Arbeitsbericht: Maschinelle Übersetzung für das Gadertalische

Samuel Frontull, Tobias Hell

1. Einführung

In den letzten Jahren wurden auf dem Gebiet der Computerlinguistik bahnbrechende Fortschritte erzielt. Dank der enormen Mengen verfügbarer digitaler Daten, fortschrittlicher Algorithmen für das maschinelle Lernen und der Steigerung der Rechenleistung konnten Sprachmodelle entwickelt werden, die nun in der Lage sind, semantische Informationen zu verarbeiten, zu verstehen und zu übertragen. Dies gilt jedoch nur für "privilegierte" Sprachen. Viele hochmoderne Technologien unterstützen nur die meistgesprochenen Sprachen der Welt, da sie auf datengesteuerten Ansätzen beruhen, die für weniger populäre Sprachen mit geringer Datenverfügbarkeit ungeeignet sind. Auch Ladinisch kann in diesem Kontext als eine "low-resource"-Sprache betrachtet werden.

Das Ladinische, das in der Region um das Sellamassiv (Dolomiten) von etwa 30.000 Menschen gesprochen wird, ist eine offiziell anerkannte Minderheitensprache, die heute durch verschiedene Maßnahmen geschützt und gefördert wird. Sie findet ihren Platz in der öffentlichen Verwaltung, in Schulen und in den Medien. Dennoch wurde sie von der UNESCO als "vom Aussterben bedroht" eingestuft, vor allem weil es immer schwieriger wird, die Sprache an die nächsten Generationen weiterzugeben. "Der Schlüssel zum Überleben kleinerer Sprachen liegt darin, sie überall, zu jeder Zeit und in allen möglichen Kontexten zu verwenden" (Minority SafePack Initiative 2022). Im digitalen Zeitalter, in dem wir leben, ist dies jedoch ein echtes Problem für kleinere Sprachen (man denke nur an die Spracherkennung).

Ladinia XLVI (2022), 203–232; DOI 10.54218/ladinia.46.203-232.
ISSN 1124–1004;

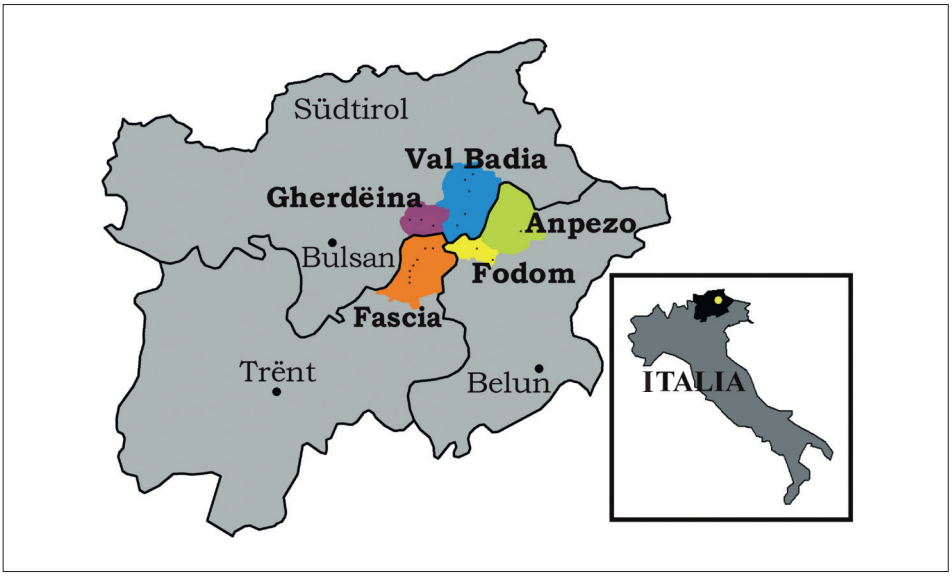


Fig. 1: Die ladinischen Täler[1].

Die Verfügbarkeit eines automatischen Übersetzungssystems für solche Sprachen würde daher nicht nur deren Zugang erleichtern, sondern auch ihre Existenz in der immer wichtiger werdenden digitalen Welt gewährleisten. Ziel der nächsten Jahre ist es deshalb, ein maschinelles Übersetzungsmodell für die ladinische Sprache zu entwickeln. Erste Schritte, die in diese Richtung bereits gemacht wurden, werden in diesem Bericht dokumentiert.

Dieses Thema ist besonders interessant, weil das Ladinische nicht nur wenig verbreitet, sondern auch in mehrere Varianten unterteilt ist, die sich von Tal zu Tal unterscheiden (Fig. 1). Außerdem wird die Schriftsprache nur von wenigen vollständig beherrscht. Daher sind im Internet kaum verwertbare Daten zu finden. Im Gegensatz zu vielen anderen "low-resource"-Sprachen wird die ladinische Sprache jedoch gefördert, ist bereits gut erforscht und hat eine Vielzahl an Publikationen vorzuweisen, darunter auch Wörterbücher.

Da es sich um den ersten Beitrag dieser Art für die ladinische Sprache handelt, gibt es noch keine vergleichbaren Ergebnisse. Es werden deshalb zwei verschiedene Ansätze evaluiert, um möglichst viele Eindrücke zu sammeln und ein Gefühl dafür zu bekommen, in welche Richtung sich die zukünftige Arbeit am Problem

[1] <https://commons.wikimedia.org/wiki/File:Ladin.png>, [31.03.2022].

einer maschinellen Übersetzung für "low-resource"-Sprachen entwickeln könnte. Da sich die Varianten des Ladinischen bereits auf syntaktischer Ebene stark unterscheiden, muss dieses Problem für jede einzelne Variante separat angegangen werden. Die Variante des Gadertals wurde als Startpunkt gewählt.

2. Verwandte Arbeiten

Die Forschung der maschinellen Übersetzung für Sprachen mit wenigen Ressourcen hat in den letzten Jahren starkes Interesse geweckt. Die jüngsten Studien von Haddow (2021), Hedderich (2020) und Ranathunga (2021) geben einen Überblick über die neuesten Entwicklungen in diesem Bereich. Im Folgenden werden wir die in diesen Studien beschriebenen Ansätze zusammenfassen und die Anwendbarkeit auf die ladinische Sprache diskutieren. Wir kategorisieren die verschiedenen Beiträge – wie in (Haddow 2021) – in drei Hauptkategorien: 1. Datensammlung, 2. Datenverwertung und 3. Modellauswahl.

2.1 Datensammlung

Für die Entwicklung eines guten maschinellen Übersetzungssystems mit *State-of-the-art*-Methoden ist im Prinzip nur eine hochwertige und umfangreiche Datenbasis erforderlich. In ressourcenarmen Szenarien sind jedoch nicht genügend Daten verfügbar, um mit modernen, datengetriebenen Ansätzen die gewünschten Ergebnisse zu erzielen. Da diese datengetriebenen Ansätze in den letzten Jahren so populär geworden sind, wurde viel darüber geforscht, wie man sie auch in ressourcenarmen Szenarien anwenden könnte. In mehreren Studien wird beschrieben, wie z.B. Daten aus dem Internet oder über *Crowdsourcing* automatisch gesammelt und abgeglichen werden können (cf. Haddow 2021). *Opus*[2], *WikiMatrix*[3] oder *Bible*[4] sind frei verfügbare Datensammlungen für über 500 Sprachen. Leider ist die ladinische Sprache, genauer gesagt die Variante des Gadertals, in den bestehenden Datensammlungen nicht enthalten. Die Bibel, die sonst eine beliebte Quelle ist, wurde bisher nur in das Grödnerische übersetzt. Auch die Möglichkeiten, automatisch Daten aus dem Web zu holen, sind sehr begrenzt. Verschiedene maschinelle Übersetzungssysteme werden in der Regel mit denselben Testdaten evaluiert, um einen Vergleich zu ermöglichen. Idealerweise sind die Testdaten

[2] <https://opus.nlpl.eu/>, [31.03.2022].

[3] <https://github.com/facebookresearch/LASER/tree/main/tasks/WikiMatrix>, [31.03.2022].

[4] <https://github.com/christos-c/bible-corpus>, [31.03.2022].

so strukturiert, dass möglichst viele Aspekte der Sprache getestet werden, sodass die Auswertungen auf den Testdaten einen ehrlichen Eindruck von der allgemeinen Leistung vermitteln. Die Erstellung solcher Testdaten ist daher ein wichtiges Thema, an dem auch geforscht wird und das besonders für Sprachen mit geringen Ressourcen von Bedeutung ist. In unseren Experimenten wurden die Testdaten aus einem Wörterbuch und zwei Kinderbüchern entnommen. Für eine bessere generelle Aussagekraft der Testdaten wäre die Erstellung von umfangreicheren Testdaten zusammen mit Experten der Sprache fundamental.

2.2 Datenverwertung

Zu dieser Kategorie gehören Ansätze, die versuchen, ein- oder mehrsprachige Daten und Modelle sowie zweisprachige Lexika und linguistische Werkzeuge zu nutzen, um die begrenzte Menge an verfügbaren parallelen Daten auszugleichen. Eine erfolgreiche Anwendung einsprachiger Daten im ressourcenarmen Bereich ist die Generierung synthetischer Daten durch die sogenannte Rückübersetzung *(backtranslation)*. Die Idee dahinter ist einfach: Einsprachige Texte in der Zielsprache werden mit dem besten verfügbaren Übersetzungsmodell zurück in die Ausgangssprache übersetzt. Auf diese Weise werden neue synthetische Paralleldaten erzeugt. Dieser Ansatz funktioniert auch iterativ. Alternativ können synthetische Daten auch durch Anpassung vorhandener paralleler Daten erstellt werden. Es wäre interessant, beide Ansätze auch für die ladinische Sprache zu evaluieren.

