Contents

1 Introduction 3

1.1 Motivation . . . 3

1.2 Research Question and Goals . . . 3

2 Romansh 5

2.1 Raeto-Romance . . . 5

2.2 Romansh . . . 6

2.3 Rumantsch Grischun . . . 7

2.3.1 Lia Rumantscha . . . 7

2.3.2 Rumantsch Grischun . . . 7

2.3.3 Properties . . . 7

2.3.4 Today . . . 8

3 Compiling the Corpus 9

3.1 Introduction . . . 9

3.2 Collecting the Data . . . 9

3.3 Web Scraping . . . 9

3.4 Building the Corpus . . . 11

3.4.1 HTML Parsing . . . 11

3.4.2 Document Alignment . . . 12

3.5 Manual alignment of unlinked documents . . . 14

3.6 SQLite database . . . 14

3.7 Summary . . . 15

4 Sentence Alignment 17

4.1 Introduction . . . 17

4.2 Methods . . . 17

4.3 Statistical methods . . . 18

4.4 Recent methods . . . 18

5 Gold standard 19

5.1 Introduction . . . 19

5.2 Sure and Possible Alignments . . . 19

5.3 Evaluation Metrics . . . 20

5.4 Gold standard for German-Romansh . . . 20
5.4.1 Annotation tool . . . 20
5.4.2 Guidelines . . . 20
5.4.3 General priniciples . . . 21
5.4.4 Examples . . . 22
5.5 Flaws . . . 24
6 Evaluation 26
7 Contributions 27
8 Summary 28
Bibliography 28

Chapter 1

Introduction

1.1 Motivation

The Romansh language is a Romance language spoken in Switzerland, primarily in the Canton of Grisons (henceforth *Graubünden*), by around 60,000 speakers. Graubünden is the only canton in Switzerland with three official languages – German, Italian and Romansh.

When traveling by train, the announcements are heard in German, Romansh or Italian according to which part of the canton one is currently at. It is enough to travel to the next valley to suddenly be greeted on the street in a different language. Newspapers, radio and television exist in all three languages, but also official documents, laws and press releases are published trinlingually. While I was resident in Graubünden, I was fascinated by this multilinguality and it was my wish to somehow capture it and make it available to others. This is why I decided to build a multilingual corpus, a parallel collection of sentences in German, Romansh and Italian, in which the sentences are translations of each other.

Having such a low number of speakers makes it a so-called low resource language. Having so little speakers means there is also little data, be it corpora or research data. Most of the reasearch in NLP focuses on high resource languages.

1.2 Research Question and Goals

jalili-sabet-etal-2020-simalign were able to show that their algorithm for word alignment outperforms all the statistical baseline models. Contrary to statistical models, their model uses vectors of word representations learned by a neural net (also commonly known as word embeddings) and, by using some sort of similarity measurement (e.g., cosine similarity), aligns the most similar words in the source and the target sentence.

But not only that the model outperforms the existing stastical models, its biggest advantage as propogated by **jalili-sabet-etal-2020-simalign** is that it requires no training data. Statistical models will only reach a threshold of good performance with enough training data (TODO: cite numbers from SimAlign). Using word embeddings can be used to align words in just a single sentence with high precision. Of course, all of this works persuming we already have a trained model whose learned embeddings we can use for this task. There exist some language models that were trained on multi-lingual data. mBERT was trained on 104 languages and LASER was trained on 93 languages. But will word embeddings based word alignment will work in zero-shot settings? That is, can the embeddings learned by a multilingual language model be used for word alignment for a language that wasn't included in the training data?

Chapter 2

Romansh

In this chapter, I will provide a short context about Romansh, the language that builds a third of the resulting corpus, but conceptually the main motivation for this work.

2.1 Raeto-Romance

In 1873, an Italian linguist by the name of Grziadio Ascoli pointed out to a shared number of chracterizing phenomena in a number of Romance dialects spoken in parts of Switzerland and Italian (but without a geographical continuum) and named this group of dialects “Ladino”. Since 1883, influenced by Theodor Gartner’s publication *Raetoromanishce Grammatik* on this group of dialects, this name (German *Rätoromanisch*, English “Raeto-Romance”) became associated with them.

Raeto-Romance is spoken in three separated areas and is made up of three super-dialects: Romansh, spoken in parts of the Swiss canton of Graubünden, Ladin, spoken in the Dolomotic Alps in northern Italy (Südtirol), and Friualian, spoken around the drainage basin of the Tagliamento river, between Venice and Trieste (**haiman1992**).

There have been long discussions in Romance linguistics about whether Raeto-Romance can be seen as a unity of dialects, or whether such a unity is merely a linguistic construct, lacking a sociolinguistical-historical basis. This dispute is referred to as the *questione ladina* “the Ladin question” (**liver1999**).

Ascoli, the grounder of the idea of one Raeto-Romance unity, made his classifications at a time where language researchers were fascinated by the regularity of sound changes and common historical sound changes were used to group languages and dialects together. He therefore based his grouping of these three dialects on the grounds of sound changes common to all three dialects. His followers propagate a narrative according to which the three dialects once occupied one geographical area, but were seperated by the Germanic incursions in the years CE 250-800 (**bossong2008**; **haiman1992**).

An opposing group of reserchers believes that the three Raeto-Romance dialects show decisive features common with their respective neighboring Italian dialects. They should therfore be classified as an Italian dialect and part of the Italian dialect continuum (**bossong2008**).

