Technical Report CS2016-01

The Efficacy of Esperanto as a Pivot Language in Statistical Machine Translation

James Ballinger

Submitted to the Faculty of
The Department of Computer Science

Project Director: Dr. Janyl Jumadinova
Second Reader: Dr. Robert Roos
Third Reader: Dr. Barbara Riess

Allegheny College
2016

I hereby recognize and pledge to fulfill my responsibilities as defined in the Honor Code, and to maintain the integrity of both myself and the college community as a whole.

James Ballinger

JAMES BALLINGER. The Efficacy of Esperanto as a Pivot Language in Statistical Machine Translation. (Under the direction of Dr. Janyl Jumadinova.)

Abstract

This thesis presents the motivation, design, and evaluation of a series of experiments that assess the quality and effectiveness of Esperanto as a pivot language under resource constraints in statistical machine translation. This is done by collecting freely available parallel corpora from the web and training the open-source machine translation system Moses. Trials translating between all combinations of English, French, German, and Esperanto were run as null cases, and pivot translations conducted between French and German using either English or Esperanto as a pivot were compared using a combination of BLEU scores, METEOR scores, and human appraisals of translation intelligibility and fluency. The results indicate that Esperanto may be a favorable candidate for the pivot role. The thesis concludes with a call for the creation of more Esperanto NLP resources, and an exploration of the role of machine translation both culturally and in the greater field of translation studies in addressed throughout.

Acknowledgements

A building is the work of the ensemble; it is not solely the work of the architect—there are the clientele who specify their requirements, the workers who extract the stones from the earth, the drivers and captains who ship them around the world, the merchants who negotiate their prices, and the builders, plumbers, electricians, and landscapers, each with their own specialized skill and unique contribution. Likewise, this thesis could not have been possible without the help of many individuals. What follows is but a small sample:

- **Dr. Janyl Jumadinova**, for her constant guidance, patience, and for letting me go on Spring Break when I most needed to step away from this project. I apologize for any grey hairs you may have acquired on my behalf.

- My cohort members, Cathal Chaffee, Brandon Ginoza, Cody Kinneer, Keegan Shudy, and especially **Luke Smith** for help with debugging and for keeping comp group lighthearted, and **Katie Beisler** for her wonderful energy and courage, and for her technical expertise with R.

- My other professors **Verónica Dantán**, **Dr. Linda DeMeritt**, **Dr. Gregory Kapfhammer**, **Dr. Briana Lewis**, **Dr. Anthony Lo Bello**, **Dr. Laura Reeck**, **Dr. Barbara Riess**, **Dr. Robert Roos** , **Dr. Sinha Roy**, and **John Wenskovitch** for believing in this project. You have taught me so much.

- To **Ina Stock**, **Matthew Bolen**, **Hana Adus**, and **Maggie Dugan** for their help with identifying and cleaning the German text used in this research.

- To my friends **Daniela Cuéllar**, **Paul Willison**, **Forrest Stuckey**, and **Lexi Ashbrooke** for their constant morale, support, and *bonheur.*

Fine, ni iru!

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Translators are like actors who speak the lines as the author would if the author could speak English.

— Edith Grossman, Why Translation Matters [41]

> “The whole earth was of one language, and of one speech. And it came to pass, as they journeyed from the east, that they found a plain in the land of Šin’âr, and they dwelt there. And they said one to another, “Come, let us make bricks and burn them thoroughly.” And they had brick for stone, and slime had they for mortar. And they said, “Come, let us build us a city and a tower whose top may reach unto heaven; and let us make us a name, lest we be scattered abroad upon the face of the whole earth.”
>
> “And the Lord came down to see the city and the tower which the children of men built. And the Lord said, “Behold, the people are one and they have all one language, and this they begin to do; and now nothing will be withheld from them which they have imagined to do. Come, let Us go down, and there confound their language, that they may not understand one another’s speech.” So the Lord scattered them abroad from thence upon the face of all the earth; and they left off building the city.
>
> “Therefore is the name of it called Bâbel (that is “Confusion”) because the Lord did there confound the language of all the earth; and from thence did the Lord scatter them abroad upon the face of all the earth.” (Genesis 11:1-9)

Communication between all of the people in the world is so powerful that even the God of the Judeo-Christian tradition feared it, so he confounded the tongues of his people on Earth in the famous story of the Tower of Babel. The veracity of this story is irrelevant, but what is not is the reality it presents: the world is divided into many different languages, and they are confounded so that we cannot understand one another's speech. What happens, then, when we need to communicate?

1.1 The Importance of Translation

The primary use of language is to communicate information. However, when humans do not share a common language, information must be translated. When there is a bilingual human present who is able to effectively translate and preserve the clarity and integrity of the information, this translation problem presents few difficulties. Unfortunately, highly capable bilinguals are not always present to provide translation services, nor are there always sufficient resources to compensate them. These human translators are often pruned by the quality of their translations; poor translations do much more harm than no translation, and it is much preferable to pay more for a good translation then to receive a poor one.

The importance of an accurate translation cannot be understated. Consider the problem of translating the Qur'an into English. In a paper by Dehlia Sabry and Ibrahim Saleh, the dangers of a problematic translation are clear: "...the libels levelled against Islam are deeply rooted in the misconceptions propagated by the first Latin Qur'an translations perverted on purpose out of fear that Islam would shake the established faith of Christians" [90]. In other words, human translators purposely changed the translations of the Qur'an in order to subdue what they saw as a threat to Christianity. In doing so, they essentially fabricated the holy book of Islam and used that fabrication as a basis for the long history of tension between the Christians and Muslims of the Middle Ages, an enmity which has continued to the present day. Sabry and Saleh specifically comment on a poor translation concerning the treatment of women as laid out by the Qur'an, and thus challenge the harmful misconception that Islam degrades women because of their sex. Consider two translations of *aya* 34 of *sura* 4 from the Qur'an:

> "Men have charge of women because Allah has preferred the one above the other and because they spend their wealth on them." [15]

> "The men are supporters of the women, by what God has given one more than the other." [27]

The first translation perpetuates the idea that women are to be subservient to men; the second—the more accurate translation, according to Sabry and Saleh [90]—teaches men that women should be respected as free-thinking individuals and were not created to be treated as second-class citizens. How radically different these ideas are!

The accurate translation of literature is also extremely important. Straumanis tells the story of how a Latvian publisher became interested in acquiring the rights to a Latvian translation of a work—a novel called *High Tide* by Ethopian author Inga Âbele—that was in fact a translation from Amharic to English [99]. How many Latvian-Amharic translators are there? Probably not many, but there are many translators from Amharic to English, and many again from English to Latvian. Had someone never translated Âbele's book to English, it never would have reached its Latvian-speaking audience. Had no group of people ever bothered to explore the translation of literature and culture from one nation to another, not only would culture be lost, but economic growth would not be as efficient as it currently is because goods would not enter foreign markets so seamlessly. In a different case, Pulitzer-prize winning author Jhumpa Lahiri, who normally writes in English, decided to learn a foreign language (Italian) and write exclusively in it; without translation, the English-speaking market would never again read her works [63]. Perhaps more evident is the loss of the great classics—the works of Ovid, the writings of Moses, and the epics of

India would be lost with their worlds. Without translation, the West never would have found Yoga, and the scientific and mathematical knowledge of the Greeks and the Arabs never would have made the European scientific revolution possible.

That said, there are disastrous pitfalls for the careless translation. For example, when Schweppes tried to introduce their tonic water to the Italian market, they found that they had translated “tonic water” as “toilet water”—as one can surmise, sales did not take off, and the brand lost money in the blunder [43]. Other examples of these international branding mishaps include the American Motors Matador, a car that was introduced to Puerto Rico. Unfortunately, in Spanish, the word *matador* means *killer*, which did not inspire confidence in the Puerto Rican market [70]. Another blunder, also by a car manufacturer, was when Ford tried to introduce their car to the Belgian market with the slogan “Every car has a high-quality body.” However, when translated, the slogan read, “Every car has a high-quality corpse”—not quite the best marketing campaign [20]. All of these could have been avoided by a simple translation check.