2.3 Modellauswahl

Es geht nicht nur um die Daten selbst, sondern auch darum, wie sie am besten genutzt werden können. In verschiedenen Beiträgen wird die Wirkung verschiedener Modelle und Konfigurationen analysiert. Dabei geht es mitunter nur um unscheinbare Feinheiten, aber auch grundlegende Methoden werden hinterfragt, etwa Suchverfahren, Trainingsziele oder Architekturen. Obwohl es sich um das älteste Paradigma der maschinellen Übersetzung handelt und dieses statistischen Ansätzen in Bezug auf Genauigkeit und Sprachfluss unterlegen ist, ist die regelbasierte Übersetzung eine gute Alternative, wenn das notwendige Expertenwissen vorhanden ist und die zu übersetzenden Sprachen verwandt sind. Wir sind der Meinung, dass dies bei Ladinisch–Italienisch der Fall ist, und da ein Wörterbuch zur Verfügung steht, kann dieser Ansatz definitiv auch für diese Sprachen und für die Übersetzung zwischen den Varianten des Ladinischen evaluiert werden.

3. Datengrundlage und Evaluierungsmethode

In diesem Kapitel beschreiben wir die Datengrundlage der Experimente und erörtern die BLEU-Metrik *(Bilingual Evaluation Understudy)*, die zur Evaluierung verwendet wird.

3.1 Parallele Daten

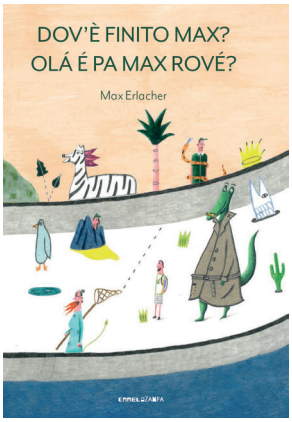
Die Grundlage (Trainingsdaten) von datengesteuerten Algorithmen für die maschinelle Übersetzung sind so genannte *parallele* Daten. Parallele Daten sind eine Sammlung von Sätzen in einer Ausgangssprache und deren Übersetzung in der Zielsprache. Moderne maschinelle Übersetzungssysteme und Sprachmodelle werden auf mehreren Millionen solcher Sätze trainiert. Was das in diesem Projekt untersuchte Problem jedoch so herausfordernd macht, ist die Tatsache, dass die Menge der für das Gadertalische verfügbaren parallelen Daten sehr begrenzt ist. Die einzigen Quellen sind im Grunde das Wörterbuch *Ladin Val Badia – Talian* (Moling 2016) und einige wenige Publikationen.

Trainings- und Testdaten
Das Wörterbuch *Ladin Val Badia – Talian* (Fig. 2a) enthält rund 18.000 parallele Sätze. Wir haben aus diesem Datensatz tausend Sätze für das Testen entnommen und den Rest als Trainigsdaten für den statistischen Ansatz verwendet. Auch der Wörterbuch-basierte Ansatz stützt sich auf dieses Corpus. Da die Sätze in diesem Wörterbuch eher einfach und meistens im Präsens geschrieben sind, können sie – vor allem beim statistischen Ansatz – wenig über die generelle Performanz des Übersetzungssystems aussagen. Deshalb wurden zusätzlich noch Testdaten aus den Kinderbüchern *Les aventöres de Pinocchio* (Collodi 2017) – in Fig. 2b abgebildet – sowie *Dov'è finito Max? Olá é pa Max rové?* (Erlacher 2021) – in Fig. 2c abgebildet – gesammelt. Somit kann eine *in-domain* (Testdaten ähnlich zu den Trainigsdaten) und eine *out-of-domain* Performanz (Testdaten unterscheiden sich stark von den Trainingsdaten) evaluiert werden. Die Übersetzungsqualität in "ressourcenarmen Szenarien" fällt bei *out-of-domain* Testdaten in der Regel stark ab (cf. Shen 2019), deshalb haben diese Testdaten auch eine größere Aussagekraft.

Die genaue Anzahl der gesammelten Sätze wird in Tab. 1 angegeben. Fig. 3 listet einige Satzpaare aus diesen Testdaten auf. Die generierten Übersetzungen dieser Sätze werden später für die jeweiligen Ansätze aufgelistet.

a) Moling 2016

b) Collodi 2017



c) Erlacher 2021

Fig. 2: Quellen der Trainings- und Testdaten.

Quelle	Sätze
Dizionar *Ladino Val Badia – Talian* (Moling 2016)	1.000
Les aventöres de Pinocchio (Collodi 2017)	3.127
Dov'è finito Max? Olá é pa Max rové? (Erlacher 2021)	101

Tab. 1: Größe der Testdatensätze.

3.2 Die BLEU-Metrik

Die Bewertung der Leistung eines maschinellen Übersetzungssystems muss vor allem in der Entwicklungsphase kostengünstig und schnell sein. Eine automatische Bewertungsmethode ist daher unerlässlich. BLEU (Papineni 2002) ist eine weit verbreitete Evaluierungsmethode, die diesen Anforderungen gerecht wird. Sie untersucht, wie nahe eine Übersetzung an einer Referenzübersetzung liegt, indem sie die Anzahl der Überschneidungen von *n*-Grammen (typischerweise 1-, 2-, 3- und 4-Gramme)[5] zählt. BLEU hat sich zu einer Standardmetrik zur automatischen Bewertung der Qualität eines maschinellen Übersetzungssystems entwickelt. Sie ist nicht nur kostengünstig, sondern auch sprachunabhängig und hat eine hohe Korrelation mit der menschlichen Bewertung (cf. Papineni 2002). Darüber hinaus kann diese Metrik auch dazu verwendet werden, Vergleiche maschineller Übersetzungssystemen für verschiedene Sprachen anzustellen.

[5] Cf. <https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu-detok.perl>, [31.03.2022].

[ITA] Moling 2016	[LVB] Moling 2016
1. di chi è questa casa?	de che é pa chësta ciasa?
2. uno spirito satirico	n spirit satirich
3. una gara di pattinaggio	na gara da jadiné
4. fare il pane	fá pan
5. scrivere quello che viene in mente	scrí chël che toma ite
6. scavare con la pala	ciavé cun le badí
7. ti ricordi di me?	te recordeste de me?
8. le montagne innevate	i crëps da nëi
9. il salvatore della patria	le salvadú dla patria
10. macchiare i pantaloni con il vino	tacé la braia cun vin
11. grosso come una noce	gran sciöche na nusc
12. ingoiare una medicina	dlotí na medejina
13. quelli sono funghi velenosi	chi é fonguns da tosser
14. proseguire con la lettura	jí inant cun la letöra
15. strisciare una grossa valigia sul pavimento	streflé n gran cufer ia por funz
16. sono notizie inventate	al é notizies inventades
[ITA] Erlacher 2021	[LVB] Erlacher 2021
1. Mi sto nascondendo nel mio ufficio.	I m'ascogni te mi ofize.
2. I baffi sono solo per precauzione.	I snauzeri ái ma por precauziun.
3. Ne ho sempre un paio di ricambio.	I nen á dagnora n per da baraté jö.
4. Lo zoo era pieno di gente.	Le zoo ê plëgn de jënt.
5. Qualche volta ci gioco a scacchi.	Val' iade sogunse a sciah.
6. Se vuoi ho qualcosa da mangiare.	Sce t'os ái iö val' da mangé.
7. Non dirmi che hai dei pesci nello zaino?	No me dí che t'as pësc tl rucsoch?
8. Ma allora da chi ti nascondi?	Mo da che t'ascogneste pa spo?
9. Da piccolo volevo fare il pianista.	Da pice orôi deventé n pianist.
10. Pesci rossi, gialli, verdi e blu.	Pësc cöci, ghei, vërc y bröms.
11. Puoi dirlo forte!	Chël poste propi dí dadalt!
12. disse Max per consolarlo.	â Max dit por le consolé.
13. Il direttore si calmò.	Le diretur s'â calmé jö.
14. C'è un coccodrillo! Aiuto! Aiuto!!!	Al é n cocodril! Aiüt! Aiüt!!!
15. Max si guardò intorno.	Max s'â ciaré incërch.
16. E perché no? Lo zoo è vuoto.	Porciodí pa nia? Le zoo é öt.
[ITA] Collodi 2021	[LVB] Collodi 2021
1. diranno subito i miei piccoli lettori.	dijará atira mi pici leturs.
2. No, ragazzi, avete sbagliato.	No, mituns, i s'ëis falé.
3. C'era una volta un pezzo di legno.	Al ê n iade n tlapun.
4. Io dico che siete stato voi.	Spo sarunsi bëgn sté iö!
5. Si fermò e stette in ascolto.	Al s'â archité a d'ascuté.
6. Che cosa sia questa musica?	Ci sará mo mai chësta musiga?
7. E rimase lì perplesso.	Y al ê sté n pez dailó a pisimé.
8. Allora te lo leggerò io.	Spo t'al lii mefo iö dant.
9. GRAN TEATRO DEI BURATTINI...	GRAN TEATER DI BURATINS...
10. È molto che è incominciata la commedia?	À la comedia bele metü man da n pez?
11. Vuoi comprare le mie scarpe?	Oste me cumpré jö mi cialzá?
12. Sono buone per accendere il fuoco.	Chi podess oghé da fá füch.
13. Quanto mi dai del berretto?	Tan me daste pa söla ciüria?
14. Pinocchio era sulle spine.	Pinocchio stô sön spines.
15. esitava, tentennava, pativa.	al bazilâ, al pisimâ, al patí.
16. È il nostro fratello Pinocchio!	Al é nosc fre Pinocchio!

Fig. 3: Auszüge aus den Testdaten.

Es ist wichtig zu erwähnen, dass diese Metrik nur als Corpusmetrik mit der menschlichen Einschätzung korreliert und deshalb auch nur auf diese Weise sinnvoll verwendet werden kann. Wenn man BLEU-Ergebnisse für einzelne Beispiele betrachtet, gibt es mehrere Aspekte, die kritisiert werden können – zum Beispiel kann diese Metrik nicht zwischen Synonymen oder komplett falschen Wörtern in einer Übersetzung unterscheiden. Dieses Problem wurde bereits in Papineni 2002 beschrieben. Tab. 2 zeigt, wie die verschiedenen BLEU-Werte interpretiert werden können.

BLEU	Interpretation[6]
< 10	nahezu unbrauchbar
10–19	schwierig, das Wesentliche zu verstehen
20–29	das Wesentliche ist klar, aber erhebliche grammatikalische Fehler
30−40	verständliche bis gute Übersetzungen
40−50	hochwertige Übersetzungen
50–60	sehr gute, angemessene und flüssige Übersetzungen
> 60	Qualität oft besser als menschliche Übersetzung

Tab. 2: Interpretation von BLEU-Werten.

