosme, and Yves Lepage. 2010. Bilingual lexicon induction: Effortless evaluation of word alignment tools and production of resources for improbable language pairs. In *Proceedings of LREC 2010*, La Valletta, Malta.

Ricarda Liver. 1999. *Rätoromanisch: Eine Einführung in das Bündnerromanische*. Gunter Narr, Tübingen.

Luka Nerima and Eric Wehrli. 2008. Generating bilingual dictionaries by transitivity. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

M.F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Heinrich Schmid. 1989. Richtlinien für die Gestaltung einer gesamtbündnerromanischen Schriftsprache Rumantsch Grischun. *Annalas de la Societad retorumantscha*, 102.

Oliver Streiter, Mathias Stuflesser, and Isabella Ties. 2004. CLE, an aligned tri-lingual Ladin-Italian-German corpus. corpus design and interface. In *Proceedings of the LREC 2004 Workshop on First Steps for Language Documentation of Minority Languages*, Lisbon, Portugal.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Proceedings of NAACL-HLT 2007*, Rochester, NY, USA.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL07)*, volume 21, pages 856–863, Prague, Czech Republic.