There are more benefits to translation—The Guardian suggests that translation can help preserve dying languages, and efforts to manually translate endangered languages like Quechua are already underway [25, 109]. Perhaps even more importantly, since “Millions of Latin Americans lack health, employment or education services because they do not speak Spanish but instead one of the hundreds of indigenous languages of the region”[25], translation helps governments provide basic services for their people who may not speak the national language. Linguistic discrimination based on one’s native language is a very real experience, and research shows that preserving one’s native language leads to personal, economic, social, intellectual, and educational advantages [50]. Lahiri also notes the strange isolation of living without one’s native language: “When you live in a country where your own language is considered foreign, you can feel a continuous sense of estrangement. You speak a secret, unknown language, lacking any correspondence to the environment. An absence that creates a distance within you." [63]

Not all languages enjoy equal status. For example, the politico-linguistic situation on the African continent is strikingly unbalanced. Africa is home to more than 2,000 tongues, but only 242 of those are used in the mass media, only 63 are used in the judicial system, and only 56 are used in public administration [58, 82]. Since there is such a disparity of language in Africa, and since there is no one language that unifies Africa, other avenues, such as machine translation, must be explored. In fact, according to a report by Kelly et al., published by the Common Sense Advisory, "Translators of African languages may benefit greatly from machine translation advances if they view it as an additional productivity tool to add to their arsenal." [58] The report concludes that "Translation Will Power Africa's Future Socioeconomic Development" [58]. The phenomenon is true even on the level of individual countries, for Spanish does not unite Spain, Mandarin does not unite China, Portuguese does not unite Brazil, and Hindi certainly does not unite India. English does not even unite the United States or Canada—one only has to read the literature of immigrant communities to learn this. The 2013 American Community Survey (administered by the US Census Bureau) found that one in five Americans (20%) didn't speak English at home [115].

Monolingualism has its benefits. For examples, individuals living in monolingual communities spend less on translation and avoid potentially wasteful bilingual protections [105]; in the Saskatchewan province of Canada, there are more German-speakers than French speakers, but companies based in Saskatchewan are still required to be able to do business in both English and French [39]. There is also evidence that dominantly monolingual communities are more cost-effective than their bilingual counterparts. The rational-choice theorist Francois Grin has posited that societies choose a language that both maximizes their economic wealth and has a low cost of integrating the language into their culture [40]. A shared language also increases feelings of national pride and patriotism; for example, Bangladesh was born from Pakistan in part because of the desire for linguistic unity; the same struggle is currently taking place both in Turkey and Iran with the Kurdish people and in Spain with the Catalan people [34, 56, 71].

But societies struggling with multilingualism tend not to need translation because they have enough multilingual persons in their governments and businesses; translation, then, is only really necessary for the monolinguals.

1.2 Machine Translation

In an effort to provide translation services to the general population, many machine translation systems have appeared, most notably Google Translate. Google Translate, like many other web-based services, uses a data-driven approach to decode language; that is, Google Translate treats the input text as something that has probably been said before and combs the *parallel corpora* of the web—two texts in different languages, one of which is a translation of the other—to determine the best translation [12]. Most machine translation systems are like Google Translate—automatic translation tools, commonly found online, that use statistical methods to create mappings between the user-input texts and the most likely output translation. While Google Translate and its brethren improve every year, they are still known to make errors. And indeed, sometimes these errors are harmful, as Lavoipierre notes was the case when the Australian State Department tried using machine translation to translate tweets about ISIS into Arabic [64].

This raises an important question: when is machine translation necessary? Opinions are vastly divergent. According to an article from the Australian Institute of International Affairs, "Google translate can also be used for informal or casual communication... Anything official within a company would risk miscommunication or other related issues."[89] Others are not so strict, such as LondonTranslations, a London-based translation firm, which recommends that "If comprehension—rather than 100% accuracy—is the end goal of an exercise, machine translation can provide an effective way of getting rapid results."[7] Meanwhile, the U.S. Department of Labor advises that "It is seldom, if ever, sufficient to use machine translation without having a human who is trained in translation available to review and correct the translation to ensure that it is conveying the intended message" [75] and that machine translation is only appropriate in emergency services (with the help of a human translator), leaving one wondering why to bother using the machine at all. Another translation company, the Comprehensive Language Center, states that "Machine translation is best when used for unofficial and informational purposes or when material is perishable but too voluminous to be translated by humans ... Machine translation is also useful for technical materials that use very consistent terminology and writing styles,"[52] meaning that if the data that needs translation is time-sensitive and it is too expensive to translate by hand, machine translation is recommended. They do not recommend machine translation for material containing "nuanced or sensitive information, such as legal agreements, marketing material, business correspondence, or hand-written text." [52] As such, opinions on the proper usage of machine translation vary widely, and there is no real consensus on whether or not its usage is appropriate or inappropriate, except that it is not appropriate to process sensitive information where accuracy is the end goal.

1.2.1 Ethical Concerns

A strong criticism of machine translation is that it will do to human translators what automated tools did to factory workers in the 19^{th} and 20^{th} centuries—replace them. Even writers at MIT have pondered this [88]. However, these fears are largely exaggerated; reporters from the Huffington Post [55], The Guardian [57], and The Economist [110] cast their doubts. These fears do not have to exclude the automatic translation of spoken text, but according to Philipp Koehn, the chair of the Machine Translation research group at the Johns Hopkins University in Baltimore, MD, "Automatic spoken translation is a particular problem because you're working with two imperfect technologies tied together—speech recognition and translation." Thus it doesn't seem as though interpreters will be out of work in the near future, either [110]. Written works, which can be fed into a computer and translated much more quickly than a human could hope to do, present their own set of concerns, particularly in terms of pre-processing and formatting; copyrighted works may face additional legal barriers. In this case, I point to the current quality of translated materials, which exhibit a statistical phenomenon known as translationese. (See Section 2.1.4 for more information about translationese.) Computers simply do not have the ability to refactor beautiful prose in one language into nuanced, elegant writing in another like humans do.

Critics of machine translation point out that it encourages harmful social ideas that devalue less-common languages and perpetuates the misconception that English is the only language worth learning, since everything can be translated into it [55]. Although a moral and ethical argument supporting machine translation practices is well beyond the scope of this thesis, I point out that machine translation can be invaluable in removing human biases implicit in human translations of sensitive texts (such as in the previously-cited example of the Qur'an); additionally, the machine translation of less-common languages makes the culture of that language much more accessible to the greater world population, which in turn can encourage others to learn that language or to help inform international policies in regions where that language is spoken, for language informs culture and culture informs law.

These limitations notwithstanding, machine translation systems have shown rapid improvements from their introduction. Although machine translation was first introduced in the 1960s, the idea of a data-driven approach to machine translation, whereby the quality of the output is directly influenced by the quality of the input, was first proposed by Brown et al. in 1990 [21]. These inputs are in the form of parallel corpora, and this data-driven method to machine translation is known as *statistical machine translation.*

1.3 Pivot Translation

It is a principle of statistics that the more data one has, the more reliable statistical predictions are. This is one of the reasons that college admissions committees are more heavily weighing the high school GPA of applicants than their standardized test scores [48]—the GPA is influenced by several data points (final grades in a class), which in turn were composed of even more data points (the individual assignments in a class), but standardized test scores are incidental.[1][2] We can map this understanding to statistical machine translation: the more data the system has, the better the translation will be.

[1] Interested readers should reference the important statistical theorem called the Central Limit Theorem, which essentially states that for any normal distribution (i.e., where the graph of all the plot points is a bell curve, like IQ), as the sample size increases, so does the accuracy of one's predictions.

[2] On the other hand, SAT scores are helpful because they are *standardized*, meaning one does not see the large variation in the different GPA systems of various high schools and undergraduate institutions; this is helpful in its own way, but it is still an incidental measure. Perhaps the perfect solution to college admissions would be for students to take the SAT multiple times.

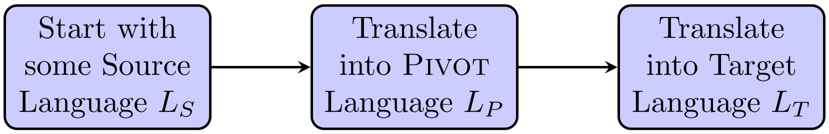

Start with some Source Language L_S
Translate into Pivot Language L_P
Translate into Target Language L_T

Figure 1.1: A basic pivot translation

Consider the problem of translating language L_a into language L_b, where there exists little to no shared literature or parallel corpora between them, such as the previously-cited example about the novel called *High Tide*, which was translated from Ethopian to English and then from English to Latvian. Using current statistical methods, the likelihood of obtaining a stellar translation is low [111]. Now consider that there exists some language with which both L_a and L_b share a large amount of parallel corpora, the pivot language L_p. Statistically speaking, it would make much more sense to first translate L_a into L_p and then from L_p into L_b; this is the pivot model that was first proposed by Wu and Wang in 2007 [111].