In diesem Bericht werden die Ansätze nur mit der BLEU-Metrik evaluiert. Der Leser sollte sich jedoch darüber im Klaren sein, dass man sich nicht nur darauf konzentrieren sollte, einen höheren BLEU-Wert zu erreichen.

4. Wörterbuch-basierte Übersetzung mit *Apertium*

In diesem Kapitel stellen wir einen Wörterbuch-basierten Ansatz für die maschinelle Übersetzung zwischen dem Gadertalischen und dem Italienischen vor, der mit *Apertium* (Forcada 2011) implementiert wurde. Der Vorteil dieses Ansatzes liegt darin, dass er nicht datengetrieben, sondern wissensbasiert ist und somit die geringe Verfügbarkeit an parallelen Daten keine Rolle spielt. Als Wissensbasis kann in diesem Fall das Wörterbuch herangezogen werden.

[6] Cf. <https://cloud.google.com/translate/automl/docs/evaluate>, [31.03.2022].

4.1 Wörterbuch-basierte Übersetzung

Eine Wörterbuch-basierte Übersetzung ist im Grunde eine Wort-für-Wort-Übersetzung, die nur in einer idealen Umgebung gut funktionieren kann, die folgende Bedingungen erfüllt:

a) die Wortfolge in der Ausgangs- und Zielsprache ist identisch,
b) Quell- und Zieltextelemente stimmen eins-zu-eins überein,
c) jedes Ausgangswort hat nur eine mögliche Übersetzung (cf. Schaeffer/Carl 2014).

Nichts davon wird für das Sprachenpaar Gadertalisch–Italienisch erfüllt. Es ist deshalb klar, dass dieser Ansatz keine guten Ergebnisse liefern kann. Dennoch ist es angesichts der Ähnlichkeiten zwischen diesen Sprachen interessant zu sehen, wie gut ein solches System funktionieren kann. Dieses Experiment kann wertvolle Einblicke geben und vor allem das Verständnis für dieses Problem fördern.

4.2 *Apertium*

Apertium ist eine freie/offene Plattform für regelbasierte maschinelle Übersetzung (cf. Forcada 2011). Die Regeln, die darin definiert werden, sind typischerweise Operationen auf Gruppen von lexikalischen Einheiten (und z.B. Operationen auf grammatikalischen-Bäumen)[7]. *Apertium* ist daher besser für verwandte Sprachen geeignet. Da Gadertalisch und Italienisch zwei romanische Sprachen sind, ist das definitiv der Fall.

Obwohl dieses System auch die Definition von Übersetzungsregeln,[8] *Constraint*-Grammatiken[9] oder sogar Anapher-Auflösungsregeln[10] unterstützt, haben wir es nur zur Implementierung eines Wörterbuch-basierten maschinellen Übersetzungssystems verwendet.

Eine Wörterbuch-basierte maschinelle Übersetzung kann in *Apertium* implementiert werden, indem man die folgenden drei Komponenten definiert:

[7] Cf. <https://wiki.apertium.org/wiki/Apertium_New_Language_Pair_HOWTO>, [31.03.2022].

[8] Cf. <https://wiki.apertium.org/wiki/Transfer_rules_examples>, [31.03.2022].

[9] Cf. <https://wiki.apertium.org/wiki/Constraint_Grammar>, [31.03.2022].

[10] Cf. <https://wiki.apertium.org/wiki/Anaphora_resolution_module>, [31.03.2022].

- ein morphologisches Wörterbuch für Gadertalisch,
- ein morphologisches Wörterbuch für Italienisch,
- ein zweisprachiges Wörterbuch Gadertalisch–Italienisch.

Diese Wörterbücher sind in *Apertium* als XML-Dateien definiert. Im Folgenden werden wir ihren Zweck und ihre Struktur erläutern.

piccol/o adj	Singular	Plural
maskulin	*piccolo*	*piccoli*
feminin	*piccola*	*piccole*

Tab. 3: Beugungen des Adjektives *piccolo*.

```
<sdefs>
  <!-- Adjektiv -->
  <sdef n="adj"/>
  <!-- Verb -->
  <sdef n="v"/>
  <!-- Feminin -->
  <sdef n="f"/>
  <!-- Maskulin -->
  <sdef n="m"/>
  <!-- Praesens -->
  <sdef n="pres"/>
  <!-- Imperfekt -->
  <sdef n="imperf"/>
...
  <!-- 1. Person -->
  <sdef n="1p"/>
  <!-- 2. Person -->
  <sdef n="2p"/>
  <!-- 3. Person -->
  <sdef n="3p"/>
  <!-- Singular -->
  <sdef n="sg"/>
  <!-- Plural -->
  <sdef n="pl"/>
</sdefs>
```

Fig. 4: Definition von Symbolen.

4.2.1 Das morphologische Wörterbuch

In einem morphologischen Wörterbuch definieren wir die lexikalischen Formen der Wörter einer Sprache. Betrachten wir das italienische Adjektiv *piccolo* und seine in Tab. 3 aufgeführten Beugungen: Je nachdem, in welchem Kontext es verwendet wird oder auf welches Subjekt es sich bezieht, können wir entweder die singuläre männliche Form *piccolo*, die männliche Pluralform *piccoli*, die weibliche Form *piccola* im Singular oder die weibliche Pluralform *piccole* verwenden. Wir definieren diese Beugungen im morphologischen Wörterbuch. Um zwischen den verschiedenen Formen zu unterscheiden, definieren wir *Symbole*.

Fig. 4 zeigt, wie Symbole in einem morphologischen Wörterbuch in *Apertium* definiert werden. Zum Beispiel können wir Symbole definieren, die die Angabe des Typs eines Wortes *adj*, *v* ermöglichen oder die zeigen, ob es sich um die Singular- *sg* oder die Pluralform *pl*, um das Tempus *pres*, *imperf*, das Genus *m, f*, die Person *1p*, *2p*, *3p* usw. handelt. Diese Symbole werden dann verwendet, um die lexikalischen Formen zu beschreiben, wie in Tab. 4 dargestellt. In *Apertium* entspricht dies einem sogenannten *Paradigma*.

piccolo	↔	Singular, männliche Form des Adjektivs *piccolo*	piccolo<adj><m><sg>
piccoli	↔	Plural, männliche Form des Adjektivs *piccolo*	piccolo<adj><m><pl>
piccola	↔	Singular, weibliche Form des Adjektivs *piccolo*	piccolo<adj><f><sg>
piccole	↔	Plural, weibliche Form des Adjektivs *piccolo*	piccolo<adj><f><pl>

Tab. 4: Unterscheidung zwischen den verschiedenen Flexionsformen mit Hilfe von Symbolen.

Es gibt viele Adjektive, die wie das Adjektiv *piccolo* dekliniert werden. Zum Beispiel folgen die Adjektive *anziano*, *bello*, *curioso*, *sicuro* und viele andere demselben Muster. Es besteht deshalb keine Notwendigkeit, für jedes dieser Wörter ein eigenes Paradigma zu definieren. Wir bestimmen den Wortstamm[11] und müssen nur beschreiben, wie sich das verbleibende Suffix ändert. So ist zum Beispiel bei dem Wort *piccolo* der Stamm *piccol* allen Formen gemeinsam und das Flexionssuffix *-o, -i, -a, -e* ist der variable Teil. Dies wird in einem Paradigma definiert. Fig. 5 zeigt die Kodierung des Paradigmas für *piccolo* in *Apertium*.

[11] In *Apertium* auch *Identität* genannt.

```
<pardef n="piccol/o_adj">
  <e><p><l>o</l><r>o<s n="adj"/><s n="m"/><s n="sg"/></r></p></e>
  <e><p><l>i</l><r>o<s n="adj"/><s n="m"/><s n="pl"/></r></p></e>
  <e><p><l>a</l><r>o<s n="adj"/><s n="f"/><s n="sg"/></r></p></e>
  <e><p><l>e</l><r>o<s n="adj"/><s n="f"/><s n="pl"/></r></p></e>
</pardef>
```

Fig. 5: Das Paradigma *piccol/o adj*.

Mit den definierten Paradigmen können wir nun Lemmata, die auf dieselbe Weise flektiert werden, demselben Paradigma zuordnen. Diese Zuordnung von Basiswörtern zum zugehörigen Paradigma ist auch Teil des morphologischen Wörterbuchs. Basiswörter werden als *Lemmata* bezeichnet. Fig. 6 zeigt, wie Lemmata in *Apertium* definiert werden. Wir definieren ein Lemma, indem wir das Lemma selbst (lm="lemma"), die Identität (d.h. den konstanten Teil, der allen Formen gemeinsam ist, Identität) und das ihm zugeordnete Paradigma (<par n="paradigm_name"/>) angeben.

```
<e lm="piccolo">
  <i>piccol</i>
  <par n="piccol/o adj"/>
</e>
...
  <e lm="anziano"><i>anzian</i><par n="piccol/o adj"/></e>
  <e lm="bello"><i>bell</i><par n="piccol/o adj"/></e>
  <e lm="curioso"><i>curios</i><par n="piccol/o adj"/></e>
  <e lm="sicuro"><i>sicur</i><par n="piccol/o adj"/></e>
...
```

Fig. 6: Definition von Lemmata in Apertium.

In Fig. 6 können wir sehen, dass jedes Adjektiv in dieser Abbildung dem Paradigma *piccol/o adj* zugeordnet wird.

Die Erstellung eines morphologischen Wörterbuchs von Hand wäre mühselig. Daher wurde dieser Prozess automatisiert, so dass die Paradigmen automatisch aus einem *Flexionscorpus* (Lemma-Inflexionspaare für beide Sprachen, Gadertalisch und Italienisch) abgeleitet wurden. Fig. 7 zeigt Ausschnitte aus diesen Flexionscorpora.

