

# Contents

<b>Abstract</b>	<b>i</b>
<b>Acknowledgements</b>	<b>i</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Research Question and Goals . . . . .	2
1.2.1 Research Questions . . . . .	2
1.2.2 Goals . . . . .	3
1.3 Structure . . . . .	3
1.4 GitHub repository . . . . .	3
<b>2 Romansh</b>	<b>4</b>
2.1 Rhaeto-Romance . . . . .	4
2.2 Romansh . . . . .	5
2.3 Rumantsch Grischun . . . . .	6
2.3.1 Lia Rumantscha . . . . .	6
2.3.2 Rumantsch Grischun . . . . .	6
2.3.3 Features . . . . .	7
2.3.4 Today . . . . .	7
<b>3 Compiling the Corpus</b>	<b>9</b>
3.1 Introduction . . . . .	9
3.2 Collecting the Data . . . . .	9
3.3 Web Scraping . . . . .	10
3.4 Building the Corpus . . . . .	11
3.4.1 HTML Parsing . . . . .	11
3.4.2 Document Alignment . . . . .	12
3.5 SQLite database . . . . .	13
3.6 Summary . . . . .	15
3.6.1 Statistics . . . . .	15

<b>4</b>	<b>Sentence Alignment</b>	<b>19</b>
4.1	Introduction . . . . .	19
4.1.1	Formal definition . . . . .	19
4.2	Method Overview . . . . .	20
4.2.1	Length Based . . . . .	20
4.2.2	Partial Similarity Based . . . . .	20
4.2.3	Translation based . . . . .	21
4.2.4	Hybrid models . . . . .	21
4.2.5	Summary . . . . .	22
4.3	More Recent methods . . . . .	22
4.3.1	Bleualign . . . . .	22
4.3.2	Vecalign . . . . .	23
4.4	Sentence Alignment Pipeline . . . . .	24
4.4.1	Tool of choice . . . . .	24
4.4.2	Pipeline . . . . .	25
4.4.3	Database Query and Sentence Segmentation . . . . .	25
4.4.4	Aligning Language Pairs . . . . .	26
4.4.5	Filtering and Tokenizing . . . . .	27
4.5	Results . . . . .	28
<b>5</b>	<b>Word Alignment</b>	<b>29</b>
5.1	Introduction . . . . .	29
5.2	Overview of Methods . . . . .	30
5.2.1	IBM Model 1 . . . . .	30
5.2.2	Higher IBM Models . . . . .	32
5.3	Word Embeddings . . . . .	33
5.3.1	Excursion: Words . . . . .	33
5.3.2	Word Embeddings . . . . .	34
5.3.3	Word Similarity . . . . .	35
5.3.4	Multilingual Word Embeddings . . . . .	36
5.3.5	Summary . . . . .	36
5.4	Similarity Based Word Alignment . . . . .	37
5.4.1	Method . . . . .	37
5.4.2	Summary . . . . .	38
<b>6</b>	<b>Gold Standard</b>	<b>40</b>
6.1	Introduction . . . . .	40
6.2	Sure and Possible Alignments . . . . .	41
6.3	Gold standard for German-Romansh . . . . .	41

6.3.1	Annotation tool . . . . .	42
6.3.2	Guidelines . . . . .	42
6.3.3	General principles . . . . .	42
6.3.4	Examples . . . . .	43
6.4	Flaws . . . . .	46
<b>7</b>	<b>Results</b>	<b>48</b>
7.1	Evaluation Metrics . . . . .	48
7.2	Baseline Systems . . . . .	49
7.2.1	fast_align . . . . .	49
7.2.2	eflomal . . . . .	50
7.2.3	Performance . . . . .	50
7.3	SimAlign . . . . .	50
7.3.1	Performance . . . . .	51
7.4	Discussion . . . . .	52
7.4.1	General Problems with Evaluation . . . . .	54
7.5	Summary . . . . .	55
<b>8</b>	<b>Concluding Words</b>	<b>56</b>
8.1	Goals . . . . .	56
8.2	Corpus Compiliation . . . . .	56
8.3	Gold Standard . . . . .	57
8.4	Evaluation . . . . .	57
8.5	Future . . . . .	57
	<b>List of Tables</b>	<b>59</b>
	<b>List of Figures</b>	<b>60</b>
	<b>List of Listings</b>	<b>62</b>
	<b>Bibliography</b>	<b>63</b>
<b>A</b>	<b>Algnment Examples</b>	<b>70</b>
A.1	Compounds . . . . .	70
A.2	Perfect–Perfect . . . . .	71
A.3	German Preterite–Romansh Perfect . . . . .	73
A.4	Double Negation . . . . .	74
A.5	Differing Word Order . . . . .	74
A.6	Summary . . . . .	78