The trade-offs of using a pivot language have been well-discussed in the literature [84, 104, 111, 112]. A study by Dr. Michael Paul and his colleagues [84] determined that "the selection of the optimal pivot language largely depends on the SRC-PVT and PVT-TRG translation performance"; that is, if the pivot language performs well when translated from the source language and performs well when translated into the target language, it is a good pivot. Consider the problem of translating between Arabic and Basque, which is becoming more necessary as the Basque Country (a large region of Northern Spain) is experiencing an influx of Arabic-speaking immigrants [108]. In this case, it might make sense to use Spanish as a pivot language, given the geographic proximity of Spanish to both Basque and Arabic and therefore the better chance of the existence of high-quality corpora. However, in other instances, such as when translating between Berber and Vietnamese, parallel corpora are rarer, less extensive, and the choice of the pivot language less obvious.

	Source Language	Pivot Language	Target Language
Forward	Basque	Spanish	Arabic
Reverse	Arabic	Spanish	Basque

Table 1.1: Pivot translations can be bidirectional.

Another consideration when identifying a viable pivot language is the consolidation of resources. Consider the European Union, which has 24 official and working languages [28, 86]. It is not realistic to expect every delegate sitting on the EU to be fluent in all 24 of these languages. It would also be difficult—and overwhelming—to have multiple machines or humans concurrently translating documents between all of these languages [37], as often in the EU, fast conversations need to happen between speakers of different languages [31]. Unfortunately, the speech-to-text problem is nowhere near reliable enough to provide real-time translations, and thus combining that technology with machine translation may lead to disastrous miscommunications [110]. In fact, there would need to be $\binom{24}{2}$ or 276 connections and systems set up to do all of these translations. The problem can be simplified to $\binom{23}{1}$, or 23, connections through one larger system by choosing a common pivot language—usually English—which can serve as an intermediary to all of these languages.

Although English is the most commonly-used pivot language, another language called Esperanto, which was developed in 1887 as an international auxiliary language, may be better-suited to serve as a pivot language. Esperanto's grammar is highly regular and rarely ambiguous, two strengths it has over English. But while the linguistic case for Esperanto is strong, the computational case is lacking in evidence—one large concern is whether or not there exists enough data in the form of parallel corpora between Esperanto and other languages to produce quality translations. Yet there has been only one study to determine Esperanto's efficacy [29].

1.4 Goals of This Thesis

This study determines whether Esperanto performs well as a pivot by using the Moses statistical machine translation system. For more information about Moses, see [60]. This system translates into Esperanto from a source language, and then from Esperanto into a target language. Translation quality is evaluated with a BLEU score [83], a METEOR score (designed to fix the weaknesses inherent in the BLEU metric) [9, 33], and through human appraisals of the translation quality.

The goals of this thesis are as follows:

1. To demonstrate that Esperanto is more effective as a pivot language than English, even under data constraints;

2. To justify the creation of new Esperanto NLP resources, including specialized tokenization and truecasing algorithms; and

3. To explore the relationships between human judgements of Esperanto-pivot machine translated text and automatic evaluations thereof.

1.5 Outline of This Thesis

Section 2 of this proposal gives the reader background information on SMT, translation studies, and justifies Esperanto as a candidate pivot language. Section 3 gives the methods. Section 4 describes my results and discusses my findings. Section 5 concludes the document.

Chapter 2

Related Work

Every time I fire a linguist, my performance goes up.

— *Frederick Jelinek*[1]

This sections adumbrates the topics to be discussed later in this thesis and familiarizes the reader with the relevant foundational studies and information in the fields of machine translation, language model creation, Esperanto grammar, and translation studies. The information presented here is referenced in Chapter 3 and in the discussion in Chapter 4, and I have tried to provide enough of the fundamentals so that the reader not familiar with machine translation, linguistics, or translation studies, but with a background in computer science and both basic discrete mathematics and probability, can understand the outlined information.

2.1 Statistical Machine Translation

Brown, in his genesis of statistical machine translation, wrote that he only considered the problem of translating sentences. More importantly, he was not concerned with reaching the "pinnacles of the translator's art," or the artistic nuances that good translations often exhibit [21]. In the beginning, the expectation was only that the computer would render a comprehensible translation, but nothing of great quality.

[1] Although no one disputes that he said this, the exact wording is unclear. I have chosen this version as it seems to be the generally accepted version [47].

But Brown framed the problem in a way that might seem counter-intuitive: he considered that any sentence in a language is a possible translation of any sentence in another. To say that *J'adore ma chatte* and *Last night we watched the fireworks from our neighbor's back porch* are translations of each other would raise some eyebrows, but in the strictest mathematical sense we can still assign a probability to this pair. Intuitively, we assign this pair the probability of T given S; that is, $\mathrm{P}(T|S)$, or the probability of producing sentence T in the target language when given sentence S in the source language.

In machine translation, this is flipped. Thus, instead of searching for $\mathrm{P}(T|S)$, we search for $\mathrm{P}(S|T)$, or the probability of producing sentence S in the source language when given sentence T in the target language. Imagined differently, we seek the sentence from which the translator produced the translation. Mathematically, we can model this conditional probability using Bayes' rule:

$$P(S|T) = \frac{P(T|S) \times P(S)}{P(T)}, \tag{2.1}$$

where we seek to maximize the numerator $P(T|S) \times P(S)$, that is, the probability of obtaining the translated sentence given the source multiplied by the probability of obtaining the source. For this reason, searching is very important in machine translation.

To frame the problem in this way, we need to create **language models** (See Section 2.2) to represent those probability distributions. Brown calls the first factor, $\mathrm{P}(T|S)$, the *translation probability of* T *given* S, and the second factor, $\mathrm{P}(S)$, the *language model probability of* S. Essentially, the translation probability suggests words from the source language that probably produced the words in the target language translation, while the source language model established the rules for arranging these words in a sentence.

2.1.1 Pivot Statistical Machine Translation

The use of a pivot language is described by Wu and Wang in 2007, where they note its helpfulness when there exists little parallel corpora between two languages (L_S and L_T) but where there also exists a large amount of parallel corpora between both the L_S and L_P (the source and pivot languages) pair and the L_P and L_T (the pivot and target languages) pair [111].

Wu and Wang outlined their translation steps as follows: (1) segment a sentence into phrases, (2) translate each of the phrases into the target language according to the phrase translation distributions, and (3) reorder the translated phrases according to some distortion model. They formalized the phrase "translation probability" as a generative probabilistic process[2].

Wang and Wu observed that using more than one pivot language to improve translation performance is a possibility and an easy next step, saying that the ambiguities present in one language are often accounted for in another, and thus the overall translation quality for the language pair improves [111]. This combination of pivot models is called *linear interpolation*. Interpolated models can use the original source and target languages regardless of how small the parallel corpora between them are. Perhaps most interestingly, Wang and Wu found that a pivot language in a different language family produced a better translation than a pivot language in the same or in a similar language family. This finding was corroborated by Snyder and Barzilay in 2010, when it was extended to general NLP (natural language processing) tasks [94]. Overall, Wu and Wang found that the pivot model outperforms the standard model training on a small bilingual corpus, and their results consistently showed that more data correlated with better translations.

[2] Generative algorithms try to model conditional probabilities instead of trying to learn direct mappings. An algorithm that learns mappings directly is called a *discriminative algorithm* [78].

Wu and Wang revisited their pivot language approach for machine translation two years later in a 2009 paper [112]. Here, they proposed a hybrid pivot model using a rule-based machine translation (RBMT) system and a normal statistical machine translation (SMT) system and found that the hybrid model outperformed a normal pivot model, making RBMT systems especially useful when the corpora are independent of each other [112]. The finding that hybrid models are useful was also confirmed by a study seeking to translate between two dialects of German [77].

Typically, English is used as a pivot language, and quite successfully [31]. What, then, would happen if a language other than English was used? In a study that solidified the findings of Wu and Wang (that different language families were complementary towards each other), researchers Paul et al. found that using SMT techniques to translate between twelve languages revealed that the translation quality for 61 out of 110—or slightly more than half—improved when a non-English pivot was used [84]. Thus there is a precedent for the use of a non-English pivot language.

2.1.2 Pivot Phrase-based Statistical Machine Translation

In statistical machine translation (SMT), phrase-based models outperform word-based and sentence-based models [61, 104]. A study by Koehn, Och, and Marcu found that the best SMT performance can be achieved by using heuristic learning of phrase-based translations from word-based alignments and lexical weighting of phrase translations [61]. In simpler terms, by following certain rules of thumb and designating some phrases as more likely to occur than others, the system will produce a better translation. However, Koehn et al. also noted that what constitutes the best heuristic varies between both the language pairings and the size of the training corpus.