This question, as interesting as it may be, is not of importance to this thesis and will not bother

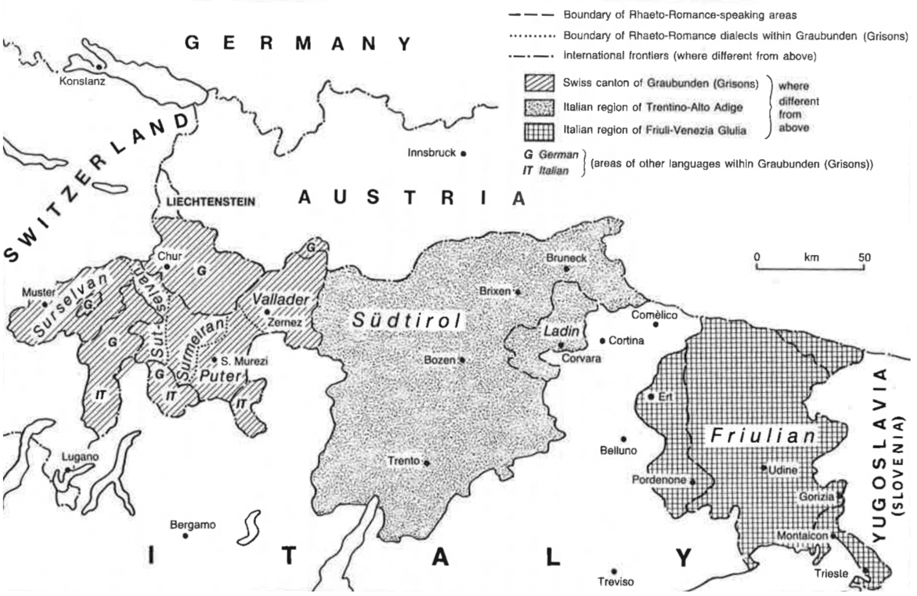

Boundary of Rhaeto-Romance-speaking areas
Boundary of Rhaeto-Romance dialects within Graubunden (Grisons)
International frontiers (where different from above)
Swiss canton of Graubunden (Grisons)
Italian region of Trentino-Alto Adige
Italian region of Friuli-Venezia Giulia
where different from above
G German
IT Italian
(areas of other languages within Graubunden (Grisons))
0 km 50
G E R M A N Y
Konstanz
SWITZERLAND
LIECHTENSTEIN
A U S T R I A
Innsbruck
Chur
Muster
Sursilvan
Sutselvan
Surmeiran
Vallader
Zernez
S. Murezi
Puter
Südtirol
Brixen
Bruneck
Bozen
Ladin
Cortina
Corvara
Comèlico
Ert
Belluno
Friulian
Udine
Pordenone
Gorizia
Montalcon
Trieste
Trento
Lugano
Bergamo
Treviso
YUGOSLAVIA
(SLOVENIA)
I T A L Y

Figure 2.1: Distribution of Rhaeto-Romance, taken from **haiman1992**

us for the rest of it. It is nonetheless important to remember that names and definitions posed by researchers are never as simple as they might same, nor do not always correspond to the feelings of the speakers and their own sense of identity. In this case, speakers of these dialects do not feel as though they all belong to some greater unity (**bossong2008**).

2.2 Romansh

The term Roamansh is a collective name referring to the Raeto-Romance dialects spoken in Switzerland which are recognized as a single language. There are five different dialects (Surselvan, Sutselvam, Suermeran, Puter, Vallader), with each having normative grammars and distinct orthographic norms (motivated by the Reformation, for translating the Bible and other religious texts) (**haiman1992**; **bossong2008**).

Romansh was officially acknowledged as a fourth official language in Switzerland (besides German, French and Italian) in a federal referendum that took place in 1938, in the eve of the Second World War, with a whopping majority of 92% Yes votes. It has been shown that this referendum played in the hands of the Raeto-Romans in Graubünden to promote their nationalistic political postulate, but was also instrumentalised by the Swiss federal government to counteract Mussolini's pretensions to enquoteItalian territories in Switzerland (referred to as the Italian irredentism) (**valaer2012**).

Romansh is currently spoken by around 40,000 people (**bundesamt2020**). This number has been diminishing constantly – 30 years ago there were 50,000 speakers (**haiman1992**). There is however hardly a single person who only speaks Romansh. In Switzerland, as in the other regions of Raeto-Romance, there is always a "prestige" language surrounding Raeto-Romance, in which Raeto-Romance speakers are fluent in (**haiman1992**).

2.3 Rumantsch Grischun

2.3.1 Lia Rumantscha

In the last hundred years there have been a Raeto-Romance revival. In Switzerland, a major force in this movement has been the founding of the Lia Rumantscha ("The Romansh League") in 1919, which was also a counter-force to the Italian irredentism [1]. It is an umbrella organisation devoted to promoting and perserving the Raeto-Romance language and culture. Its goals include creating and promoting a common language awareness and identity amongst the Raeto-Romans, be responsible for developing a language standard and language renovation, and generally represent the interests of the Romansh and its speakers, in Graubünden and in the Swiss diaspora (**dazzi2012**).

2.3.2 Rumantsch Grischun

The endeavours of the Lia Rumantscha in the field of language planning and standardization led to the official launching of a pan-Romansh language – "Rumantsch Grischun" (**haiman1992**). Its goal was not to replace the local dialects, but be available for persons, institutions, government agencies, companies etc., that want to use Romansh but require a language variant that would be interrgional and intelligible by speakers of all dialects. The main motivation for planning an interregional standard was the failure of Romansh to establish itself as a fourth national language due to the lack of a written standard, despite the great willingness of the people. The existence of a written standard was intended to Romansch be better respected and incorporated in the canton of Graubünden, as well as on a federal level, and would elevate its prestige in the eyes of its speakers (**schmid1982**).

2.3.3 Properties

Rumantsch Grischun was suggested in 1982 by the Zurich born Romance linguist Heinrich Schmid. It was, however, not the first attempt to harmonize the Romansh dialects. In the 19th century, a high school teacher named Gion Antoni Bühler, made failed attempts to propagate for a *Romonsch fusionau*; in the 1960's a Swiss author from the canton of Graubünden, Leza Uffer, suggested *Interrumantsch*, which was mainly based on the Surmiran dialect, but failed similarly (**liver1999**).

Rumantsch Grischun's success has been hypothesized to be mainly due to the favorable timing – the socioeconomical situation at the time as well as a change in the approach of many Raeto-Romans to their own language; but also due to the fact that Rumantsch Grischun, contrary to previous suggestions of a standard language, is more consistent and balanced between the dialects (**liver1999**). It never consitently favors one dialect over the other.