...	...
acordé, acordun, v.ind.pres.1p	*accordare, accordiamo,* v-tr.ind.pres.noi
acordé, acordunse, v.cong.pres.1p	*accordare, accordiamo,* v-tr.cong.pres.noi.che
acreditamënt, acreditamënc, sost.masch.plur	*accredito, accrediti,* s.m.pl.
acreditamënt, acreditamënt, sost.masch.sing	*accredito, accredito,* s.m.sing.
acredité, acreditá, agg.masch.plur	*accreditato, accreditato,* ag.m.sing.
acredité, acreditá, v.part.pass.masch.plur	*accreditare, accreditato,* v-tr.part.pass
acredité, acreditá, v.ind.imperf.3p	*accreditare, accreditavamo,* v-tr.ind.imperf.noi
acredité, acreditá, v.ind.imperf.3s	*accreditare, accreditava,* v-tr.ind.imperf.lei/lui
acredité, acreditá, v.ind.imperf.1s	*accreditare, accreditavo,* v-tr.ind.imperf.io
...	...

Fig. 7: Beispiele aus dem ladinischen (links) und italienischen (rechts) Flexionscorpus.

Für das Gadertalische haben wir einen Datensatz[12] von 397.219 Beugungen aus Moling 2016 und für die italienische Sprache einen Datensatz von 546.616 Beugungen aus dem Internet gesammelt. Für jedes Lemma wurde die Identität extrahiert, und die Lemmata, bei denen die verbleibenden Suffixe für jeden Flexionstyp im selben Paradigma identisch waren, wurden gruppiert. Auf diese Weise haben wir für alle ladinischen und italienischen Lemmata 1.177 bzw. 605 Paradigmen abgeleitet. Tab. 5 zeigt diese Zahlen noch einmal in einer Übersicht. Die Spalte *Dateigröße* gibt die Größen der generierten XML-Dateien an.

	Beugungen	Paradigmen	Regeln	Lemmata	Dateigröße
Gadertalisch	397.219	1.177	128.434	22.087	13.6 MB
Italienisch	546.616	605	27.364	23.887	4.3 MB

Tab. 5: Statistik der morphologischen Wörterbücher für Gadertalisch und Italienisch.

4.2.2 Zweisprachiges Wörterbuch

Der Zweck des zweisprachigen Wörterbuchs ist es, die Lemmata des ladinischen morphologischen Wörterbuchs mit jenen des italienischen morphologischen Wörterbuchs zu verbinden. Wir haben das Wörterbuch Moling 2016 verwendet, um diese Verbindungen herzustellen. Fig. 8 zeigt ein Beispiel dafür, wie diese Verbindungen in *Apertium* definiert werden. In diesem Beispiel geben wir an, dass das Adjektiv *gran* im Ladinischen dem italienischen *grande* entspricht. Analoges gilt für *pice* und *piccolo*. Insgesamt haben wir 40.066 Einträge zum zweisprachigen Wörterbuch hinzugefügt.

[12] Zur Verfügung gestellt von *smallcodes* <http://smallcodes.com/index.php/en/home-eng/>, [31.03.2022].

```
...
<e>
  <p>
    <l>gran<s n="adj"/></l>
    <r>grande<s n="adj"/></r>
  </p>
</e>
<e><p><l>pice<s n="adj"/></l><r>piccolo<s n="adj"/></r></p></e>
...
```

Fig. 8: Eintragsbeispiele des zweisprachigen Wörterbuchs.

Das zweisprachige und die beiden morphologischen Wörterbücher sind die Basis für eine Wörterbuch-basierte Übersetzung. Für jedes Wort, sei es ein Verb im Infinitiv oder z.B. im Futur, ein Adjektiv im männlichen Plural oder im weiblichen Singular usw., kann man mit diesen Informationen die passende(n) Übersetzung(en) finden.

4.3 Evaluierung

Die Wörterbuch-basierte Übersetzung stößt schnell an ihre Grenzen. Meist gibt es mehr als nur eine mögliche Übersetzung. Das kann man schon daraus schließen, dass es in beiden Sprachen nur etwa 23.000 Lemmata gibt, aber 40.000 Einträge im zweisprachigen Wörterbuch. Dies zeigt, dass wir es nicht mit den oben erwähnten idealen Bedingungen zu tun haben. In Fällen, wo mehrere Möglichkeiten gefunden wurden, wurde immer der erste Vorschlag in der Liste verwendet (was natürlich keine intelligente Lösung ist).

Mehrdeutige Übersetzungen lassen sich nur dann auflösen, wenn entweder der Kontext berücksichtigt wird, oder wenn sie auf grammatikalischer Ebene auseinandergehalten werden können. Das ladinische Wort *üsc* kann zum Beispiel auf Italienisch "Türen" *(porte)* oder "ihre" *(vostri)* bedeuten – *laurs* kann "Bär" *(orso)* oder "Arbeiten" *(lavori)* bedeuten. Ersteres könnte auf grammatikalischer Ebene disambiguiert werden, letzteres unter Berücksichtigung des Kontextes. Darüber hinaus wurden keine Übertragungsregeln definiert – zum Beispiel gibt es im Ladinischen die Interrogativform für Verben, im Italienischen aber nicht (zumindest in unseren Daten). Solche Fälle könnte man mit Regeln abdecken. In Fällen, wo die Übersetzung eines Verbs gefunden wurde, die passende Beugung aber nicht, wurde die Infinitivform als Übersetzung verwendet.

[ITA] Moling 2016 (mit *Apertium*)	[LVB] Moling 2016 (mit *Apertium*)
1. di chi vincerla pa costei casa	da che é chësc vila
2. un animo satirico	n am satirich
3. una gara di pattinare	na gara da jadiné
4. stipulare pane	fá al bina
5. stilare ciò chi cadere dentro	scrí ci che vëgn de mënta
6. scavare in lo pala	ciavé cun A badí
7. ti evocare di mi	te recorc da me
8. li rupi di neve	i munts inovëi
9. lo salvatore di patria	al salvatur della patria
10. tacé la braca in vino	macé i braies cun al vin
11. alto come una noce	gros co na nusc
12. digerire una medicina	dlotí na medejina
13. chi vincerla funghi di veleno	ci é fonguns velenoso
14. camminare oltre in la lettura	proseguire cun A letöra
15. camminare un alto baule a in base	arissé na grossa cufer sul funz
16. a vincerla notizie coniare	é notizies inventé
[ITA] Erlacher 2021 (mit *Apertium*)	[LVB] Erlacher 2021 (mit *Apertium*)
1. Li m' '\ascogne mio ufficio	Me vá ascognon nel mi büro
2. Li baffi avere maggio in cautela	I snauzeri é ma por precauziun
3. Li ne ha sempre un paio di mutare a	Un á tres n per da baratada
4. Lo zoo sì saziato di razza	Le zoo era colm da jënt
5. Valle' ' volta giocare a scacco	Zacotan iade nes jüch a sciah
6. Se t' 'lei avere io valle' ' di mangiare	Sce os á cizé da mangé
7. No mi enunciare chi t"asso pesce a zaino	Ne dirmi che as \dei pesci nello sacados
8. Ma di chi t' 'celarsi \pa \spo?	Ma dailó da che te ascogne
9. Di bimbo volere tornare un pianista	Da bas orô fá al pianista
10. Pesce cuore, , gialli, , verdi e blu	pësc cöci, , crimi, , vërc y ble
11. Ciò potere infatti enunciare forte	Pos dirlo sëch
12. aveva Max detto in lo consolare	dí Max por consolarlo
13. Lo preside s' 'aveva acquietare a	Al diretur H achité
14. A vincerla un coccodrillo! ! Aiuto! ! Aiuto	C' 'é n cocodril! ! Aiüt! ! Aiüt
15. Max s' 'aveva assistere attorno	Max H cuché incëria
16. Perché \pa non? ? Lo zoo vincerla öt	Y ciodí no? ? Le zoo é vacuum
[ITA] Collodi 2021 (mit *Apertium*)	[LVB] Collodi 2021 (mit *Apertium*)
1. enuncerà subito mio bimbi lettore	dijará doré i mi basc letur
2. No, , proli, , li s' 'avete illudersi	No, , jogn, , ëis sgaré
3. A sì un volta un ceppo	C' 'era na iade n sona da lëgn
4. Poi sarunsi bene stare io	Iö diji che sëis ester i
5. A s' 'aveva arrestare a d' 'ascolto	H archité y jí de ascuté
6. Ciò chiuso ma mai costei musica	Che roba \sia chësc musiga
7. E a sì stare un pezzo là a esitare	Y romagne iló dubius
8. Poi t' 'a leggere cioè io prima	Dailó te le liará iö
9. Alto Teatro di Buratins	GRAN TEATER \DEI CASPERLI
10. HA la spasso già messo mano di un pezzo	É tröp che é incominciata A comedia
11. Volere mi comprare a mio scarpa	Os cumpré i mi strëfli
12. Chi potere calzare di stipulare falò	É bon por impié al füch
13. Così mi concedere \pa a cuffia	Tan me dai del ciüria
14. Pinocchio stava a spine	Pinocchio era sulle spines
15. a esitava, , a esitava, , a subiva	dubitâ, , dubitâ, , sofrî
16. A vincerla nostro fratello Pinocchio	é al nost fre Pinocchio

Fig. 9: Mit dem Wörterbuch-basierten Ansatz generierte Übersetzungen der Sätze aus Fig. 3.

Tab. 6 listet die BLEU-Werte auf, die mit dem statistischen Ansatz auf den Testdaten erzielt wurden. Eine mögliche Erklärung, warum die Übersetzung vom Gadertalischen ins Italienische schlechter abschneidet, ist, dass italienische Wörter im Ladinischen oft in mehrere Wörter aufgespalten werden (z.B. *lavorativa* → *de laur*). Daher funktioniert es in der Gegenrichtung auch nicht immer, wenn man versucht, jedes einzelne ladinische Wort zurück zu übersetzen (z.B. *de laur* → *di lavoro* – auch wenn es in diesem Fall eine korrekte Übersetzung wäre).

Apertium	Italienisch → Gadertalisch	Gadertalisch → Italienisch
Moling 2016	8.41	6.66
Erlacher 2021	3.62	2.64
Collodi 2017	3.77	2.70

Tab. 6: BLEU-Ergebnisse des Wörterbuch-basierten Ansatzes.