<b>B</b>	<b>JSON examples</b>	<b>79</b>
<b>C</b>	<b>Aligning Romansh to Italian</b>	<b>83</b>
C.1	Examples . . . . .	83
C.2	Summary . . . . .	86

# Glossary

**Graubünden** The Canton of Grisons. 1, 5, 9, 56

**HTML** Hypertext Markup Language. A language containing display instructions for web browsers and the format in which web pages are usually saved . 10

**JSON** JavaScript Object Notation. A format for organizing data in a hierarchical form. 11

**Standeskanzlei** State Chancellery of Grisons. 9, 56

**URL** Uniform Resource Locator. A reference to an internet resource, a web address. 10

# Acronyms

**AER** average error rate. 38, 48, 50, 51, 52, 54, 61

**EM** expectation-maximization. 31

**gen** genitive. 45

**HTML** Hypertext Markup Language. 10, 11, 15

**JSON** JavaScript Object Notation. 11, 12, 13, 15, 56, 58, 79

**NER** named entity recognition. 2

**NMT** neural machine translation. 24, 27

**part** participle. 45

**pl** plural. 45

**POS** part of speech. 2, 43, 55

**pres** present. 45

**SMT** statistical machine translation. 23

**URL** Uniform Resource Locator. 27, 56

# List of Tables

2.1	Examples for choosing the forms for Rumanstch Grischun . . . . .	8
3.1	Description of the table corpus in corpus.db . . . . .	14
3.2	Description of the table raw in corpus.db . . . . .	14
3.3	Number of parallel documents per year . . . . .	16
3.4	Number of documents per language and year . . . . .	17
3.5	Twenty most frequent tokens in each language in the corpus . . . . .	18
4.1	Parallel corpus in numbers . . . . .	28
6.1	Translation examples of German compounds into Romansh . . . . .	44
7.1	Word alignment quality of the baseline models . . . . .	51
7.2	Word alignment quality using SimAlign . . . . .	52
7.3	Comparison of the best performance of the three SimAlign methods . . .	52

# List of Figures

2.1	Map of the distribution of Rhaeto-Romance . . . . .	5
3.1	Directory tree of corpus_builder . . . . .	10
3.2	Directory scheme for saving the HTML files . . . . .	11
3.3	Portion of automatically aligned press releases . . . . .	12
3.4	Corpus creation pipeline . . . . .	15
4.1	Sentence alignment pipeline . . . . .	25
5.1	Word alignment example . . . . .	29
5.2	Similarity matrix . . . . .	37
5.3	Alignment matrix . . . . .	37
5.4	The resulting word alignment . . . . .	38
6.1	Aligning German compounds to a Romansh noun phrases . . . . .	44
6.2	Aligning German perfect to Romansh perfect . . . . .	44
6.3	Alignment of German preterite to Romansh perfect . . . . .	45
6.4	Aligning a German present participle to a Romansh relative clause . . . . .	45
7.1	Comparing precision between the systems for different dataset sizes. . . . .	53
7.2	Comparing recall between the systems for different dataset sizes. . . . .	53
7.3	Comparing AER between the systems for different dataset sizes. . . . .	54
A.1	Word alignment example for the case of perfect tense in German and Ro- mansh . . . . .	71
A.2	Word alignment example with compounds. . . . .	72
A.3	Word alignment example with compounds. . . . .	72
A.4	Word alignment example for the case of perfect tense in German and Ro- mansh . . . . .	73
A.5	Word alignment example for the case of German preterite . . . . .	74
A.6	Word alignment example for the case of German preterite . . . . .	75
A.7	Word alignment example with a German separable verb in preterite . . . . .	75
A.8	Word alignment example with Romansh double negation ( <i>na ... betg</i> ) . . . . .	76



A.9	Word alignment example with Romansh double negation ( <i>na ... betg</i> ) . . .	76
A.10	Word alignment example with differing word order . . . . .	77
A.11	Word alignment example for a long sentence with differing word order . .	77
A.12	Word alignment example for a long sentence with differing word order . .	78
C.1	Word alignment example Romansh–Italian and Romansh–German . . . .	84
C.2	Word alignment example Romansh–Italian and Romansh–German . . . .	84
C.3	Word alignment example Romansh–Italian and Romansh–German . . . .	85
C.4	Word alignment example Romansh–Italian and Romansh–German . . . .	85
C.5	Word alignment example Romansh–Italian and Romansh–German . . . .	85