Without going too much into detail, Rumantsch Grischun favors the greatest common denominator, by taking the word forms common to the three most important written dialects (Sursilvan, Surmiran and Vallader). For instance, in all three dialect the word for "key" is *clav*, hence, this is also the Rumantsch Grischun word for "key". In case the dialects do not agree, it uses the word

[1] The nationalistic claim of lands inhabited by persons who the Italian nationalists saw as ethnic Italians

Sursilvan	Surmiran	Vallader	Rumantsch Grischun	Principle
clav	**clav**	**clav**	**clav** “key”	Greatest common denominator
tschiel	**tschiel**	tschel	**tschiel** “sky”	Majority vote
siat	**set**	**set**	**set** “seven”	"
cor	**cor**	cour	**cor** “heart”	"
vendiu	vendia	vendü	**vendi** “bought”	Favor simplicity
sg./pl. *iert/orts*	**iert/ierts**	üert/üerts	iert/ierts “garden”	"

Table 2.1: Examples for choosing the forms for Rumanstch Grischun, based on **liver1999**

form common to the majority of dialects in a sort of “majority vote”. That way, it never prefers one dialect over the other throughout.

Clarity and transperence also play a major role. This means that forms which exhibit stem alternations, for instance between singular and plural, are abandoned in favor of the simpler, more constant form. Also phenomenons that are specific to just one dialect are left outside, such as the rounded front-vowels [y] and [ø] typical of the dialects of the Engadine or the closing diphthong [w][2], which is unique to Sursilvan (**liver1999**). See table 2.1 for some examples.

This new language fulfills the requirements of its authors: it can be read and understood by any Raeto-Roman without them having to elaborately learn it and the differences to the specific dialects are minimal (**liver1999**).

2.3.4 Today

Rumantsch Grischun has become one of the most ambitious endeavours in the history of Romansh. Since its invention, Romansh and the people promoting it have had notable success achieving their goals. In 1999, Romansh became a “partially official language” (*Teilamtssprache*) of the Swiss confederation. In 2003, it was recognized in the cantonal constitution of Graubünden as an equal cantonal language, and the protection of the traditional language regions was guaranteed. Nowadays, Romansh is in use in many domains, not only in the public administration, but also in economy. Many writings and works were written in Rumantsch Grischun. People learn to read and write in Rumantsch Grischung and in some schools, classes are held in Rumantsch Grischun. The extent of radio and television in Romansh has been growing. There is a radio station broadcasting 24/7, television programs broadcast in all public channels of the Swiss Broadcating Corporation (SSG SSR), as well as internet portals, e.g., `https://www.rtr.ch/`. All of this were not possible if it weren’t for the political “upgrade” that was aspired for by the Romansh language movement (**cathomas2012**).

The canton of Graubünden has been releasing most or all of its press releases since 1997, which build up this corpus, in three languages: German, Italian and Romansh using the Rumantsch Grischun standard.

[2] The diphthong starts with an open vowel [] and ends with a closed vowel [w]

Chapter 3

Compiling the Corpus

3.1 Introduction

The corpus at hand incorporates the press releases published by the Canton of Grisons/Graubünden. These press releases are a means of the cantonal government to publish news and information about topics such as politics, economy, health and culture. Graubünden, which is made up of German speaking, Italian speaking and Romansh speaking regions, is the only trilingual canton in Switzerland. As such, virtually all press releases are published in German, Italian and Romansh. This trilingual setting lends itself to be collected to a parallel trilingual corpus.

3.2 Collecting the Data

At first, I contacted the *Standeskanzlei* ("State Chancellery of Grisons") which is the "the general administrative authority for questions of office, coordination and liaison with the cantonal parliament ('Grosser Rat'), government and cantonal administration" (**staka**). The *Standeskanzlei*, with its *Übersetzungsdienst* ("Translation service"), is responsible for translating documents in service of the canton. I was hoping to receive the data directly from them – after all, we are not talking about private or commercial data, but about public translation work financed with taxpayers' money.

I spoke to Mr. Mirco Frepp from the communication sevices (*Kommunaktionsdienst*), which, although very friendly, had to inform me that it would be impossible for me to receive the data. The explanation was that the documents are not saved locally somewhere, but are saved in a database. The documents are extracted from the database and are generated as ad-hoc HTML documents whenever the website is accessed. It was also not possible to receive a dump of the database.

3.3 Web Scraping

Not being able to receive a dump of the database meant I had to scrape the canton's website, extract the relevant content from the HTML files and construct my own database. In order to achieve this, I wrote a series of Pyhton scripts that would take care of these tasks. All the scripts can be found on my GitHub/Gitlab repository. The scripts relevant for the database building are saved under the folder `corpus_builder`.

```
corpus_builder
└── corpus_builder
│   ├── access_db.py
│   └── create_corpus.py
└── web_scraper
    ├── test_web_scraper.py
    └── web_scraper.py
```

Figure 3.1: Directory tree of `corpus_builder`

```
html
├── 1997
│   ├── 1997_12924_DE.html
│   ├── 1997_12936_IT.html
│   └── ...
├── 1998
├── 1999
├── ...
└── 2022
    ├── 2022_2022010301_DE.html
    ├── 2022_2022010301_IT.html
    ├── 2022_2022010301_RM.html
    └── ...
```

Figure 3.2: Directory scheme for saving the HTML files

Web Scraper

The script `web_scraper.py` goes to the index web page for each year and language. This page contains a links pointing to all the press releases that were released that year. It collects all those links and then downloads the HTML from each link. The HTML pages are saved to seperate folders for each year. The filenames have the following format: `year_file-id_language`, e.g., `1997_12924_DE.html`. The file-id is taken from the URL and will be later used to align the documents.

Since the script makes many requests to the website, one has to expect that server might stop responding, which will result in a request time-out. To avoid downloading HTML pages that were already downloaded, the script will skip and press realease that already exists locally, providing the file size is greater than 0 bytes. This way, the script can be run at a later stage, after new press releases were published, to complete the local repository. To make sure the local copy of the press releases is complete, the script can be simply run until a message is printed that no new press releases were downloaded.