Koehn et al. carried out experiments to compare the performance of three different methods of building phrase-translation probability tables. First, they learned phrase alignments[3] from a corpus that was word-aligned using Giza++, a word-alignment tool. Second, they used a word-aligned corpus annotated with parse trees, generated by statistical syntactic parsers[4]. Third, they directly learned from phrase-level alignments of the parallel corpora [61]. The results showed that learning all phrases consistent with the word alignment was the best method, even with small phrases of at most three words.

2.1.3 Pivot Sentence-based Statistical Machine Translation

While the phrase-based technique is concerned with the translation of individual phrases and treats sentence endings as delimiters between phrases, the sentence-based technique ignores the concept of "phrases" altogether and instead sees the entire sentence as a unit. Utiyama and Isahara found that phrase-based techniques significantly outperform sentence-based techniques [104], and thus I excluded them from my model.

2.1.4 Translationese

A great many works have been published pointing out the artificiality of translated material, so much that it has been given its own moniker—*translationese* [38, 101, 102]. Translationese is best defined as "the influence of properties of the source language in a translated text in a target language [92]. The two most notorious

[3] Alignments create a correspondence between words in language A and words in language B. This type of mapping is not necessarily one-to-one or onto [22].

[4] A statistical syntactic parser works out the grammatical structures of sentences in a given language. It is statistical because it is data-driven; i.e., there is no linguist or programmer who prescribes rules for it to follow [113]. If they did, it would be called a rule-based syntactic parser.

aspects of translationese are *explicitation*[5] and *simplification*[6] [6, 16, 17]. In fact, these differences are so striking that they can be automatically classified with alarmingly high accuracy [10, 51, 62]. Resnik, a professor at the University of Maryland, notes that a key feature of translationese is the existence of phrases that are logical but sentences that are globally nonsensical [87]. This betrays the phrase-based nature of current machine translation methods: the phrases themselves make perfect sense. But when joined together, the result is an incoherent mess.

Studying translationese is an active area of research, and this phenomenon is currently being improved upon by adapting translation models to translationese [66]. This means that a good next step, after this study, would be to adapt the pivot model with Esperanto to account for the translationese effect. Lembersky, in his PhD Thesis at the University of Haifa, noted that the two major factors that influence the quality—and therefore recognizability—of translationese are the language model and the translation model. Lembersky and Twitto both noted that accounting for the effect of translationese in the training data dramatically improves the quality of the machine translation system [66, 103].

After translation, there exists the problem of correcting the grammatical errors in machine translations result from grammatical differences in the source and target languages [49]. At the very least, human translators will be needed to look at the source material and use that to correct and improve upon the machine translation.

A group of researchers at the University of Pittsburgh designed a system called *The Chinese Room*, which functioned as an interface fraught with linguistic resources that allowed users to decode and correct poor machine translations. Anecdotal evi-

[5] Given as Hypothesis 3 in [16]: "cohesive patterns in TL [target language] texts are neither TL nor SL [source language] norms oriented, but form a system of their own, possibly indicating a process of explicitation".

[6] Given as Hypothesis 1 in [16]: "cohesive patterns in TL [target language] texts text to *approximate* the norms of TL texts of the same register"—see Section 2.4 for an more in-depth discussion.

dence suggested that the insights gained from use of The Chinese Room could help researchers develop better machine translation systems; however, the study did not provide enough evidence to make any firm conclusions [1].

Finally, although the purpose of language is to communicate information, it is important to remember that language is also an art form. The machine is concerned with making sure that the nuts and bolts of the language are held together; it has no concept of how to make musical or poetic phrases, nor does it take artistic liberties. Human translators will need to continue to be artists so that the great works of literature are not lost in machine translation.

2.2 Language Models

The language model, defined as the probability distribution over a sequence of words, estimates the *a priori* probability of a sentence in the target language [65]. (This is the same as Brown's equation of conditional probability introduced in Section 2.1.) Most language models today are n-grams that model language as a Markov chain[7] of order $n - 1$; these language models were developed for the problem of speech recognition and were later adopted by the machine translation community [5, 21].

2.2.1 n-gram-based Machine Translation

Researchers from the Universitat Politècnica de Catalunya (the Polytechnic University of Catalonia) present what they call the Tuple n-gram Model [72]. According to the authors—henceforth referred to as Mariño et al.—a **tuple** is bilingual unit—a sample of "bilinguage" [30]. Hence, researchers are able to model transition proba-

[7] A Markov chain is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event. In the case of n-grams, each word only cares about the $n - 1$ words that directly proceed it.

bilities at the sentence level using n-grams of tuples, as described by the following:

$$p(T, S) \approx \prod_{k=1}^{K} p((t, s)_k | (t, s)_{k-1}, (t, s)_{k-2}, \ldots, (t, s)_{k-n+1}), \tag{2.2}$$

where t refers to the target, s to the source, and $(t, s)_k$ to the kth tuple of a given bilingual sentence pair. In plain English, this means that the probability of getting the tuple (T, S), where T and S are sentences, is approximately equal to the conditional probability of getting the bilingual tuple (t, s), where t and s are words in T and S respectively, given all of the previous bilingual tuples in those sentences. For our earlier example of the French *J'adore ma chatte* $= F$ and the new English sentence *Mary eats fish* $= E$, we can say:

$$p(F, E) \approx \prod_{k=1}^{K} p((f, e)_k | (f, e)_{k-1}, (f, e)_{k-2}, \ldots, (f, e)_{k-n+1}), \tag{2.3}$$

can be mapped (remembering that f and e are words where $f_i \in F$ and $e_i \in E$) to:

$$p(F, E) \approx p(chatte, fish) | p(ma, eats) \times p(J'adore, Mary) \tag{2.4}$$

the probability of which would be close to zero. (*J'adore ma chatte* is French for "I love my (female) cat.")

Mariño et al. extract tuples from word-to-word bilingual aligned corpora [72]. In their study, Mariño et al. found that the single best translation, as determined by BLEU scores[8], was n-gram-based; however, the majority of the best translations were from pivot-based systems. Thus, we can attribute the word-to-word n-gram method a higher level of variability than the pivot paradigm.

Then it seems the trade-off can be condensed into the following consideration:

[8] A BLEU score determines how close a machine translation is to a professional human translation; see Section 3.9.1 for more details.

Does the researcher value greater variability in translation quality, with a few very good translations and a few very bad ones, or would they rather opt for a more reliable and still high-performing system? In this case, given that the best n-gram translation did not significantly nor consistently outperform the best pivot translation, the greater consistency of the pivot model make it the more appealing choice.

2.2.2 KenLM

The KenLM was developed by Kenneth Heafield and is shipped with Moses [46]. It was designed to function as a language model but at a faster speed and through consuming less memory than traditional models; Heafield compares his system with SRILM, IRSTLM, MITLM, RandLM, BerkeleyLM, Sheffield, and TPT, and KenLM performs *substantially better than all of them.* Furthermore, the KenLM is open source, and it interfaces with both Java and C++.

KenLM presents two new data structures: the first is PROBING, which uses linear probing hash tables and is designed for speed—it is 2.4 times faster than SRILM and uses 57% of the memory. The TRIE data structure is a reverse trie with sorted records, interpolation search, bit-level packing, reverse n-gram lookup, and optional quantization, all aimed at lowering memory consumption. It is also fast: in his trials Heafield found that the TRIE simultaneously used less memory and outperformed the strictest benchmarks.

One of the reasons for this speed is that KenLM does not use a trie[9] filled with hashtables, which Heafield notes makes his implementation unique among its competition. The interpolation search is another reason for this speed—interpolation search is a formal implementation of the idea that one opens the dictionary near to end to find the word "zebra". Thus in using Heafield's KenLM, we can minimize our search

[9] A trie is essentially a tree that is optimized for storing and searching for strings. The preference for a trie over a tree is therefore obvious: in machine translation, all of the data is text-based.

time by ignoring the part of the search space that is unlikely to contain the object of our search.

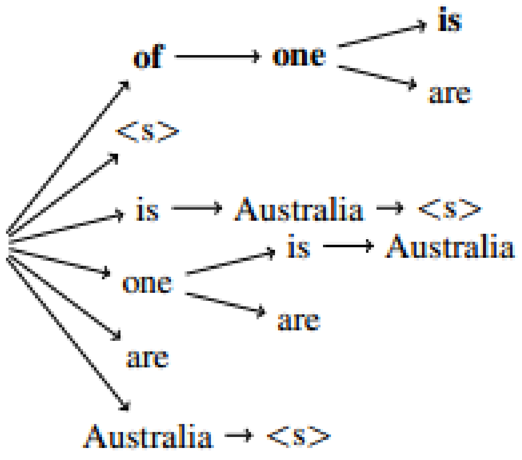


Figure 2.1: Lookup of "is one of" in a reverse trie [46].