# List of Listings

- 4.1 Excerpt from a file containing sentences for alignment. . . . . 26
- 4.2 Excerpt from the file containing sentence pairs in German–Romansh . . . 28
- B.1 Example for a JSON file containing the press releases extracted from the  
HTML files. . . . . 79
- B.2 Example for a JSON file containing aligned documents . . . . . 80

# Bibliography

- Brown, Peter F. et al. (1993). “The Mathematics of Statistical Machine Translation: Parameter Estimation”. In: *Computational Linguistics* 19.2, pp. 263–311. URL: <https://aclanthology.org/J93-2003>.
- Pires, Telmo, Eva Schlinger, and Dan Garrette (July 2019). “How Multilingual is Multilingual BERT?” In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 4996–5001. DOI: [10.18653/v1/P19-1493](https://doi.org/10.18653/v1/P19-1493). URL: <https://aclanthology.org/P19-1493>.

# Appendix C

## Aligning Romansh to Italian

Due to the nature of my research question, I virtually ignored in the course of this work the issue of word alignments using embeddings (i.e., SimAlign) between Romansh and Italian. Therefore, I would like to curtly attend this issue in this appendix part.

Romansh and Italian share many similarities. Both of them are Romance languages and some researchers even consider Romansh to be a part of the Italian dialect continuum (see Section 2.1).

Since 1-to-many alignments and differing word order are more challenging to model than 1-to-1 alignments and similar or identical word order—word order or 1-to-many alignments are not modeled by IBM Model 1, but only by higher models (Brown et al. 1993)—one might expect that it should be easier to word-align languages that are more similar in structure, word order and grammar. That is, word-aligning Romansh to Italian should be easier than aligning Romansh to German due to the higher similarity between the former languages. Further, when dealing with unseen languages, as in the case of Romansh, multilingual language models have been shown to favor language similarity and vocabulary overlaps (Pires, Schlinger, and Garrette 2019). All this gives rise to the assumption that word alignment for Romansh–Italian might perform better.

I randomly hand-picked a few examples<sup>1</sup> and compared SimAlign’s performance on the pairs Romansh-Italian and Romansh-German in order to unempircally<sup>2</sup> test this notion.

### C.1 Examples

Figure C.1 is an example for a word alignment that works perfectly both with Italian and with German. In Figure C.2<sup>3</sup>, word alignment works well with Italian and German exactly

---

<sup>1</sup>The only precondition was that the sentences be short; Visualization for longer sentences leaves something to be desired.

<sup>2</sup>Obviously, a gold standard for Romansh-Italian would be needed.