By deafult, the script will download the press releases for the entire year range (1997 to the current year) and in all three languages. This can be limited by using the following optional arguments:

- `--year` – limit the scraping to a year or to a range of years seperated by a comma, e.g., `--year 2022` or `--year 2020,2022`
- `--lang` – limit the scraping to one or more languages (comma seperated), e.g., `--lang de,it`

3.4 Building the Corpus

All the scripts responsible for buidling the corpus can be found under the folder `corpus_builder`.

3.4.1 HTML Parsing

After the creation of a local copy of the HTML files containing the press releases, the content needs to be extracted from the HTML files and saved in a format that would be suitable for later processing.

Using the Python package BeautifulSoup to parse the HTML files, I extracted from each HTML file the title and the text of the press release, as well as some meta data: date, language and the original file-id and the original file name (for debugging purposes). The data was then saved to a JSON file, one per each year. See listing 3.1 for an example.

```
{
    "0": {
        "id": "12924",
        "orig_file": "html/1997/1997_12924_DE.html",
        "lang": "DE",
        "title": "25 Jahre Arge Alp: Graubünden feiert tüchtig mit",
        "date": "31.12.1997",
        "content": "Die Arge Alp feiert heuer ihr 25-Jahre-Jubiläum. Aus diesem Grund finden vom 27. September bis 12. Oktober 1997 die Festwochen des Alpenraums in Telfs-Mösern, Tirol, statt. ..."
    },
    "1": {
        "id": "12926",
        "orig_file": "html/1997/1997_12926_DE.html",
        "lang": "DE",
        "title": "Kanton will auch personelles Engagement bei der Bündner Kraftwerke AG verstärken",
        "date": "31.12.1997",
        "content": "Nachdem der Kanton Graubünden letzten Herbst die Aktienmehrheit der Bündner Kraftwerke AG übernommen hat, will er nun auch seine Vertretung im Verwaltungsrat stärken...."
    },
    "2": {
        "id": "12927",
        "orig_file": "html/1997/1997_12927_DE.html",
        "lang": "DE",
        "title": "Graubünden trifft präventive Massnahmen zur Bekämpfung der illegalen Einwanderung",
        "date": "31.12.1997",
        "content": "Die Fremdenpolizei des Kantons Graubünden trifft im Einvernehmen mit dem kantonalen Sozialamt, dem Amt für Zivilschutz sowie der Kantonspolizei Graubünden Massnahmen, um die illegale Einwanderung in den Südtälern des Kantons Graubünden zu bekämpfen...."
    },
```

Listing 3.1: Example for a JSON file

3.4.2 Document Alignment

After extracting the relevant data from the HTML files and saving them in JSON files, the core task can begin: aligning the documents to get document-triples which are translations of each other.

Linked vs. unlinked

For all releases published after mid-2009 this is pretty simple. The file-id extracted from the URLs is common to all three releases in the three languages (see figure 3.2). This file-id can be used to link the press releases with each other. I shall refer to these press releases as "linked releases".

For releases published prior to that, each release has a unique file-id. This means it can't be used for document alignment. I shall refer to these releases as "unlinked releases". For unlinked releases I used a simple heuristic: if on one single date exactly three releases were published in three different languages, I assume they are translations of each other. The titles of press releases that weren't aligned are saved to a CSV file which can be used for manual alignment.

Unfortunately, this means around 40% (TODO: what is the exact average?) of the releases each year prior to 2009 cannot be automatically added to the corpus, cf. figure 3.4.2.

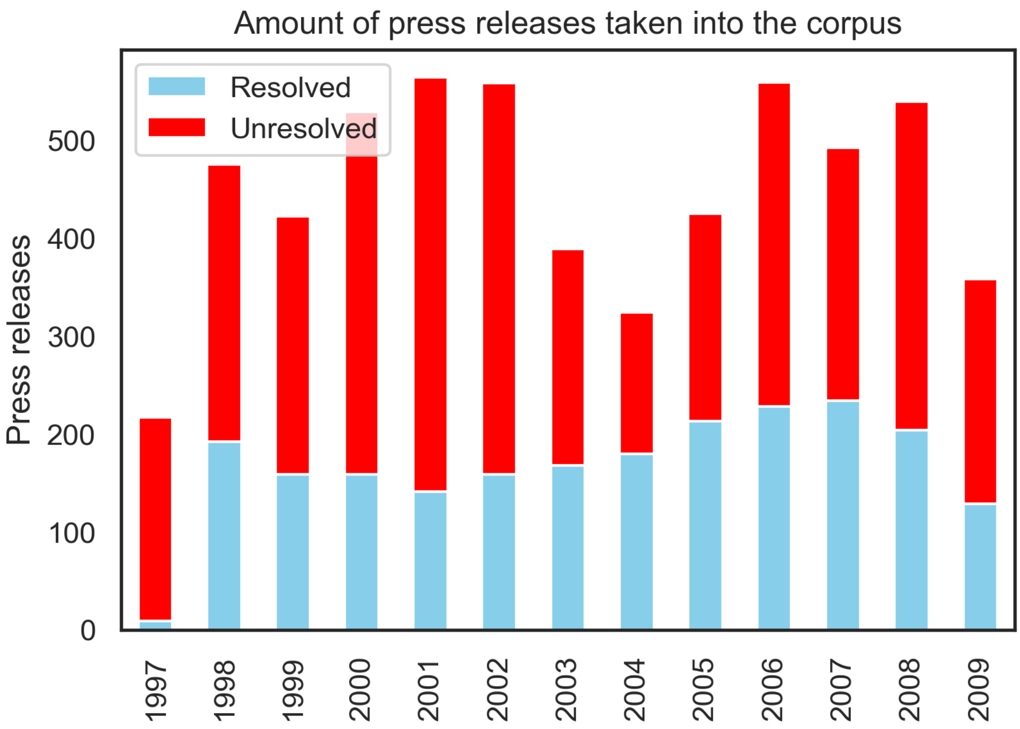


Figure 3.3: Portion of automatically aligned press releases up to 2009

Since the year 2009 contains both linked and unlinked releases, I wrote the script `split_2009.py` to split the data accordingly. It uses a very simple heuristic: if the file-id of a press release is longer than 5 digits, it is a linked press releases.