Fig. 9 gibt einen Eindruck von den generierten Übersetzungen (für die Ausgangs- und Referenztexte siehe Fig. 3). Nur wenige Wörter wurden nicht übersetzt, weil sie im morphologischen Wörterbuch nicht zu finden waren. Dies zeigt, dass die Datenbasis gut ist, und dass die größte Herausforderung hier bei der Disambiguierung der Wörter liegt. Da in solchen Fällen immer die erstvorgeschlagene Übersetzung verwendet wurde, wurden oft Übersetzungen gewählt, die nicht zum Kontext passen.

Um diese Ergebnisse zu verbessern, wäre der nächste Schritt die Definition von sogenannten Transferregeln, um z.B. die Übereinstimmung der verschiedenen Beugungen zu verbessern. Außerdem kann das entwickelte System mit allen in *Apertium* unterstützten Regeln zu einem regelbasierten Übersetzungssystem ausgebaut werden. Die jüngste Arbeit von Khanna (2021) bestätigt, dass *Apertium* nicht veraltet und für Sprachen mit geringen Ressourcen immer noch attraktiv ist. Darauf basierende Übersetzungssysteme sind bereits für mehr als 40 Sprachen veröffentlicht.

5. Phrasen-basierte, statistische maschinelle Übersetzung mit *Moses*

Das Paradigma der statistischen maschinellen Übersetzung war lange Zeit die beste Methode der maschinellen Übersetzung. Dieser Ansatz wurde 2017 mit der Einführung der *Transformer*-Architektur überholt (cf. Vaswani 2017). Dennoch bleibt die Übersetzung von Sprachen mit geringen Ressourcen ein offenes Problem, da *Transformer* große Datenmengen benötigen, um gut zu funktionieren. Der statistische Ansatz der maschinellen Übersetzung ist und bleibt daher ein Ansatz, der für ressourcenarme Sprachen evaluiert werden muss.

Es gibt mehrere Methoden der statistischen maschinellen Übersetzung. Die leistungsstärkste Methode ist die Phrasen-basierte, statistische maschinelle Übersetzung (cf. Koehn 2009, 127). In diesem Kapitel beschreiben wir, wie wir ein Phrasen-basiertes, statistisches Übersetzungssystem für Gadertalisch – Italienisch mit *Moses* trainiert haben und diskutieren die Ergebnisse.

Wir werden nur die Grundidee der Phrasen-basierten, statistischen maschinellen Übersetzung wiedergeben und kurz die Herausforderungen aufzeigen, die bei diesem Ansatz zu bewältigen sind. Für detailliertere Informationen wird der Leser auf Koehn 2009 verwiesen. Die Dokumentation für *Moses* (Koehn 2016) ist auch *online* verfügbar.[13]

5.1 Phrasen-basierte statistische maschinelle Übersetzung

Bei der Phrasen-basierten, statistischen maschinellen Übersetzung werden die Sätze im Parallelcorpus in Sequenzen von so genannten Phrasen aufgeteilt. Eine Phrase muss keine sprachliche Einheit sein, sie kann auch nur einen Teil eines Wortes oder mehrere zusammengefasste Wörter umfassen. Wie dies der Fall ist, hängt davon ab, ob die einzelnen Wörter im Parallelcorpus häufig zusammen vorkommen oder nicht. Jede Phrase wird dann einzeln übersetzt und die übersetzten Phrasen werden dann miteinander verbunden. Das Verknüpfen der übersetzten Phrasen ist jedoch nicht trivial. Für eine Phrase kann es mehrere Übersetzungskandidaten geben, und die einzelnen Phrasen müssen möglicherweise neu geordnet werden.

[13] <https://www.statmt.org/moses/>, [31.03.2022].

Das macht die Suche nach der perfekten Übersetzung teuer – der Suchraum kann schon bei kurzen Sätzen sehr groß sein. Es braucht gute Heuristiken, Statistiken und Modelle, um den Suchraum so weit wie möglich einzugrenzen, ohne die Qualität (zu sehr) zu beeinträchtigen. Ein Phrasen-basiertes statistisches Übersetzungsmodell – und statistische Übersetzungsmodelle im Allgemeinen – bestehen aus zwei Hauptkomponenten: einem *Übersetzungsmodell* und einem *Sprachmodell*. Das Übersetzungsmodell ist für die Auswahl der Phrasen eines gegebenen Satzes, die Bestimmung der wahrscheinlichsten Übersetzung und schließlich die Neuordnung der jeweiligen Phrasen verantwortlich. Das Sprachmodell hingegen bewertet, wie wahrscheinlich es ist, dass ein bestimmter Satz in einer bestimmten Sprache gültig ist. Die Konstruktion des Sprachmodells und des Übersetzungsmodells erfolgt in der Trainingsphase und die Bestimmung der Übersetzung in der Dekodierungsphase.

Im Folgenden stellen wir *Moses* kurz vor und gehen dann genauer auf die Trainingsphase und die Dekodierungsphase ein.

5.2 *Moses*

Moses ist ein statistisches maschinelles Übersetzungssystem, das es u.a. ermöglicht, ein Phrasen-basiertes, statistisches Übersetzungsmodell zu trainieren. Es müssen dabei nur die Daten bereitgestellt werden und das *Skript Experiment.perl* mit der gewünschten Konfiguration darauf ausgeführt werden. Aber im Hintergrund passiert vieles – von der Corpusvorbereitung über das Training des Sprach- und Übersetzungsmodells bis hin zur Feinabstimmung. Im Folgenden werden diese Schritte im Detail besprochen.

5.2.1 Corpusvorbereitung

Die parallelen Daten können nicht so verwendet werden, wie sie sind, sondern müssen für den Trainingsprozess aufbereitet werden. Die folgenden Schritte werden durchgeführt:

- *tokenization*: es werden Leerzeichen eingefügt, um Wörter *(tokens)* von Interpunktionssymbolen zu trennen. Zum Beispiel wird *da d'ël.* zu *da d␣ '␣ël␣ .* (drei Leerzeichen ␣ eingefügt),
- *truecasing*: jedes Wort wird in seine wahrscheinlichste Schreibweise umgewandelt. So werden beispielsweise Wörter, die nur deshalb großgeschrieben werden, weil sie das erste Wort im Satz sind, kleingeschrieben (vorausgesetzt natürlich, sie kommen im Corpus häufiger kleingeschrieben vor),

- *cleaning*: in dieser Phase werden lange Sätze (> 80 Zeichen) und leere Sätze entfernt.

Anschließend werden die einzelnen *tokens* extrahiert und in absteigender Reihenfolge der Häufigkeit durchnummeriert. Die daraus entstehende "Liste" nennen wir *Vokabular*. Fig. 10 zeigt in der linken Hälfte einen Ausschnitt aus den generierten Vokabularen (links für Gadertalisch und rechts für Italienisch).

Anstelle der *tokens* werden die ihnen zugeordneten Nummern zur Darstellung der Texte verwendet. Fig. 10 zeigt in der rechten Hälfte Beispiele für die numerische Darstellung der Texte.

5.2.2 Sprachmodell trainieren

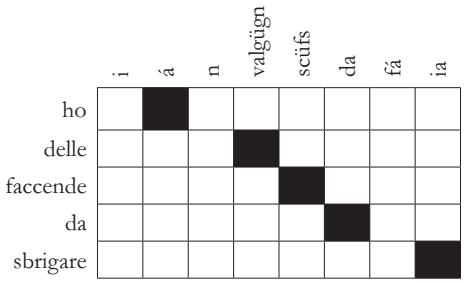
Das Sprachmodell wird verwendet, um zu prüfen, wie wahrscheinlich es ist, dass ein Satz in einer bestimmten Sprache korrekt ist. Solche Modelle basieren in der Regel auf *n*-Grammen, können aber theoretisch auch tiefe Textanalysen durchführen. Je besser ein Sprachmodell, desto hochwertiger die generierten Übersetzungen. Zum Trainieren eines Sprachmodells werden externe Tools verwendet, z.B. IRSTLM[14] oder KenLM[15]. In unseren Experimenten haben wir KenLM verwendet, das in *Moses* standardmäßig enthalten ist und es auf den Daten des Trainingscorpus für die jeweiligen Sprachen trainiert. Dieser Prozess wird in diesem Bericht nicht im Detail erklärt – der interessierte Leser wird auf Koehn verwiesen (2009, 181).

1	UNK	0	1	UNK	0		
2	da	5549	2	di	3764	bona fortüna	buona fortuna
3	de	4681	3	un	2939	99 1274	151 1039
4	n	3664	4	la	2145		
5	che	2988	5	il	2103	cütles da pom	frittelle di mele
6	se	2922	6	è	1841	5974 2 555	4678 2 758
7	le	2580	7	una	1832		
8	la	2521	8	a	1610	dënz fac ite	denti finti
9	na	2264	9	"	1332	393 1649 29	433 6828
10	al	2210	10	in	1229		
11	i	2121	11	,	1152	local da se mudé	spogliatoio
...			...			955 2 6 204	13058

Fig. 10: Ausschnitt aus den Vokabularen und numerische Darstellung der Texte (UNK = *unknown*).

[14] Cf.<https://github.com/irstlm-team/irstlm>, [31.03.2022].

[15] Cf. <https://github.com/kpu/kenlm>, [31.03.2022].



Sentence pair (27611) source length 8 target length 5 alignment score : 1.26123e-08
ho delle faccende da sbrigare
NULL ({}) i ({}) á ({1}) n ({}) valgügn ({2}) scüfs ({3}) da ({4}) fá ({}) ia ({5})

Fig. 11: Beispiel einer Wortausrichtung aus **la-it.A3.final.gz** mit Visualisierung.