Essentially, KenLM models n-grams by storing them in its reverse trie. Its speed comes from the quickness of its linear probling and interpolation search lookups, and its low memory is a result of its use of lossless compression. KenLM is also thread-safe. This method of modeling language allows the language model to retain large combinations of data.

2.2.3 Neural-Network Language Models

Recently, the pendulum has begun to swing in the favor of the neural network language model (NNLM). In this section, I give a cursory introduction of the neural network, and then I explore different neural network representations of language models.

2.2.3.1 Neural Networks

A neural network, more properly termed an *artificial neural network*, is a mathematical and statistical representation of a neuron in the human brain. Human neurons

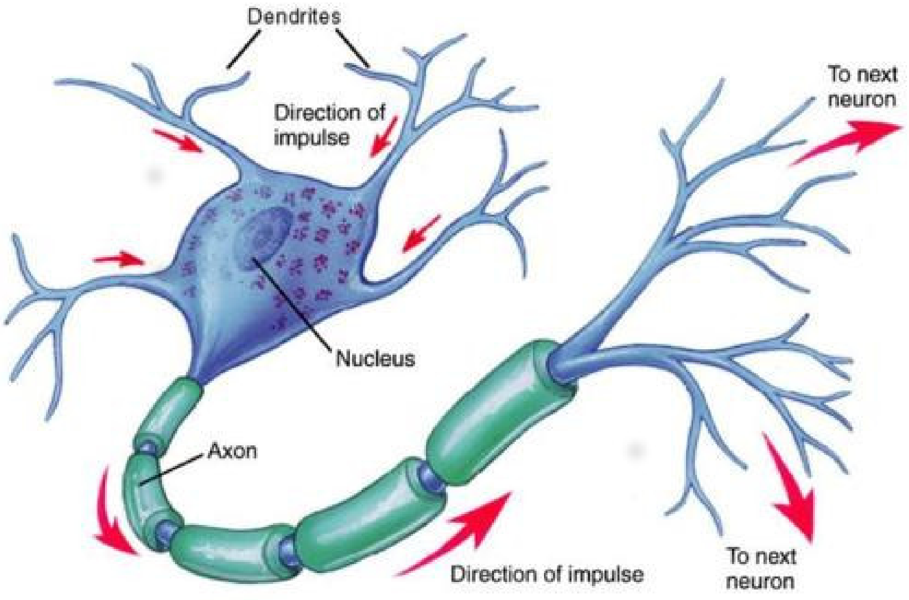


Figure 2.2: A biological neuron [69]

have input receptors (called dendrites) which aggregate electrical charges in the cell soma (the cell body); each neuron has some activation threshold (usually $-55mV$) that must be transcended before the action potential (before the neuron fires). The action potential travels through the axon and the terminal buttons to the next neuron. In the simplest case, in biological neurons a neurotransmitter (a chemical) is released at the terminal buttons and activates receptors on the dendrites of the receiving neuron; this pattern continues until a neuron's threshold is not transcended (i.e., the action potential fades out) or the action potential reaches the final neuron [11]. Neural communication is, of course, much more complex then this simplification.

Biological and psychological theories of learning mostly ascribe to what is called Hebbian learning—commonly simplified by students as "cells that fire together wire together" [24]. The idea is that the most-commonly used connections become the most powerful connections. The artificial neural network represents this in terms of *weights* [98]. Thus the output of each neuron in the artificial neural network is given some weight which determines its influence on the next neuron. In this way, artificial neural networks are very powerful and finely tunable.

We represent the input of the basic artificial neuron as follows:

$$weightedsum = \sum_{i=0}^{n} x_i w_i \tag{2.5}$$

where the *weighted sum* corresponds the the sum of all of the inputs x_i multiplied by their weights w_i. If the *weighted sum* transcends some threshold, the neuron fires.

2.2.3.2 Bengio's Neural Network Language Model

Let us briefly revisit the n-gram model. Recall from Section 2.2.1 that:

$$p(T, S) \approx \prod_{k=1}^{K} p((t, s)_k | (t, s)_{k-1}, (t, s)_{k-2}, \ldots, (t, s)_{k-n+1}), \tag{2.6}$$

which, when turned monolingual, we can rewrite as the monolingual n-gram equation:

$$P(w_t | w_1^{t-1}) = P(w_t | w_{t-n+1}^{t-1}) \tag{2.7}$$

where w_t is the t-th word and $w \in$ some sentence W and $w_i^j = (w_i, w_{i+1}, \ldots, w_{j-1}, w_j)$ [14]. Conceptually, the n-gram model is designed to take advantage of the fact that temporally closer words in a sentence or word sequence are statistically more dependent; in other words, the n-gram model constructs tables of conditional probabilities for the next word based on what Bengio calls the *context*—combinations of the last $n-1$ words. But there is an obvious problem with this approach—what happens when the system encounters a new sequence of n words that wasn't in the training corpus? The language model should do more than determine the validity of a word based on its preceding words—it should also help us generate phrases that might be similar based on parts of speech; i.e., it should help us generate other grammatical[10] sentences. For example, the sentences *Last night my sister's best friend roasted marshmallows at the firepit* and *Tuesday my brother's favorite alpaca gnawed apples underwater* have the same linguistic structure (Time Phrase + Subject + Verb + Object + Preposition), and while one is ridiculous, it is still grammatical. Indeed, the pairs {(Last night, Tuesday), (my, my), (sister's, brother's), (best, favorite), (friend, alpaca), (roasted, gnawed), (marshmallows, apples), (at the firepit, underwater)} all play similar semantic and grammatical roles, so there is no reason we cannot interchange them to create equally grammatical—if not equally nonsensical—sentences like *Tuesday my sister's best friend gnawed marshmallows underwater.* Bengio calls this problem the *curse of dimensionality* [14].

The use of neural networks to model high-dimensional discrete distributions (like natural language) to learn the joint probability of $Z_1 \ldots Z_n$, a set of random variables with possibly different natures, was already proposed by Bengio in an earlier paper and found to be decomposed as a product of conditional probabilities [13]. Bengio's approach had three tenets:

1. Associate each word in the vocabulary with a distributed word feature vector,
2. Express the joint probability function of word sequences in terms of the feature vectors of these words in the sequence, and
3. Simultaneously learn the word feature vectors and parameters of that probability function.

Essentially, we associate each word with a point in a vector space. The probability function is a product of conditional probabilities of the next word given the previous ones (much like the n-gram).

[10] For more on grammaticality, I highly recommend Steven Pinker's book *The Language Instinct*, particularly chapters 4 ("How Language Works") and 5 ("Words, Words, Words") [85].

For the general language model, let the training set be a sequence $w_1 \dots w_T$ of words such that $w_t \in V$, in which the vocabulary V is a large but finite set. The objective is to learn a good model M: $f(w_t, \dots, w_{t-n+1}) = P(w_t|w_1^{t-1})$ that gives a high out-of-sample likelihood, where the function f is a composition of two mappings ($V \rightarrow \mathbb{R}$ and from the neural network to some distribution of random variables $Z_1 \dots Z_n$).

Bengio found that this model yielded a 10-20% difference in perplexity[11] when compared with the trigram (3-gram, the most commonly used n-gram).

2.2.3.3 Neural Network Language Model for Low-Resource Languages

Unfortunately, Bengio's model assumes an abundance of data—Esperanto does not have a billion words worth of annotated corpora for this type of neural model. Gandhe et al. proposed a solution to this data problem [36]. By using a feed-forward neural network language model (ff-NNLM) and subsequently training the ff-NNLM with different techniques, they are able to subvert the curse of dimensionality and minimize the training time. In this model, the probability of the current word depends on the word history of n words, computed as an n-gram:

$$p(w_j|history) = p(w_j|w_{j-1}, w_{j-2}, \dots, w_{j-n+1}) \tag{2.8}$$

In this way, history is presented in the ff-NNLM by many sparse vectors at the input layer. Each vector is then linearly mapped to a continuous word projection by an $N \times P$ weight matrix (F)—this concatenation of continuous word vectors results in a *projection layer*. The second layer—the hidden layer—uses a hyperbolic tangent function[12] as a non-linearity. The output layer has N units (where N is the size

[11] *Perplexity* is the geometric average of $\frac{1}{P(w_t|w_{t-1}^1)}$.