Aligned corpus

The aligned press releases are saved again to JSON files, with each row containing the three press releases in the three languages, along with metadata such as date and file-id. In the rare case that one language is missing (TODO: how rare?), i.e., the press releases wasn't translated into that language for some reason, it is simply left blank. Press releases that are available only in one language are discarded from the corpus.

The script `create_corpus.py` deals with this task. Using the Python library Pandas, the JSON files are read into a Dataframe. For linked releases, all the unique ID's are taken, and then for each ID the three languages are collected and saved into a new row. The dates are converted from their original format (DD.MM.YY) to an ISO 8601 format (YYYY-MM-DD) (**enwiki:1095673391**) for better compatibility and easier processing later.

For JSON files contatining unlinked documents, the script `create_corpus` has to be run with the switch `--by-date`, which tells the program to use the date for aligning the documents instead of the file ID.

For an example of the resulting JSON files, with each row containing the aligned documents, see listing 3.2.

```
{
    "0": {
        "id": "2010010501",
        "date": "2010-01-05",
        "DE_title": "Neues Online-Angebot für das Bündner Rechtsbuch",
        "DE_content": " Das im Internet verfügbare Bündner Rechtsbuch ist neu gestaltet
            worden und enthält neue Funktionalitäten. ...",
        "IT_title": "Nuova offerta online per la Collezione sistematica del diritto
            cantonale grigionese",
        "IT_content": " La Collezione sistematica del diritto cantonale grigionese
            disponibile in internet è stata ristrutturata e contiene nuove funzioni. ...
             ",
        "RM_title": "Nova purschida d'internet per il cudesch da dretg grischun",
        "RM_content": " Il cudesch da dretg grischun che stat a disposiziun en l'
            internet ha survegnì in nov concept e novas funcziuns. ... "
    },
    "1": {
        "id": "2010010502",
        "date": "2010-01-05",
        "DE_title": "Staupe bei Füchsen und Dachsen im Puschlav",
        "DE_content": " Nachdem sich im Verlaufe des letzten Herbstes die Staupe-
            Krankheit bei Wildtieren in Nord- und Mittelbünden verbreitete, sind im
            Laufe der letzten Wochen nun auch im Puschlav bei Füchsen und Dachsen
            Infektionen mit dem Staupevirus nachgewiesen worden. ... ",
        "IT_title": "Volpi e tassi affetti da cimurro in Valposchiavo",
        "IT_content": " Dopo che nel corso dell'autunno il cimurro si è diffuso tra gli
            animali selvatici del Grigioni settentrionale e centrale, nelle ultime
            settimane la presenza del virus è stata rilevata anche tra volpi e tassi
            della Valposchiavo. ...  ",
        "RM_title": "Pesta dals chauns tar vulps e tar tass en il Puschlav",
        "RM_content": " Suenter che la pesta da chauns è sa derasada tar la selvaschina dal Grischun dal nord e central en il decurs da l'atun passà, èn vegnidas cumprovadas en il decurs da las ultimas emnas ussa er infecziuns cun il virus da questa malsogna tar vulps e tar tass en il Puschlav. ... "
    },
    "2": {
        "id": "2010010801",
        "date": "2010-01-08",
        "DE_title": "Projekt Sicherheitsfunknetz POLYCOM Graubünden mit Vertragsunterzeichnung offiziell gestartet",
        "DE_content": " Die Vorsteherin des Departements für Justiz, Sicherheit und Gesundheit, Regierungsrätin Barbara Janom Steiner, und der Chef des Grenzwachtkorps, Jürg Noth, haben heute in Chur eine Vereinbarung zur Realisierung des Sicherheitsfunknetzes POLYCOM im Kanton unterzeichnet. ... ",
        "IT_title": "Avviato ufficialmente con la sottoscrizione del contratto il progetto di rete radio di sicurezza POLYCOM Grigioni",
        "IT_content": " La Consigliera di Stato Barbara Janom Steiner, direttrice del Dipartimento di giustizia, sicurezza e sanità, e il capo del Corpo delle guardie di confine, Jürg Noth, hanno sottoscritto oggi a Coira un accordo per la realizzazione nel Cantone della rete radio di sicurezza POLYCOM. ... ",
        "RM_title": "Il project per la rait radiofonica da segirezza POLYCOM dal Grischun è vegnì lantschà uffizialmain cun suttascriver il contract",
        "RM_content": " La scheffa dal departament da giustia, segirezza e sanadad, cusségliera guvernativa Barbara Janom Steiner, ed il schef dal corp da guardias da cunfin, Jürg Noth, han suttascrit oz a Cuira ina cunvegna per realisar la rait radiofonica da segirezza POLYCOM en il chantun. ... "
    },
}
```

Listing 3.2: Example for a JSON file containing aligned documents

3.5 Manual alignment of unlinked documents

As can be seen in figure 3.4.2, a big portion of the unlinked documents cannot be automatically aligned using the simple heuristic described in section 3.4.2. To deal with that, the script `create_corpus.py` will write the titles and file IDs of all the discarded documents to a CSV file, one for each year.

These CSV files can be used for manually aligning the corresponding documents. This is done by enumerating the documents while using the same digit for corresponding documents. The CSV files containing the now enumerated documents can be given to the script `create_corpus.py` using the argument `add-from-csv` to combine the enumerated documents as linked documents into the JSON file.

3.6 SQLite database

The query language SQL offers flexible and complex way to query datbases. For this reason, I decided to save the resulting corpus in an SQLite database. I opted for SQLite because it doesn't require running a seperate server and can be SQLite databases can be easily built, edited and accessed using `sqlite3`[1], a Python module delivered in the Python standard library[2].

The SQLite database contains two tables, `corpus` and `raw` with the exact same structure as the two JSON files described in listings 3.1 and 3.2.