5.2.3 Training des Übersetzungssystems

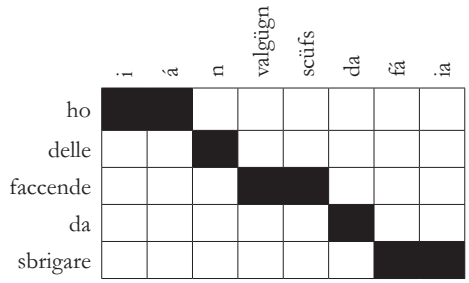
Dieser Prozess umfasst die folgenden Schritte:

- *Wortausrichtungen*: Das Ziel des *Word-Alignment*-Prozesses ist es, die Wörter des Satzes in der Ausgangssprache mit den Wörtern des Satzes in der Zielsprache zu "verbinden", d.h. es soll aus den Daten abgeleitet werden, welche Wörter Übersetzungen voneinander sind. Dies ist keine triviale Aufgabe, da die Wörter der jeweiligen Sätze nicht immer eins zu eins übereinstimmen. Oft entspricht ein Wort in der Ausgangssprache mehreren Wörtern in der Zielsprache oder umgekehrt. *Moses* verwendet das weit verbreitete Tool GIZA++,[16] um Wortausrichtungen *(word-alignments)* zu ermitteln. GIZA++ ist eine Implementierung der IBM wortbasierten Modelle, die in Koehn 2009 ausführlich beschrieben werden.

- Die Wortausrichtungen werden für beide Richtungen separat aufgestellt und dann vereinigt. Das ist deshalb notwendig, weil ein Wort in der Ausgangsprache mit mehreren Wörtern in der Zielsprache ausgerichtet werden kann und umgekehrt. *Moses* generiert die Dateien **giza.1/la-it.A3.final.gz** (für die Richtung Gadertalisch → Italienisch) sowie **giza-inverse.1/it-la.A3.final.gz** (für die Richtung Italienisch → Gadertalisch). Fig. 11 und Fig. 12 zeigen einen Ausschnitt daraus mit der zugehörigen Visualisierung.

[16] Cf. <http://www.fjoch.com/GIZA++.html>, [31.03.2022].

- Die Vereinigung dieser Wortausrichtungen erfordert die Anwendung von Heuristiken, da es für dieses Problem keine einheitliche Lösung gibt. Verschiedene Methoden können angewandt werden. In unseren Experimenten wurde die *grow-diag-final-and* Methode verwendet. Diese übernimmt (für die einzelnen Satzpaare) die Ausrichtungen der Schnittmenge der beiden Wortausrichtungen und von der Vereinigungsmenge jene, die diagonal oder benachbart zu einer Ausrichtung aus der Schnittmenge liegen. Wörter für die keine Ausrichtung diese Kriterien erfüllt, werden anschließend ergänzt. Die vereinigten Wortausrichtungen werden bei *Moses* nach **model/aligned.1.grow-diag-final-and** exportiert. Fig. 13 zeigt die Wortausrichtungen für die Sätze aus Fig. 10, so wie sie in dieser Datei abgespeichert werden.



Sentence pair (27611) source length 5 target length 8 alignment score : 1.41597e-08
i á n valgügn scüfs da fá ia
NULL ({ }) ho ({ 1 2 }) delle ({ 3 }) faccende ({ 4 5 }) da ({ 6 }) sbrigare ({ 7 8 })

Fig. 12: Beispiel einer Wortausrichtung aus **it-la.A3.final.gz** mit Visualisierung.

```
0-0 1-1
0-0 1-1 2-2
0-0 1-1 1-2
0-0 0-1 0-2 0-3
0-0 0-1 1-2 1-3 2-4 3-5 4-6 4-7
```

Fig. 13: Auszug aus **model/aligned.1.grow-diag-final-and**.

Die Wortausrichtung in Fig. 14 ist das Ergebnis der Vereinigung der Wortausrichtungen aus Fig. 11 bzw. Fig. 12 für die Sätze *ho delle faccende da sbrigare* (Italienisch) und *i á scüfs da fá ia* (Gadertalisch) und entspricht der letzten Zeile des Auszuges in Fig. 13. Generell gilt: je mehr Daten verfügbar sind und je häufiger einzelne Wörter darin vorkommen, desto besser sind auch die Wortausrichtungen.

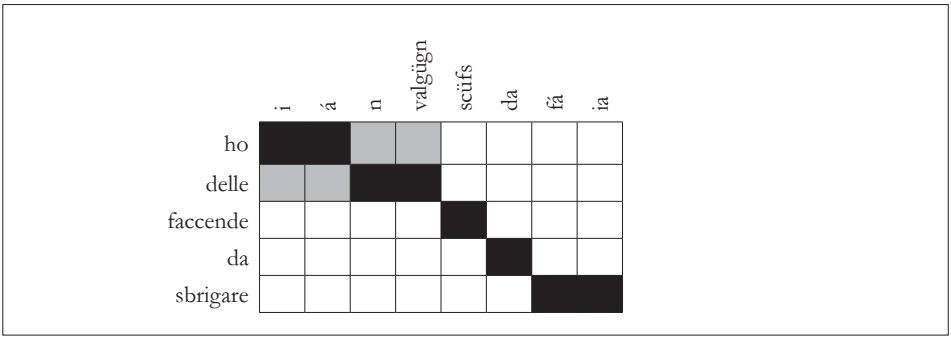


Fig. 14: Visualisierung der vereinigten Wortausrichtung.

- *Lexikalische Übersetzungstabellen*: Auf Basis der Wortausrichtungen lassen sich die wahrscheinlichsten Übersetzungen für die einzelnen Wörter ableiten, indem Kookkurrenzen analysiert werden. Fig. 15 zeigt Auszüge aus den abgeleiteten Übersetzungstabellen **model/lex.1.e2f** (in diesem Fall Gadertalisch → Italienisch) und **model/lex.1.f2e** (Italienisch → Gadertalisch). In Fig. 15a sieht man die verschiedenen Übersetzungseinträge für das Wort *valgügn*, wobei *alcuni* die wahrscheinlichste davon ist (28%). Analog in Fig. 15b für das Wort *delle*, für die das Wort *dles* mit 52% die wahrscheinlichste Übersetzung wäre.

alcuni valgügn 0.2800000	dles delle 0.5275229
qualcuno valgügn 0.1200000	di delle 0.1330275
diversi valgügn 0.0800000	les delle 0.0458716
. valgügn 0.0400000	de delle 0.0412844
qualcun valgügn 0.0400000	NULL delle 0.0366972
qualche valgügn 0.0400000	dala delle 0.0229358
qcn. valgügn 0.0400000	tles delle 0.0091743
qcn valgügn 0.0400000	n delle 0.0091743
po valgügn 0.0400000	incompatibilités delle 0.0091743
lavoretti valgügn 0.0400000	valgügn delle 0.0045872
a) *valgügn* in **lex.1.e2f.**	b) *delle* in **lex.1.f2e**

Fig. 15: Auszüge aus den lexikalischen Übersetzungstabellen.

- *Phrasenextraktion*: auch die Phrasen werden aus den Wortausrichtungen extrahiert. Betrachten wir erneut Fig. 14, so sehen wir, dass *ho* an *i á* ausgerichtet ist. Die Wörter *i*, *á* bilden also eine Einheit und werden deshalb zu einer Phrase gruppiert. Außerdem kann man ablesen, dass *ho* und *i á* ein Übersetzungspaar bilden. Analog könnte man aber auch *ho delle*, die mit *i á n valgügn* (in Fig. 14

durch den grauen Bereich hervorgehoben) ein Übersetzungspaar bildet, als Phrasen extrahieren usw. bis zu dem Punkt, wo man die gesamten Sätze als Phrase betrachtet. Wichtig ist dabei nur, dass zusammengehörende Wörter – d.h. horizontale oder vertikale Verbindungen – nicht gespalten werden (z.B. *i* und *á* als separate Phrasen). Der Algorithmus zur Phrasenextraktion ist in KOEHN (2009, 132) zu finden. Beim Schritt der Phrasenextraktion werden alle möglichen Phrasen extrahiert und nach **model/extract.1.sorted.gz** exportiert. Fig. 16 zeigt einen Ausschnitt daraus. In der ersten Spalte werden die extrahierten italienischen Phrasen, in der zweiten Spalte die dazugehörigen ladinischen Phrasen und in der dritten Spalte die Wortausrichtungen angegeben. Analog wird auch für die Gegenrichtung Gadertalisch → Italienisch eine solche Datei **model/extract.1.inv.sorted.gz** erstellt.

```
aumentabile ||| che an pó aumenté ||| 0-0 0-1 0-2 0-3
aumentabile ||| che se lascia aumenté ||| 0-0 0-1 0-2 0-3 aumentano ||| vá sö ||| 0-0 0-1
aumentare di peso ||| jí sö de pëis ||| 0-0 0-1 1-2 2-3 aumentare di ||| jí sö de ||| 0-0 0-1 1-2
aumentare il valore ||| aumenté le valur ||| 0-0 1-1 2-2 aumentare il ||| aumenté le ||| 0-0 1-1
aumentare la dose ||| aumenté la dosa ||| 0-0 1-1 2-2 aumentare la velocità ||| aumenté la velo-
zité ||| 0-0 1-1 2-2 aumentare la ||| aumenté la ||| 0-0 1-1
aumentare la ||| aumenté la ||| 0-0 1-1 aumentare ||| aumenté ||| 0-0
```

Fig. 16: Ausschnitt aus **model/extract.1.sorted.gz**.