[12] The hyperbolic tangent function can be used where a sigmoid function would normally be used in a neural network [81].

of the vocabulary of the model). In order to produce the posterior probabilities, Gandhe applied a softmax function to the activation of each unit to ensure that the probabilities all sum to 1 [36].

2.2.3.4 Recurrent Neural Network-Based Language Model

More recently, Mikolov proposed the most state-of-the-art NNLM—the *recurrent* neural network language model (RNNLM), the main difference in which is the representation of history [74]. The advantages of the RNNLM over the ff-NNLM are as follows:

1. History: For the ff-NNLM, history is the previous $n-1$ words. For the RNNLM, history is contained in the hidden layer of the neural network and is therefore much more comprehensive.

2. Pattern Recognition: The RNNLM can represent more advanced patterns in sequential data than the ff-NNLM, meaning that patterns relying on words that could have occurred at variable positions in the history can be much more encoded (example: consider the problem of placing the word "only" in the sentence "he told him he was sorry"[13]).

3. Training Time: Following Mikolov's thesis [74], training the RNNLM can take less time than training the ff-NNLM, despite the larger hidden layer.

The details of Mikolov's RNNLM and backpropogation through time (BTTT) algorithms can be found in [74], and an open-source implementation of his neural networks can be found at [73].

[13] Although this is a problem only in colloquial English, because in proper, written English we place the word "only" close to the word that only "only" modifies, as opposed to only placing the world "only" next to the word that "only" modifies. Another example: an earlier version of this footnote opened as "Although this is only a problem in colloquial English".

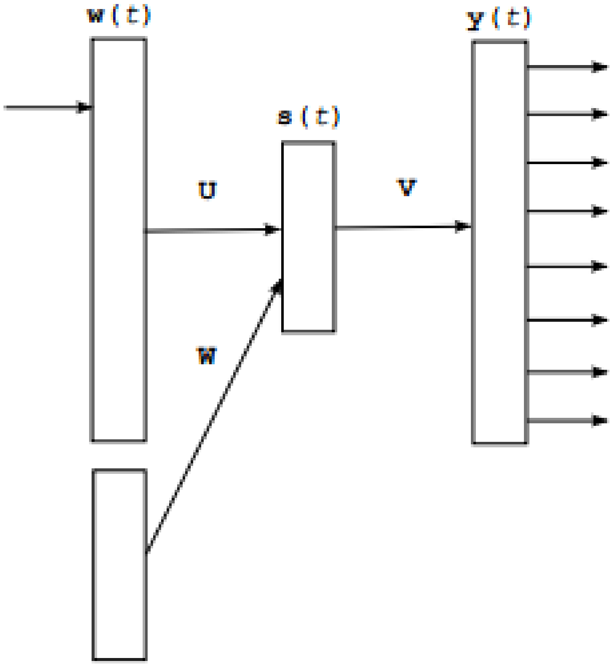


Figure 2.3: A simple recurrent neural network. [74].

2.3 The Case for Esperanto

It is important to have a basic understanding of the language being modeled by the LM. Esperanto is a constructed (planned) language invented by Leyzer Ludwik Zamenhof in 1887 [114]. This section outlines why Esperanto is a good candidate for a pivot language in an SMT system and addresses concerns about its viability.

As Wu and Wang noted, and as was later confirmed by Snyder and Barzilay, the lexical ambiguities present in one language may be missing in another language, thus improving the overall translation quality [94, 111]. This finding was particularly strong when the pivot language was in neither the source nor target language family; that is, languages in different families are complimentary. Esperanto is a constructed language: it has no organic language family.

As a constructed language, Esperanto's classification is a little ambiguous: It is very lexically Romantic, but somewhat agglutinative (a feature of languages such as

Turkish and Finnish, where suffixes and endings can be added to words and roots to create a string of characters representing a new word) and has 28 letters. Additionally, Esperanto contains some features not seen in Romance languages—for example, the accusative case (which is used in German). Words are derived by stringing together prefixes, roots, and suffixes, and compound words are formed as they are in English, with the modifier first. Regardless, the official classification of Esperanto is likely not important, for Snyder's observation that languages in different linguistic families are complimentary (in terms of machine translation) references the biological evolution of that language and less so the human classification it was given [94].

As Esperanto is a relatively new language, one might be concerned that it is lacking in resources for statistical machine translation. These concerns may include incidences of parallel corpora or whether the language lends itself well to NLP tasks. However, according to Dellert, "since the language has developed into a full replacement for natural languages in all situations, all the aspects of semantics and pragmatics that NLP (natural language processing) wants to address are present in Esperanto as much as in any natural language" [32]. Additionally, as van Cranenburgh and colleagues point out, "Although it was designed as an easy-to-learn language, with regular and transparent syntax and morphology, its semantic and pragmatic components have evolved naturally," [106] indicating that Esperanto is a viable candidate for NLP tasks.

The fact that Esperanto is now a possible language for Google Translate indicates that there is a sufficient amount of parallel corpora available to create a operable translation system between it and many other languages [23, 68]. Additionally, since nearly everything written in Esperanto was either first written in Esperanto and then translated into the author's native language or written in the author's native language and then translated into Esperanto, the phenomena of "translationese" should be greatly reduced [102]. Remember that 2.1.4 gave an overview of translationese.

2.3.1 Esperanto's Linguistic Properties

Like many languages, such as English, French, and Chinese, Esperanto most commonly uses an S-V-O (Subject-Verb-Object) word order. However, unlike most other languages, Esperanto includes an accusative case, where direct objects and their modifiers (i.e., adjectives) are marked with an -n suffix. All plural forms are marked with a -j suffix, and thus any plural direct object will end in -jn. Nouns denoting places end with the -ejo suffix, and those denoting people end with -ulo. Thus Esperanto for "coffee" is *kafo*, but Esperanto for "café", which is really a place for coffee, is *kafejo.* Another example of word building in Esperanto is the derivation of related words following the pattern we just learned. For example:

1. *sana* = healthy
2. *malsana* = (bad healthy) = sick
3. *malsanulo* = a sick person
4. *malsanulejo* = sick person place = hospital

Additionally, verbs have no conjugations—the verb for *to be*, *esti*, will never change except for to indicate a change in tense. Thus the complex rules of conjugations so many students struggle with—like the French verb for *to be*, *être*, whose present-tense conjugations are *je suis, tu es, il/elle est, nous sommes, vous êtes, ils/elles sont*—are nonexistent in Esperanto: in the present, one drops the -i suffix in *esti* to form the root *est*, and add the -as suffix; thus, everyone simply *estas.* I am *mi estas*, you are *vi estas*, he/she/it is *li/ŝi/ĝi estas*, we are *ni estas*, and they are *ili estas.* There are no exceptions to these rules.

	Question (K)	**Pointer (T)**	**Indefinite ()**	**Universal (Ĉ)**	**Negative (NEN)**
Individual **(IU)**	KIU who, which	TIU that one	IU some(one)	ĈIU every(one)	NENIU no one, none
THING **(IO)**	KIO what	TIO that thing	IO something	ĈIO everything	NENIO nothing
Kind **(IA)**	KIA what kind of	TIA that kind of	IA some kind of	ĈIA every kind of	NENIA no kind of
Place **(IE)**	KIE where	TIE there	IE somewhere	ĈIE everywhere	NENIE nowhere
Motion **(IEN)**	KIEN where to	TIEN there	IEN somewhere	ĈIEN everywhere	NENIEN nowhere
Time **(IAM)**	KIAM when	TIAM then	IAM sometime	ĈIAM always	NENIAM never
Amount **(IOM)**	KIOM how much, how many	TIOM so much, so many	IOM some	ĈIOM all	NENIOM no amount
Manner **(IEL)**	KIEL how	TIEL so	IEL somehow	ĈIEL in every way	NENIEL in no way
Reason (IAL)	KIAL why	TIAL so	IAL for some reason	ĈIAL for every reason	NENIAL for no reason
Possession **(IES)**	KIES whose	TIES that one's	IES somebody's	ĈIES everybody's	NENIES nobody's

Table 2.1: Correlatives in Esperanto

Transitive/intransitive verbs have their own markers: verbs with the *-ig* suffix are transitive; i.e., they require a direct object (marked with the -n suffix). Likewise, verbs with the *-iĝ* suffix are intransitive and never take a direct object. This is another aspect of Esperanto's expressive power. For example, the work *weak* in Esperanto is *senforta*, a compound word coming from the prefix *sen*, or without, and the root word *forta*, or strong. (All adjectives in Esperanto end with an -a.) Without strong, (i.e., without strength) means weak. And much like in English, we can modify this word to create two new words which can also be expressed in English: to weaken (*senfortigas*) and to become weak (*senfortiĝas*). The reader can probably split these words into their constituent parts: *sen* (without), *fort* (strength/strong), *ig* (transitive marker), *as* (present tense verb marker). This way of building words contributes to a highly regular system of grammar.