3.7 Summary

For compiling the corpus, the following steps were taken:

1. Scrape website and save HTML documents locally
2. Extract relevant content from HTML files (date, language, title and content) and save to JSON files
3. Read the JSON files using Pandas Dataframe, align the documents and save to new JSON files
4. Feed both types of JSON files to an sqlite database

The final result is an sqlite database (`corpus.db`) containing two tables:

- `corpus`: all the aligned documents from 1997 until today. Each row contains following columns:
  - id: automatically incremented unique ID
  - file_id: original file ID
  - date: Release date
  - DE_title: Title of German document
  - DE_content: Content of German document
  - IT_title: Title of Italian document
  - IT_content: Content of Italian document
  - RM_title: Title of Romansh document
  - RM_content: Content of Romansh document
- `raw`: all the documents contained in the HTML files scraped from the website. Each row contains the following columns:
  - id: Automatically incremented unique ID
  - file_id: Original file ID
  - orig_file: Original filename
  - lang: Document language (DE for German, IT for Italian, RM for Romansh)

[1] `https://docs.python.org/3/library/sqlite3.html`

[2] `https://docs.python.org/3/tutorial/stdlib.html`

- title: Document title
- date: Release date
- content: Document content

Chapter 4

Sentence Alignment

4.1 Introduction

The corpus presented in chapter 3 is a raw parallel corpus, that is it is a corpus of aligned documents without any further processing. In order to use the corpus for tasks such as training a machine translation model, another processing step is needed: sentence alignment (**koehn2009**).

Formally, the task can be described as follows: We have a list of sentences in language e, $e_1, ...e_{n_e}$ and a list of sentences in language f, $f_1, ..., f_{n_f}$. (Note that the number of sentences in each language is not necessarily identical.) A sentence alignment S consists of a list of sentence pairs $s_1, ..., s_n$, such that each sentence pair s_i is a pair of sets:

$$s_i = (\{e_{\text{start-e}(i)}, ..., e_{\text{end-e}(i)}\}, \{f_{\text{start-f}(i)}, ..., f_{\text{end-f}(i)}\})$$

(**koehn2009**)

This means each set in the pair of set can consist of one or more sentences. The numer of sentences in each set is referred to as alignment type. A 1-1 alignment is an alignment where exactly one sentence of language e is aligned to exactly one sentence of language f. In a 1-2 alignemnt, one sentence in lanauge e is a aligned to two sentences in langauge f. There are also 0-1 alignments, in which a sentence of language f is not aligned to anything of language e. Sentences may not be left out and each sentence may only occur in one sentence pair (**koehn2009**).

4.2 Methods

One early method for sentence alignment is the one described in **gale-church-1991-program** which is “based on a simple satistical model of character lengths” (**gale-church-1991-program**). The method arose out of the need to design a faster, computationally more efficient algorithm[1].

The method uses the fact that longer sentences in language e are usually translated into longer sentences in language f and vice-versa – shorter sentences correspond to shorter sentences.

The method combines a distance measure based on the lengths of the sentence with a prior probility of the alignment type (1-1, 1-0 or 0-1, 2-1 or 1-2, 2-2) to a probabilistic score. It assigns

[1] With the algorithms that existed up to that time, it took 10 days to extract 3 million sentence pairs, 12,500 sentences per hour.

this score to possible sentence pairs in a dynamic programming framework to find the best (most probable) pairs. A program based on this method was tested against a human-made alignment on two pairs of languages: English-German and English-French. The program made a total of 55 errors out of a total of 1316 alignments (4.2%). By taking the best scoring 80% of the alignments, the error rate could be reduced to 0.7%

The method was also much faster than the algorithms that existed up to that time. It took 20 hours to extract around 890,000 sentence pairs, around 44,500 sentence pairs per hour, around 3.5 times faster than former algorithms.

4.3 Statistical methods

4.4 Recent methods

Chapter 5

Gold standard

5.1 Introduction

In order to measure the quality of words alignments, a model's performance is measured on a test set which is a gold standard created by human annotators. For the gold standard to be of good quality and consistent with itself, annotators have to follow strict guidelines. These guidelines address issues of ambiguity in word alignments. (**koehn2009**).

Some problematic cases that might occur are function words (TODO) that have no clear equivalent in the other language. **koehn2009** gives as an example the German-English sentence pair: *John wohnt hier nicht John does not live here*. What German word should the English word *does* be aligned to? Three different choices can be made:

1. The word should remain unaligned since it has to clear equivalent in German.
2. The word *does* is connected with *live*; it contains the number and tense information which is in German contained in one word *wohnt*, so it should be aligned to *wohnt*, together with *live*.
3. *does* is part of the negation; without it, the sentence would not contain this word. Therefore, *does* should be aligned with *nicht* (the German negation).

5.2 Sure and Possible Alignments

An approach for solving problematic cases is the distinction between *sure* (s) and *possible* (p) alignments (**och-ney-2000-improved**), which are also sometimes referred as fuzzy alignments (**clematide2018**). Generally, these labels allow to distinguish between ambiguous and unambiguous links. Ambiguous links are labeled *possible* and unambiguous links are labeled *sure* (**lambert2005**). The *possible* label was conceived to be used especially for aligning words within idiomatic expressions, free translations and missing function words (**och-ney-2000-improved**). This distinction also has an impact on the way the evaluation metrics are computed (more on that later).

There seems to be no clear global definition about which alignments should be considered as umabiguous and marked as *sure* and which should be considered ambiguous marked as *possible*. For some created gold standards, no distinciton between *sure* and *possible* alignments was made (**clematide2018**). In another case, annotators were asked to first label all alignments as *sure* and then refine their alignments with confidence labels (**holmqvist-ahrenberg-2011-gold**). In the creation of the English-Icelandic gold standard in **steingrimsson-etal-2021-combalign**, annotators used only *sure* links. Their annotations were then combined, with all 1-1 alignments both annotators agreed upon (i.e., the intersection of their annotations) makred as *sure* and differences all other alignments made by either one or both were marked as *possible* (**steingrimsson-etal-2021-combalign**).

5.3 Evaluation Metrics

TODO: move to results/evaluation part Four types of measures have become standard for evaluating word alignment. Three of them – precision, recall and F-measure – are well known in Information Retrieval metrics **mihalcea-pedersen-2003-evaluation**. The fourth, alignment error rate (AER) one was introduced by **och-ney-2000-improved**.