<sup>3</sup>Apologies for the somewhat unreadable edges in Romansh–German

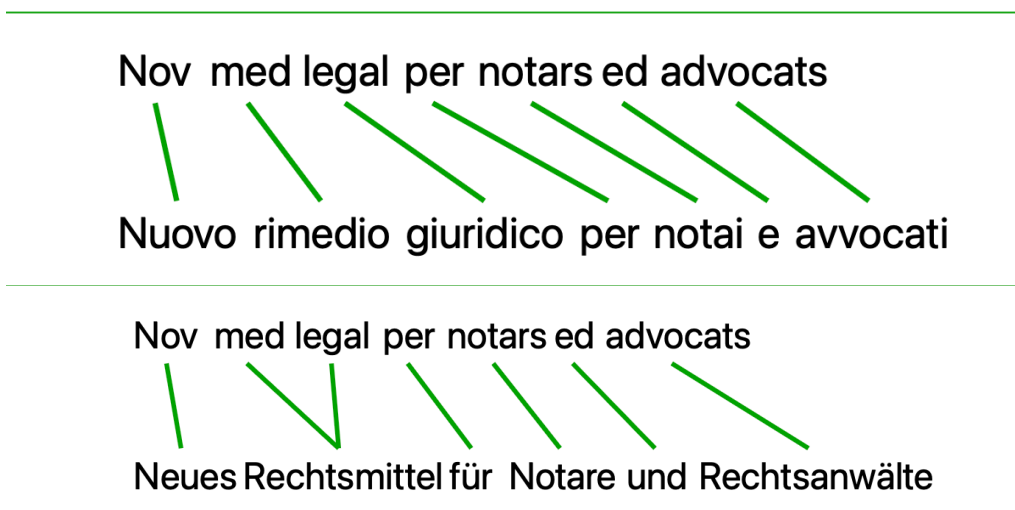


Figure C.1: Word alignment example Romansh–Italian and Romansh–German

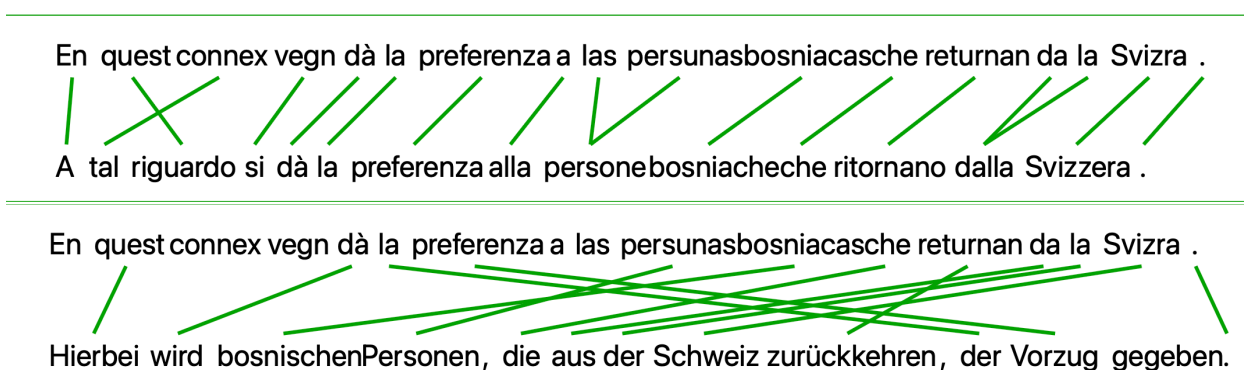


Figure C.2: Word alignment example Romansh–Italian and Romansh–German

for the same Romansh words, and it is exactly the same words where SimAlign fails: Romansh *en quest connex* (“in this context/matter”) is not aligned correctly, neither in German nor in Italian. The same applies for Romansh *vegn* (literally “come”, but here part of the passive construction), which is misaligned both times. This is also the case in Figure C.3. The same words are aligned correctly with German and with Italian, but in both cases Romansh *chantun* (“canton”) remains unaligned.

In Figure C.4 word alignment with German is even better than with Italian. Here, every alignment is correct, whereas in the Italian example, Romansh *schilar* (“tackle”) is not aligned to Italian *affronatare*, which should have been the case.

Finally, Figure C.5 is an example for many misalignments. In the German example, SimAlign succeeds in aligning Romansh *la derasaziuna da infecziuns* to German *die Durchseuchung*, but the rest of the alignments are wrong. The Italian example is completely misaligned.

---

La regenza dal chantun Glaruna visita il Grischun

Il Consiglio di Stato del Cantone di Glarona in visita nei Grigioni

---

La regenza dal chantun Glaruna visita il Grischun

Regierungsrat des Kantons Glarus besucht Graubünden

---

Figure C.3: Word alignment example Romansh–Italian and Romansh–German

---

Co duain ins schliar quest problem ?

Come affrontare dunque questo problema ?

---

Co duain ins schliar quest problem ?

Wie soll nun dieser Tatbestand angegangen werden ?

---

Figure C.4: Word alignment example Romansh–Italian and Romansh–German

---

La derasaziunda las infecziuns na sa lascha betg pli franar uschia .

In questomodo la diffusione del contagio non può più essere arrestata.

---

La derasaziunda las infecziuns na sa lascha betg pli franar uschia .

Die Durchseuchung lässt sich so nicht mehr aufhalten .

---

Figure C.5: Word alignment example Romansh–Italian and Romansh–German

## C.2 Summary

From observing these very few hand-picked cases, SimAlign doesn't seem to perform better when aligning Romansh to Italian. This is in spite of the higher similarity between Romansh and Italian, compared with German.

One possible explanation for this is that what mostly influences performance is the quality of the embeddings. If the Romansh word is similar enough to any of the words (or subwords) in the language model, alignment will work, regardless of the target language. Take for example Figure C.1. Here, all of the Romansh words are reminiscent of other seen languages and alignment works perfectly. However, in the case of Figure C.3, a suitable embedding for the Romansh word *chantun* cannot be looked-up for some reason, hence the word remains unaligned in both cases.