Esperanto expands on this word-building through its correlative system, where 50 common question words that require specific answers (what, how, when, who, etc.) and their general answers (something, in no way, then, everyone, etc.) are constructed by combining a set of predefined prefixes and suffixes [44]. Compound words, such as “birdsong” and “backpack”, also exist in Esperanto. “Birdsong” is *birdokanto*, a literal portmanteau of “bird” (birdo) and “song” (kanto), while “backpack” is much the same: *dorsosako*.

2.4 Translation Studies

Author	**Year**	**Title**	**Citation**
Roman Jakobson	1959	On Linguistic Aspects of Translation	[54]
Eugene Nida	1964	Principles of Correspondence	[79]
George Steiner	1975	The Hermeneutic Motion	[97]
Philip E. Lewis	1985	The Measure of Translation Effects	[67]
Shoshana Blum-Kulka	1986	Shifts of Cohesion and Coherence in Translation	[16]

Table 2.2: Overview of the five articles reviewed in this section

Roman Jakobson is a linguist and is primarily concerned with the words that one uses. Jakobson takes issue with the problem posed by Bertrand Russel: that one cannot understand the word ‘cheese’ unless he has a nonlinguistic association with cheese. However, according to Jakobson, since language can explain itself, we do not need to have a nonlinguistic association with a word in order to be able to define or describe or conceptualize it; neither you nor I have ever tasted ambrosia, but we still understand this word and know both other words with which is associates (“nectar”, “gods”, and perhaps “Greek”, “Mt. Olympus” and “mythological”) as well as the meaning ascribed to it: that ambrosia is the nectar of the mythological Greek gods who are said to inhabit Mt. Olympus. This ability of language to describe itself is what we call the *metalinguistic* quality of language.

However, Jakobson is careful not to ascribe exactitude in meaning to any pair of words and objects. He asks whether a word simply names the thing in question, as Adam was said to do when God brought the animals before him in Genesis, or does a word imply a meaning? Is a bachelor an unmarried man, or is there some other quality of *bachelorness* that makes the term *bachelor* a more accurate descriptor than *unmarried man*?

Jakobson identifies three kinds of translation: (1) intralingual translation, (2) interlingual translation, and (3) intersemiotic translation. Intralingual translation is what computer scientists would call bag translation—the practice of rewording a sentence.

Variation 1	Variation 2	Variation 3
The unmarried man was at the party.	The bachelor arrived at the soirée.	The single man went to the ball.
The chicken was good.	The chicken tasted good.	The chicken behaved well.
I argued for blue.	I made the argument for blue.	I argued in favor of blue.

Table 2.3: Examples of intralingual translation

Bag translations can differ in their linguistic register (very formal, not formal, average formality), vocabulary, and certainly in meaning (if the unmarried man went to the ball, did he ever arrive? Is a soirée more formal than a party; i.e., would you call a house party hosted by the undergraduates living on the corner a soirée? But is not a soirée a type a party?). Jakobson explains this by saying that any linguistic sign can be translated into further, alternative sign. However, it is important to note that our ambiguities may be our undoing—in the second example, is the speaker talking about a good-tasting chicken or a well behaved chicken? The sentence is ambiguous[14].

In interlingual translation we have the same problem—the English word for *cheese* also carries with it cultural connotations that other languages might not have; ad-

[14] Steven Pinker's *The Language Instinct* earns another recommendation for its extensive (and highly entertaining) discourse on ambiguity in language.

ditionally, other languages might have more words for *cheese*. Jakobson gives the example of Russian, where *cheese* cannot be identified with the Russian *сыр* because cheese according to the English sense is not cheese according to the Russian sense—cheese is only *сыр* if ferment is used. Thus, translation substitutes messages but not code-units. We can conclude that synonymy is not equivalence and that correspondences between foreign languages, even at the word level, do not carry exactly the same meanings.

Intersemiotic translation is not concerned with words: it is instead the interception of meaning through nonverbal means. Although not unimportant, here we ignore it because it is quite unrelated to statistical machine translation.

Eugene Nida, by contrast, is a structuralist and is more concerned about preserving the feeling elicited by the text and not so concerned with retaining the exact same meaning. However, Nida does echo Jakobson's conclusion: since no two languages are identical, there can be no absolute correspondence between languages, hence there can be no exact translations. If we think of translation formally, as a function, we would would say that translation is not a bijection.

Nida identifies three basic factors in translating: (1) the nature of the message; (2) the goal(s) or purpose(s) of the author and translator; and (3) the type of intended audience. Notice Nida's omission of the content itself—only the nature of the message contains its meaning. At times, this is very culturally appropriate—one would not, for example, translate the phrase "white as snow" into Wolof for any number of reasons, the least of all being that many Wolof speakers have never seen snow, but one might instead use the phrase "white as egret feathers" on the assumption that since egrets and Wolof speakers inhabit the same geographical region, the comparison would resonate better.

The most important classifications that Nida presents are his equivalences; a formal equivalence aims for identicality of meaning, but a dynamic equivalence aims for the complete naturalness of the expression. In other words, dynamic equivalence expresses how *we* would say it and is less concerned with the actual code-units used. For example, in Colombian Spanish one says *me cogió la corriente*[15] (I caught the current [of electricity]) instead of *I got shocked [by static electricity]*, much the same way that the English *bless you* is replaced by the Portuguese *saúde* (health) or the French *à tes/vos souhaits* (to your wishes).

Steiner is much more concerned with the philosophy of translation and the implications thereof. He identifies what he calls "the hermeneutic motion" of translation and identifies four steps: (1) Initiative Trust (*élancement*), (2) Aggression (penetration), (3) Incorporation (embodiment), and (4) Reciprocity (restitution).

The Initiative Trust begins with the seeing of the *other* and trusting that there is *something there* to be understood; in other words, trusting that the translation will be worth the energy expended upon it. This initial trust can be betrayed: nonsense rhymes, such as those from the works of Dr. Seuss, are untranslatable. But one has to suppose that if the author took the time to write it, it must mean something.

Aggression and Incorporation speak to the aspects of translation that are violent—cognition and understanding are aggressive in how they seek to understand a text, how they break it down, and how they reconstruct it in their own vision and image. In this way, Steiner likens decryption and translating: the act of "breaking" the code of the foreign language is a dissective decipherment. Incorporation is the practice of assimilating the translation into one's own culture: while translation adds to our means by giving up alternative energies and manners of expression, sometimes in our quest to possess a translation we erase its true heritage[16].

[15] In European Spanish (castellano), *coger* means something much less appropriate for television.

[16] If this sounds like a thinly-veiled metaphor for any number of forced assimilation incidents in history, it may be because it is [107]. Sometimes translation theorists, seeing how their field of study is essentially a cultural commentary, make their own comments about their own field.

Steiner declares that the final part of this translation motion is Reciprocity: we cannot take too much, and it is necessary to avoid this inflammatory "naturalization" of another culture's art. Genuine translations will seek to equalize and maintain the fidelity (defined by Steiner as an ethical faithfulness to the "negation of entropy") of the harmonics of the work being translated.

Ten years later, Lewis and Blum-Kulka wrote about the measure of translation effects and shifts of cohesion and coherence in translation. Lewis wrote that a good translation must always play tricks in the form of controlled textual disruptions, which, insofar as they are abusive, exert an unpacking and disseminating effect. According to Lewis, words exist to provide meaning, and much unlike the linguistic opinion of Jakobson, the word itself is not the most important, but rather it is the abstract image or feeling that it conveys that should be of concern. Sometimes words or sentences have to be changed to fit linguistic styles: French, for example, prefers long, flowing sentences—the longest five sentences from Marcel Proust's *In Search of Lost Time* (in French, *À la recherche du temps perdu*), in its English rendition, are 958, 599, 477, 426, and 398 words [19]—but English and German (especially German, with their rules of positioning verbs in a sentence) prefer shorter sentences or independent clauses joined by semicolons. English favors the actualization of sentences, so while a speaker of French would use the conditional *il faudrait*, an English speaker would assert the present-tense *it has to be.*

Blum-Kulka writes about something similar, but she focuses on what she calls *coherence* and *cohesion. Coherence* is defined as "a covert potential meaning relationship among parts of a text, made overt by the reader or listener through processes or interpretation", which *cohesion* is "a overt relationship holding between parts of the text, expressed by language-specific markers". Thus while coherence largely comes from the interpretation of the reader, cohesion is specific to the use of language. It might be helpful to recognize that these are different names for Nida's equivalences: coherence, with its focus on meaning, more closely resembles dynamic equivalence, while cohesion, with its concern with the language itself, more closely resembles formal equivalence. However, the central idea behind Blum-Kulka's essay is that Nida's equivalences do not always hold, because there are times the translator needs to make certain shifts due to linguistic factors.