- *Phrasentabelle*: Aus den extrahierten Phrasenpaaren wird eine Übersetzungstabelle generiert. Es wird berechnet, wie wahrscheinlich die einzelnen Phrasenpaare als Übersetzungen sind, in beiden Richtungen. Diese Werte werden in der sogenannte Phrasentabelle *(phrase-table)* **model/phrase-table.1.gz** abgespeichert. Fig. 17 zeigt einen Ausschnitt daraus, wobei nur die ersten drei Spalten angegeben werden. In der dritten Spalte werden folgende vier Werte angegeben:

  – inverse Phrasenübersetzungswahrscheinlichkeit (rechts nach links) $\phi(f|e)$
  – inverse lexikalische Gewichtung $lex(f|e)$
  – direkte Phrasenübersetzungswahrscheinlichkeit (links nach rechts) $\phi(e|f)$
  – direkte lexikalische Gewichtung $lex(e|f)$

  Die lexikalische Gewichtung gibt die Wahrscheinlichkeit an, dass die Übersetzung einer Wort-für-Wort Übersetzung entspricht. Das ist vor allem bei seltenen Phrasenpaaren relevant, die automatisch hohe Werte für $\phi(f|e)$ und $\phi(e|f)$ haben.

```
a qcn . ||| a valgügn ||| 0.143172 0.000741444 0.00130156 0.00133765
acquistato alcuni libri ||| cumpré n valgügn libri ||| 0.143172 0.00801064 0.071586 0.0991736
acquistato alcuni ||| cumpré n valgügn ||| 0.143172 0.00930267 0.071586 0.0991736
alcuni giorni in ||| n valgügn dis sön ||| 0.143172 0.0073457 0.143172 0.00170151
alcuni giorni ||| n valgügn dis ||| 0.456202 0.100391 0.456202 0.0891873
delle faccende da ||| n valgügn scüfs da ||| 0.071586 0.00102004 0.071586 2.69422e-05
delle faccende ||| n valgügn scüfs ||| 0.071586 0.00868828 0.071586 3.15633e-05
delle ||| n valgügn ||| 0.011931 0.0202726 0.000822828 4.20843e-05
diversi anni ||| n valgügn agn ||| 0.071586 0.019511 0.071586 0.0143382
diversi ||| n valgügn ||| 0.0760337 0.0402726 0.101378 0.015625
```

Fig. 17: Ausschnitt aus **model/phrase-table.1.gz**.

Für *a qcn.* ist z.B. *a valgügn* in 14% der Fälle (in den Trainingsdaten) die zugehörige Übersetzung. In der Gegenrichtung gilt das (*a qcn.* für *a valgügn*) aber nur in 1% der Fälle.

- *Umordnungsmodell*: Phrasen müssen möglicherweise neu geordnet werden. Bei der Suche nach der perfekten Übersetzung sollte eine gute und zweckmäßige Umordnung berücksichtigt werden. Wenn man jedoch jede denkbare Umordnung zulässt, explodiert der Suchraum. Außerdem ist ein Sprachmodell nicht in der Lage, die Grammatikalität eines ganzen Satzes zu beurteilen, weil es nur die einzelnen *n*-Gramme betrachtet und daher zu wenig Weitsicht und Übersicht hat. Es ist die Aufgabe des Umordnungsmodells, zu entscheiden, für welche Phrase eine Verschiebung notwendig ist und für welche nicht. Zu diesem Zweck werden für jede Phrase Umordnungsstatistiken gesammelt und zur Umordnungstabelle hinzugefügt. Die Umordnungstabelle wird dann in der Entschlüsselungsphase verwendet. Der Leser wird auf Koehn (2009, 142) verwiesen, um mehr über dieses Thema zu erfahren.

5.2.4 Feinabstimmung

In der Abstimmungsphase besteht die Aufgabe darin, die besten Parameter für das Übersetzungssystem zu finden. Dies geschieht durch wiederholte Evaluierung verschiedener Konfigurationen auf einer kleinen Menge paralleler Daten. Die "Kosten" der Übersetzung werden durch die Formel

$$p(t \mid s) = \phi(s \mid t)^{w_1} \cdot LM^{w_2}(t) \cdot D(t, s)^{w_3} \cdot W(t)^{w_4}$$

bestimmt, wobei $p(t \mid s)$ die Wahrscheinlichkeit für die Übersetzung t in der Zielsprache bei gegebener Referenz s (cf. Id. 2016, 62) ist. Diese Formel beschreibt auch das endgültige Dekodierungsproblem. Es geht darum, die

richtigen Gewichtungen (w_1, w_2, w_3, w_4) zu finden, die den Einfluss der verschiedenen Modelle regulieren: das Sprachmodell (*LM*), das Übersetzungsmodell (ϕ), das Umordnungsmodell (*D* - "bestraft" Übersetzungen mit Umordnungen) und die *word-penalty* (*W* - bestraft Übersetzungen, die sich in der Länge zu stark von der Referenzübersetzung unterscheiden). Diese Konfigurationen können für verschiedene Sprachen recht unterschiedlich sein. Bei verwandten Sprachen kann beispielsweise das Gewicht des Umordnungsmodells recht hoch sein, da wir eine ähnliche Satzstruktur erwarten, während es bei nicht verwandten Sprachen eher niedrig ist, da wir erwarten, dass viele Umordnungen erforderlich sind. Ähnlich verhält es sich mit der *word-penalty*.

Das gesamte System kann auch auf Geschwindigkeit getrimmt werden. Dies bedeutet, dass der Suchraum (der sehr groß sein kann) bei der Berechnung der Übersetzung reduziert wird. Dies kann geschehen, indem man die maximale Anzahl der Kandidaten für jede Phrase einschränkt (und nur die *n* wahrscheinlichsten nimmt) oder indem man die Größe des Stapels begrenzt, auf dem die *n* besten Teilübersetzungen gespeichert werden. Beide Einschränkungen können potentiell die Möglichkeit ausschließen, die bestmögliche Übersetzung zu finden.

5.2.5 *Experiment Management System*

Das oben erwähnte Skript **Experiment.perl**, *Experiment Management System* (EMS) genannt, ermöglicht die Konfiguration, Durchführung und Analyse eines Experimentes. Das EMS bietet auch eine Webschnittstelle für eine detaillierte Analyse der Ergebnisse.

Fig. 18 zeigt einen *Screenshot* dieses Tools mit einer vom trainierten (und abgestimmten) Modell durchgeführten Übersetzung. In diesem Beispiel wird der ladinische Satz *na mësa arjignada ca por cin' porsones* mit hoher Annäherung übersetzt: *Un tavolo pronta per cinque persone*. Wir sehen, wie die Wörter in Phrasen gruppiert (eingerahmt) sind, und können für jede einzelne Phrase ihr Vorkommen im Corpus nachschlagen. Dies hilft zu verstehen, warum eine bestimmte Übersetzung gewählt wurde und kann auch helfen, die Qualität der Trainingsdaten zu verbessern. In diesem speziellen Fall sehen wir, dass die Phrase *arjignada* zweimal mit dem Wort *pronta* und einmal mit dem Wort *arredata* im Parallelcorpus übereinstimmt. Diese Phrase wird jedoch nicht einzeln übersetzt, da die Sequenz *arjignada ca por* auch im Corpus vorkommt und daher die Übersetzung für die gesamte Phrase *(pronta per)* gewählt wird.

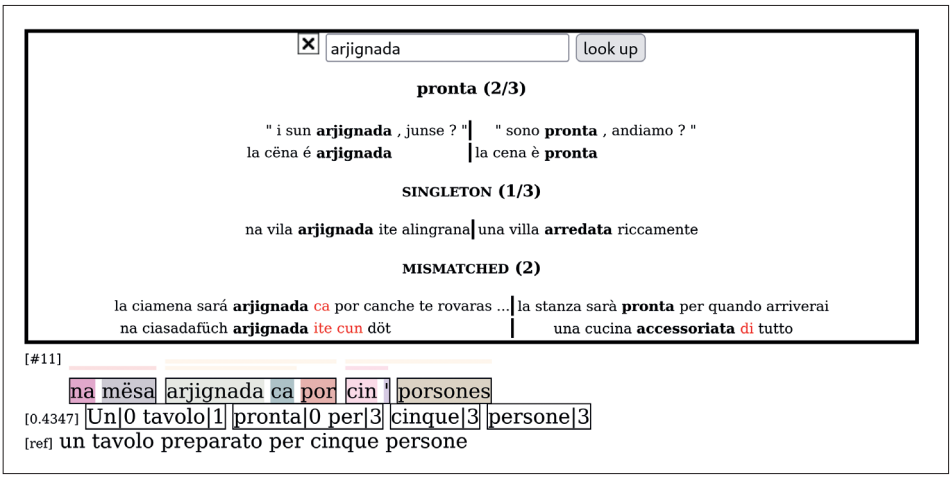


Fig. 18: *Experiment Management System* – Analysewerkzeug.

5.3 Evaluierung

Wir haben ein eigenes Übersetzungssystem für beide Übersetzungsrichtungen trainiert, abgestimmt und evaluiert. Tab. 7 listet die resultierenden BLEU-Werte auf. Laut Tab. 2 liefert dieses System *in-domain* bereits "verständliche bis gute Übersetzungen" (35 – 37 BLEU). Statistische Systeme stoßen jedoch schnell an ihre Grenzen, wenn z.B. Wörter übersetzt werden müssen, die in den Trainingsdaten nicht vorkommen. Das bestätigen auch die Ergebnisse auf den *out-of-domain* Testdaten die im Bereich 10 − 14 BLEU liegen.

Moses	Italienisch → Gadertalisch	Gadertalisch → Italienisch
Moling 2016	37.45	35.74
Collodi 2017	12.57	11.69
Erlacher 2021	10.16	13.92

Tab. 7: BLEU-Ergebnisse des statistischen Ansatzes.