5.4 Gold standard for German-Romansh

In order to measure the performance of both models, the embedding based model (SimAlign) and the stastitical model (fast_align), on the language pair German-Romansh a gold standard is needed. Since no such gold standard exists, I took upon myself to create one. Although I am not a speaker of Romansh, my experience as a trained linguist, as well as my knowledge in related languages (Latin, Italian, French), allows me to confidently tackle this task. Additionally, whenever I was in doubt, I referred to the online dictionary Pledari grond, which also offers a grammar overview. (TODO: add more grammar references)

5.4.1 Annotation tool

I used the tool *AlignMan* which was originally programmed for creating the gold standard for English-Icelandic by **steingrimsson-etal-2021-combalign**. It is quite easy to use and its code is readable. I also had to make some small changes to the code. For instance, the sentences to be aligned, while loaded into the database, were read in opposite order, such that the source language became the target language and vice versa. I fixed this issue, so that source (German) and target (Romansh) languages stay the same accross all applications.

As mentioned above, the tool does not allow labeling of links with *Sure* and *Possible*. Instead, AlignMan treats the union of 1-1 alignments made by two annotators as *Sure* alignments and all other alignments as *Possible*. This means, each annotator is expected to only annotate *Sure* alignments, which also applied to me while annotating the German-Romansh gold standard.

5.4.2 Guidelines

As mentioned above, clear guidelines need to be defined for creating the gold standard in order to ensure quality and consistency. I shall now proceed to describe the guidelines I used for my annotation of the word alignments for the gold standard.

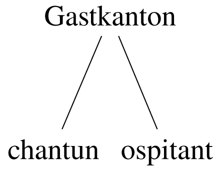
A motto cited often for annotating word alignments is “Align as small segments as possible, and as long segments as necessary” (**Vronis00evaluationof**, cited in **lines2007**). A variation of this is found in **clematide2018**: “as few words as possible and as many words as necessary that carry the same meaning should be aligned.”, referring to **lambert2005**.

In the following sections I will list some general principles as well as more specific principles involving German and Romansh.

5.4.3 General priniciples

Principle I. Use only *Sure* alignments. Since the annotating tool I was using does not provide the use of confidence labels (cf. section 5.4.1), I only aligned words which would be considered *Sure* alignments, i.e., they are unambiguous (cf. section 5.2).

Principle II. Prefer 1-1 alignments over 1-n alignments or n-n alignments. Since all alignments are seen as *Sure* alignments, 1-n alignments should be avoided, unless a single word in the source sentence lexically corresponds to several words in the target sentence (see TODO principle sth.) This means alignments of phrases should be avoided. This is also due to the fact that we are testing models for automatic word, and not phrase alignments.



Words that are repeated in one language, but not in the other, should only be linked once, leaving the repetition unaligned.

Principle III. Lexical alignments should always be preferred over all other alignments (part-of-speech alignments or morphosyntactical alignments). This means alignments should describe first and foremost lexical correspondences, i.e., they have the same lexical meaning (but not necessarily share the same grammatical function or the same part-of-speech). Only words that are translations of each other also outside of the specific context of the sentence pair at hand should be aligned. This is in line with **clematide2018**. In cases of paraphrasing during translations, words should remain unaligned (TODO: example?)

- only sure alignments
- prefer 1-1 alignments over 1-n alignments
- align words, not phrases
- only align words that are translations of each other also outside of context
- POS doesn’t matter: German often prefers a nominal style, Romansh prefers a verbal style – expect some noun-verb alignments.

German	Romansh	
Beratungsstelle	*post* ***da*** *cussegliaziun*	“consultation point”
Gebäudeversicherung	*Assicuranza* ***d’*** *edifizis*	“building insurance”
Webseite	*pagina* ***d’*** *internet*	“web site”
Kindermasken	*mascrinas* ***per*** *uffants*	“children masks”
Brandversicherung	*assicuranza* ***cunter*** *fieu*	“fire insurance”
Gastkanton	*chantun ospitant*	“hosting canton”

Table 5.1: Translation examples of German compounds into Romansh

5.4.4 Examples

I will now supply some examples to illustrate the above principles.

Compound words

Compounding is the formation of new lexemes by adjoining two or more lexemes (**bauer1988**). In German, compounds are productive and prominent means of word formation in German (**clematide2018**). In a sample of 4,500 types examined by **clematide2018**, 80% of German nouns were compounds. Romansh, in comparison, uses prepositions (usually *da*) for linking nouns, with one noun modifying the other (**valladers**). Other prepositions that can be found for linking words are *cunter* and *per*. [1] In other cases, German compounds might be translated to Romansh using an adjective + noun, e.g., German *Gastkanton* was translated to *chantun ospitant* “hosting canton”. See table 5.1 for examples.

German compounds will be aligned to their equivalent lexical words, but not to function words, resulting in a 1-n alignment: *Webseite ~pagina [d’] internet*, *Gebäudeversicherung ~Assicuranza [d’] edifizis*. This is also inline with principles I, II and III in **clematide2018**.

German preterite vs. Romansh perfect

In the corpus at hand, two tenses are used in German for referring to past events: the preterite and the perfect. The German preterite is a synthetic verb form, i.e., it is made up of a single conjugated form. Some examples are *nahm* (infinitive *nehmen* “take”) or *wurde* (infinitive *werden* “become”). The German perfect is an analytic construction made up of an auxiliary verb (*haben* “have” or *sein* “be”) and the past participle, e.g., *Die Präsidentenkonferenz* ***hat*** *nun* ***entschieden*** “The conference has decided”.

In contranst to German, Romansh only has one tense referring to past events: the perfect. It is an analytic construction made, in a similar fashion as in German, of an auxiliary *habere* “have” for transitive verbs or *esse* “be” for intransitive verbs and the past participle (**bossong2008**). The German sentence given above (*Die Präsidentenkonferenz hat nun entschieden*) was translated as *La conferenza da las presidentas e dals presidents* ***ha*** *usse* ***decidi***. *ha* is the auxiliary and *decidi*

[1] Typologically, this is inline with other Romance languages such as French, which uses prepositions (*de*, *en* and *à*) for linking two nouns, e.g., *une robe de soie* “a silk dress” (**price2008**)[510].

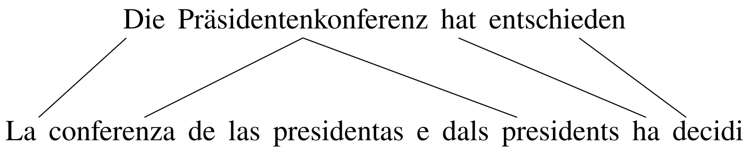

Die Präsidentenkonferenz hat entschieden
La conferenza de las presidentas e dals presidents ha decidi

Figure 5.1: Aligning German perfect to Romansh perfect

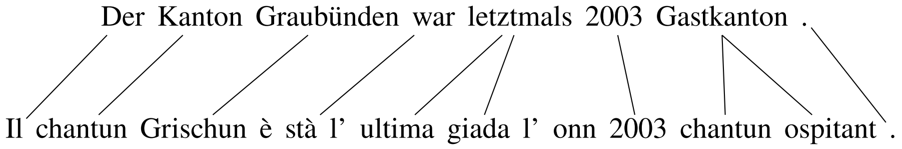

Der Kanton Graubünden war letztmals 2003 Gastkanton .
Il chantun Grischun è stà l' ultima giada l' onn 2003 chantun ospitant .

Figure 5.2: Alignment of German preterite to Romansh perfect

is the past participle. This poses no real problem since we can link the German auxiliary to the Romansh auxiliary and the German participle to the Romansh participle.

However, a German preterite is always translated using the Romansh perfect. For example, in the sentence *Der Kanton Graubünden war letzsmals 2003 Gastkanton* "The last time the Canton of Grisons was a host canton was in 2003" the verb *war* "was" is translated as *è stà*. This theoretically results in a 1-2 link. However, since the verb *è* here only carries grammatical information of tense and number, but no real lexical information, it should reamin unaligned.

The German perfect should be aligned to the Romansh perfect using a 1-1 alignment; auxiliary to auxiliary and participle to participle. **The German preterite should also be aligned using a 1-1 alignment to the Romansh participle, leaving the auxiliary unaligned and avoiding a 1-2 alignment.**

German present participle

German present participles (known in German as *Partizip I*) are translated to Romansh using relative clauses. Moreover, adjectives (and participles in the function of adjectives), can be nominalized, meaning they become the head of a noun phrase and there is no need for an actual noun. A good example for that in the corpus is the German noun phrase *nichtarbeitslose Stellensuchende* (cf. ex. 1), which was translated as a noun phrase with a relative clause: *persunas che tschertgan ine plazza che n'èn betg dischoccupadas* "persons who look for a job who are not unemployed".

(1) nicht-arbeit-s-los-e Stellen-such-end-e
not-work-gen-less-Pl job-search-pres.part-pl
"People looking for jobs who are not unemployed"

In this case, these two phrases should not be aligned as phrases, but only the content words

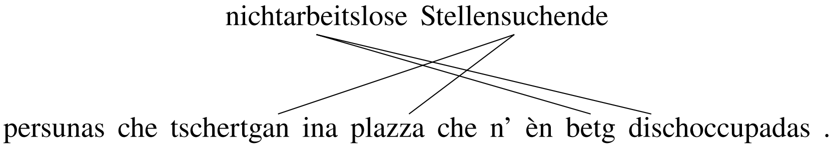

nichtarbeitslose Stellensuchende
persunas che tschertgan ina plazza che n' èn betg dischoccupadas .

Figure 5.3: Aligning German present participles to Romansh relative clauses

which lexically correspond to each other: *nichtarbeitslose* ~ *betg dischoccupadas*; *Stellensuchende* ~ *tschertgan [ina] plazza*.

Double negation

Negation in Romansh is built using two particles: *na* and *betg* to negate verbs or *nagin-* to negate nouns. Since we prefer 1-1 alignments, the German negations *nicht* (for verbs) and *kein-* for nouns should be aligned only to the second Romansh particle (*betg/nagin-*), leaving Romansh *na* unaligned. Granted, this is also in favor of the SimAlign output, but it is also linguistically motivated: when negating the imperative form, *na* can be omitted required TODO:cite Grammatica per linstrucziun dal rumantsch grischun.

Articles and prepositions

German articles inflect in case, which expresses some syntactic relations between nouns. Romansh often uses preopsitions for expressing the same relations. For instance *Zustimmung der Person* "the person's agreement" is translated as *consentiment da la persuna*. I align the German article *der* with Romansh *da*, leaving *la* unaligned. Except for my preference for 1-1 alignments, the motivation for this is that it is the preposition *da* that expresses the genitival relations between the nouns.

Separable verbs

German uses many verbs to which an adverb or a preopisition is affixed in order to delimit the verb's meaning (or sometimes completely change its meaning). In such cases, both the verb and its affix should be aligned to the corresponding Romansh verb, resulting in a 2-1 alignment.

5.5 Flaws

I shall now discuss the quality of my gold standard and some flaws it has.

The most obvious flaw is the fact that I created the gold standard alone. With more than one annotator, more intricate annotating schemes can be used in order to ensure higher quality, consistency and harmony. For instance the annotators' agreement can be measured using the so-called inner-annotator agreement (**holmqvist-ahrenberg-2011-gold**). Further, the intersection of the annotators' *Sure* alignment can be used to build the final *Sure* alignments set and the reunion of the *Possible* alignments can be used to create the final *Possible* alignments set **mihalcea-pedersen-2003-evaluation**. A third annotator can also revise and resolve conflicts between two annotaors **mihalcea-pedersen-2003-evaluation**. When several annotators work on the same task, they can also discuss conflicts and resolve them using a majority vote (**DBLP:journals/corr/cmp-lg-9805004**).

All of these possible schemes cannot be realized in my case.

Another flaw is the missing confidence labels (*Sure* and *Possible*), which may influence the evaluation scores. Doing without *Possible* links and using only *Sure* links is however precedented (**clematide2018**; **mihalcea-pedersen-2003-evaluation**) and hence defensible.

In order to test my own consistency, I have re-annotated the first 100 sentences in the sample. TODO: results

Despite of the flaws mentioned, I am certain that gold standard is of high quality and consistency, due to the fact that I was also the one to define the guidelines.

Chapter 6

Evaluation

Chapter 7

Contributions

Chapter 8

Summary