Fig. 19 gibt einen Eindruck von den mit *Moses* generierten Übersetzungen (für die Ausgangs- und Referenztexte siehe Fig. 3). Im Vergleich zum Wörterbuch-basierten Ansatz gibt es hier mehr Wörter, die nicht übersetzt werden konnten. Das sind diejenigen, die in den Trainigsdaten nicht vorkommen. Aber der Gesamteindruck ist gut, und die Qualität ist auch über diese Testdaten relativ konstant. Die größte Einschränkung sind die unbekannten Wörter, was wiederum auf zu wenige Trainingsdaten zurückzuführen ist. Es fällt auf, dass die *out-of-domain*

[ITA] Moling 2016 (mit Moses)	[LVB] Moling 2016 (mit Moses)
1. di chi è questa casa ?	de che é pa chësta ciasa ?
2. un spirito satirich	n spirit satirico
3. una gara pattinare	na gara dala dlacia
4. fare pane	fá le pan
5. scrivere quello che cade dentro	scrí ci che vëgn ite
6. scavare con il pala	ciavé cun le badí
7. ti rammenti di me ?	te recordeste da me ?
8. le montagne innevate	les munts da nëi
9. il salvatore della patria	le salvadú dla patria
10. schizzato la pantaloni con vino	smardacé sö la braia cun le vin
11. grande come una noce	gros sciöche na nusc
12. ingerire una medicina	dlotí jö na medejina
13. quei è funghi velenoso	ci che é fonguns velenosi
14. proseguire con la lettura	jí inant cun la letöra
15. strascinare un grande valigia per terra	trá ia na gran cufer sön funz
16. è notizie inventades	i sun notizies inventate
[ITA] Erlacher 2021 (mit Moses)	[LVB] Erlacher 2021 (mit Moses)
1. mi ascogni nel mio ufficio .	me stá ascognon tl me ofize
2. baffi ho solo per precauzione .	i snauzeri é ma por precauziun
3. non ne ha sempre un paio di alternato .	nen á tres n por de sostituziun
4. lo zoo era pieno di gente .	I zoo ê plëgn de jënt
5. qualche volta sogunse a scacchi .	val ' iade nes jüch a sciah
6. Sì vuoi ho io qualche da mangiare .	Sce ostí á valch da mangé
7. non mi dire che hai pesce nell zaino ?	Ne me dí che astes da pësc tl sacatin
8. Mo da che puoi ascogneste pali poi ?	Pu spo da che te ascognes
9. fin da piccolo orôi diventato un pianista .	da pice orô fá i pianist
10. pesce rosse , ghei , verdi e chiaro .	pësc granëtes , , crimi , , vërc y bröm
11. Chël puoi proprio esclamare !	postes i dí dadalt
12. aveva Max detto per il rinfrancato .	â dit Max por i consolé
13. il direttore si era tranquillizzare .	i Diretur sce achité
14. È un coccodrillo ! aiuto ! aiuto ! ! !	al é n cocodril ! ! aiüt ! ! aiüt
15. Max si era guardare attorno	Max s ' â ciaré intoronn
16. perché pali nulla ? lo zoo è vuoto	Y ciodí no ? i zoo á öt
[ITA] Collodi 2021 (mit Moses)	[LVB] Collodi 2021 (mit Moses)
1. dirò subito mio piccoli leturs .	diranno atira mi pici lettori .
2. non , figli , ve lo sbagliato . "	no , mituns , ëise pa falé .
3. era un viaggio un ceppo .	al ê n iade n toch de lëgn .
4. poi sarunsi bene stato io !	iö diji nia ci sëis sté os .
5. si aveva intrattenersi a d ' ascoltare .	se fermò y stette a ascuté sö .
6. che sarà ma mai questa musica ?	ci sides chësta musiga ?
7. e era stato un pezzo lì a indugiare .	y al ciarâ ma plü dailó dubius .
8. allora puoi al lii proprio io davanti .	inlaota t ' al leggerò Iö .
9. grande teatro dei Buratins ...	gran teater DEI BURATTINI ...
10. À la commedia già cominciato a un pezzo ?	al é dër che é incominciata la comedia ?
11. vuoi comprarmi giù mio scarpe ?	oste cumpré mies cialzá ?
12. quei potrebbe servire di fare fuoco .	i sun de bones por fá füch .
13. quanto mi dai pali sulla berretto ?	tan me daste dl ciüria ?
14. Pinocchio stava sul spine .	Pinocchio ê sön les spines .
15. al bazilâ , pisimâ , patî .	esitava , tentennava , pativa .
16. È nostro fratello Pinocchio !	al é le nost fre Pinocchio !

Fig. 19: Mit dem statistischen Ansatz generierte Übersetzungen der Sätze aus Fig. 3.

Übersetzungen wesentlich schlechter sind. Um dieses System weiter zu verbessern, könnte man das gesamte Wörterbuch einschließlich des Flexionscorpus als Trainingsdaten zur Verfügung stellen oder fehlende Übersetzungen mit einer regelbasierten Übersetzung ergänzen (hybrides System).

6. Zusammenfassung

In diesem Bericht wurden zwei verschiedene Ansätze für die maschinelle Übersetzung des Gadertalischen vorgestellt: ein Wörterbuch-basierter (Kapitel 4) und ein statistischer Ansatz (Kapitel 5). Als Datenbasis wurden das Wörterbuch *Ladin Val Badia – Talian* (Moling 2016) und die Kinderbücher *Les aventöres de Pinocchio* (Collodi 2017) sowie *Dov'è finito Max? Olá é pa Max rové?* (Erlacher 2021) verwendet. Die mit den verschiedenen Ansätzen erzielten BLEU-Werte sind in Tab. 8 noch einmal zusammengefasst.

		Apertium	*Moses*
Italienisch → Gadertalisch	Moling 2016	8.41	37.45
	Collodi 2017	3.62	10.15
	Erlacher 2021	3.77	12.57
Gadertalisch → Italienisch	Moling 2016	6.66	35.74
	Collodi 2017	2.64	13.92
	Erlacher 2021	2.70	11.69

Tab. 8: Übersicht der erreichten BLEU-Werte.

Der statistische Ansatz erzielte bei weitem die besten Ergebnisse und lieferte verständliche bis gute *in-domain* Übersetzungen. Es ist jedoch wichtig zu beachten, dass die Test- und Trainingsdaten ähnlich waren und ähnliche Themenbereiche abdeckten. Die Übersetzungsqualität der *out-of-domain* Übersetzungen war deutlich schlechter, aber immer noch besser als die *in-domain* Übersetzungen des Wörterbuch-basierten Ansatzes.

Die weitere Arbeit an diesem Thema, d.h. an möglichen Ansätzen für die maschinelle Übersetzung, könnte sich in verschiedene Richtungen entwickeln, die alle Verbesserungspotenzial bieten. Wir sind der Meinung, dass der Wörterbuch-basierte Ansatz mit den verfügbaren Ressourcen den größten Spielraum für Verbesserungen bietet. Das implementierte System könnte zu einem regelbasierten Übersetzungssystem erweitert werden. *Apertium* ist dafür gut geeignet, da es immer noch eine aktive Gemeinschaft hat, die bereits Übersetzungssysteme für über 40 Sprachen veröffentlicht hat (cf. Khanna 2021).

Auch für den statistischen Ansatz gibt es noch Raum für Verbesserungen. Zum Beispiel, indem man auch die Flexionskorpora als Trainingsdaten verwendet. Hier liegt die Herausforderung darin, die Flexionskorpora der zwei Sprachen miteinander zu verbinden. Es gibt auch Veröffentlichungen, die einen hybriden Ansatz vorschlagen. In Sánchez-Cartagena/Sánchez-Martínez/Párez-Ortiz 2012 haben die Autoren beispielsweise einen Ansatz entwickelt, um *Moses* mit Phrasenpaaren anzureichern, die mit *Apertium* generiert wurden. Es wäre interessant, diesen Ansatz für die ladinische Sprache zu replizieren.

7. Bibliographie

Collodi, Carlo: *Les aventöres de Pinocchio*, San Martin de Tor 2017.

Erlacher, Max: *Dov'è finito Max? Olá é pa Max rové?*, San Martino Buon Albergo (VR) 2021.

Forcada, Mikel L. et al.: *Apertium: a free/open-source platform for rule-based machine translation*, in: "Machine translation", 25/2, 2011, 127–144.

Haddow, Barry et al.: *Survey of low-resource machine translation*, *Co*RR, abs/2109.00486, 2021.

Hedderich, Michael A. et al.: *A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios*, *Co*RR, abs/2010.12309, 2020.

Khanna, Tanmai et al.: *Recent advances in Apertium, a free/open-source rule-based machine translation platform for low-resource languages*. *Machine Translation*, in: "Machine translation", 35/4, 2021, 475–502.

Koehn, Philipp: *Statistical Machine Translation*, Cambridge 2009.

Koehn, Philipp: *MOSES*. *Statistical Machine Translation System – User Manual and Code Guide*, Edinburgh 2016.

Minority SafePack Initiative: *Stop language in Europe from becoming extinct! – The story of Goaitsen*, 2022; <http://www.minority-safepack.eu/>, [23.02.2022].

Moling, Sara et al.: *Dizionario Italiano–Ladino Val Badia/Dizionar Ladin Val Badia–Talian*, San Martin de Tor 2016.

Papineni, Kishore et al.: *Bleu: a Method for Automatic Evaluation of Machine Translation*, in: Isabelle, Pierre/Charniak, Eugene/Lin, Dekang (eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia/Pennsylvania 2002, 311–318.

Ranathunga, Surangika et al.: *Neural Machine Translation for Low-Resource Languages: A Survey*, *Co*RR, abs/2106.15115, 2021.

Sánchez-Cartagena, Victor M./Sánchez-Martínez, Fernando I./Párez-Ortiz, Juan A.: *An open-source toolkit for integrating shallow-transfer rules into phrase-based statistical machine translation*, in: Proceedings of the Third International Workshop on Free/Open-Source Rule-Based Machine Translation, Gothenburg 2012, 41–54.

Schaeffer, Moritz/Cari, Michael: *Measuring the cognitive effort of literal translation processes*, in: Germann, Ulrich et al. (eds.), Proceedings of the EACL 2014 workshop on humans and computer-assisted translation, Gothenburg 2014, 29–37.

Shen, Jiajun et al.: *The source-target domain mismatch problem in machine translation*, *Co*RR, abs/1909.13151, 2019.

Vaswani, Ashish et al.: *Attention is all you need*, *Co*RR, abs/1706.03762, 2017.

Ressumé

Te chësta relaziun vëgnel presenté y evalué döes poscibilités desvalies da fá traduziuns "a mascin" por le ladin dla Val Badia: öna, cun *Apertium*, é basada sön le dizionar; l'atra é chëra cun *Moses*, olach'an pëia ia da na statistica basada sön frases. I esperimënc desmostra che la poscibilité che se basëia sön les frases ti sciampa dant a chëra basada sön le dizionar. Mo cun les ressurses che é a desposiziun pîta la poscibilité basada sön le dizionar impó n majer potenzial ch'an pó ciamó mioré.

Abstract

In this report, two different machine translation approaches for the low-resource Ladin of the Val Badia are presented and evaluated: a dictionary-based approach using *Apertium* and a phrase-based statistical approach using *Moses*. The experiments show that the statistical approach outperforms in-domain the dictionary-based one and can provide satisfactory translations. Out-of-domain, on the other hand, both approaches deliver poor translations. Here, the dictionary-based approach offers the greater room for improvement with the available resources.