Blum-Kulka furnishes the play *Old Times* by Pinter, the excerpt of which I've reproduced below in Table 2.4 as an example:

Excerpt from *Old Times*

	Source Language (English)	Target Language (Hebrew)
1	Kate: Dark. (pause)	kehah (dark)
2	Deeley: Fat or thin?	šmena or raza? (fat or thin)
3	Kate: Fuller than me, I think. (pause)	yoter mlea mimeni, ani xoševet (more full than me, I think)
4	Deeley: She was then?	kax haytra az? (so she was then?)
5	Kate: I think so.	kax ani xošovet (so I think)
6	Deeley: She may not be now.	yitaxen sıena kax kaet (perhaps she is not so now)

Table 2.4: *Old Times*, Pinter, 1971, Hebrew version by R. Kislev

The stage directions call for dim light so that the audience can only identify three figures on stage: Kate, curled up on a sofa, Deeley, slumped into an armchair, and Anna, perched still in the low light, looking out the window. When the lights go up, we learn that Deeley and Kate are smoking, but Anna remains in the dimness. The audience is then privy to the conversation happening on the stage. This information is deliberately unwound in stages, each line making the entire picture more apparent. By the end of these six lines, the audience learns that "dark" refers to a female person.

The shift in this example begins in the first line. In English, one can simply say "dark" without asking the question *what is dark?*, but in Hebrew (and many other languages, such as Spanish, French, German, and Portuguese) adjectives are required to agree with their nouns in gender. Thus the question of *what is dark* is partially answered—something that is feminine is dark. In Hebrew, the chosen word “kehah” can only apply to humans, and therefore the audience immediately, from the first line, knows that the subject of the discussion is a female human. The unraveling of the mystery experienced by the English-speaking audience is robbed from the Hebrew-speaking audience by the requirements of their own language. This is an example of a cohesive shift.

A coherent shift occurs in the Hebrew translations—one can see the English translation of the Hebrew in the parentheses, and in this one sees a number of “so”s. That is a coherent shift—in an attempt to preserve Kate’s uncertainty in her answers, the Hebrew translator added a number of linguistic markers that convey to the audience that Kate is not fully confident in the details of her own memory.

There are many aspects of language, linguistics, and translation studies that were not discussed in these papers but were alluded to. Modern translation theorists are concerned with the social and political implications of their translations, about the biopolitics and impacts of translation on cultural erasure, and recognize the inherent violence of translation not only as a violence against a text but potentially as a violence against a nation or culture of language speakers in the form of a neocolonial ownership of that community [95]. Modern translation theorists are increasingly interested and concerned with maintaining fidelity not only to the language within the original source material but also to the communities represented within: the article “Translating camp talk: Gay identities and cultural transfer” [45] first appeared in 1998 and has since spurred a field of LGBT translations studies. Even as recently as February 2016 (for reference, this thesis was completed in April 2016), an article was published about the censorship and manipulation of gay characters in Italian dubs of television series [91]. This erasure is the cultural violence translators talk about when they discuss translation as a type of neocolonialism and control over culture.

It is not surprising that translation is viewed as a kind of violence, a betrayal, given how closely some words or phrases pertaining to translation and betrayal represent each other. In Italian, the words for "translator" and "traitor" have a striking resemblance: *traduttore* and *traditore*. And likewise in French—*traduction* and *trahison*—mean "translation" and "betrayal", respectively. Portguese only exhibits a two-letter change (*tradução* and *traição*). German is slightly different, where a translated book (*ein übersetztes Buch*) is a prefix change away from being an injured book (*ein verletztes Buch*). What remains to be see is whether machine translation, being relatively without human influence, exhibits this kind of violence.

But if these similarities are woven into our language, then perhaps the *dépaysement* of the translated text is its own struggle. In fact, the word I just used—*dépaysement*—is an example of a word that has no English equivalent; in order for us to understand it, we must break it down and recast it in our own image. Here, a *dépaysement* is the feeling that results from being outside of one's country. More literally, it can be seen as a kind of "decountryment", which is both an accentuation and somewhat of a misrepresentation, even if it does better capture the essense of the idea in a single word.

Chapter 3

Method of Approach

A firm experimenting with an electronic brain designed to translate English into Russian fed it the words: "The spirit is willing, but the flesh is weak." The machine responded with a sentence in Russian which meant, a linguist reported, "The whisky is agreeable, but the meat has gone bad."

— *Paul Lee Tan*[1]

The vast majority of this research project was spent preparing the data and Moses for translation. This section will also address and concerns to validity and experimental flaws, as well as the concessions made from the proposal.

3.1 Moses

Moses is an open source machine translation system created by Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst with the purpose of making the field of statistical machine translations more accessible to researchers [60]. Moses is freely available on GitHub and maintains both a website and an active email-support group. It is shipped with the KenLM (see Section 2.2.2) as well as the necessary source code

[1] There are numerous tales of these kinds of mistranslations, including several where "out of sight, out of mind" was translated into Chinese and back as "invisible insanity" [93]. Another story is of the above quote being translated into Russian and back into English as "the vodka is strong, but the meat is rotten." These are all, as noted in Section 2.5, examples of intralingual translations.

for both the BLEU and METEOR evaluation metrics. We elected to use Moses because of its extensive documentation, its support infrastructure, the credentials of its creators, and the relative lack of configuration needed to run and implement it.

3.2 Configuring Moses: An Overview

I followed the instructions at the StatMT Moses Source Installation[2] page to install and build Moses and mgiza. I then attempted to create the Baseline System[3] but was ultimately unsuccessful in getting the EMS to work, so I created a bash script file with the proper command-line commands to run the translations.

My Moses configuration (or "stack", if you will) consisted of mgiza, Moses, and the default KenLM. Since I had trouble configuring the baseline system for Moses, it is worth reproducing the steps I took to configure it, since there are may small operational differences between what I did and what the EMS does. Most of these commands were taking from the Moses Baseline System website, linked to above.

The steps to setting up Moses and performing a simple direct translation (ie, source language to target language) are below:

1. Download and install Moses

2. Download mgiza

3. Compile/make mgiza and have Moses point to it

4. Identify Source and Target languages

5. Collect training data

[2] http://www.statmt.org/moses/?n=Development.GetStarted
[3] http://www.statmt.org/moses/?n=Moses.Baseline

- Constraints for the training data: the data must be sentence-aligned. Thus, the data must contain the same number of sentences and they must be in the same order. If multiple files were used to create the training corpus, join them into one file, taking care not to disrupt the order of the alignments.

6. Collect tuning data

7. Collect test data

8. Tokenize the training data: separate words and punctuation marks

9. Train the truecaser for each language

10. Truecase the training data: determine whether a letter should be upper or lower case

11. Clean the training data: eliminate sentences whose length is outside a given range, usually 1-80 tokens

12. Train the language model to produce a .arpa file

13. Binarize the .arpa file to speed up the loading times

14. Train the language model using mgiza to create phrase alignments

15. Tokenize the tuning data for each language

16. Truecase the tuning data for each language

17. Tune the translation system using the tuning data

18. Binarize, in two steps, the phrase table created from the training/tuning process

19. Copy the moses.ini output file to the same directory as the binarized phrase table

20. Redirect and rename some paths in the moses.ini file

21. Tokenize the test data

22. Truecase the test data

23. Filter the phrase table to speed up the translation

24. Translate

I have grouped these into five major categories: (1) **installation**, (2) **identification/collection**, (3) **preparation**, (4) **training/tuning**, and (5) **translation**. I will cover the first four of these in this chapter, and translation in Chapter 4, the Results.

3.3 Installation

Moses maintains an active GitHub page, but unfortunately Moses requires a couple of dependencies in order for it to make/compile and run. First, Moses is entirely dependent on Boost, a library of C++ files. Second, Moses requires the use of the word-alignment tool, and third, Moses requires a language model.

After I downloaded Moses from GitHub and followed the instructions to make it, I downloaded MGiza and Boost and compiled those; then I had to link them with Moses. Fortunately, Moses comes shipped with the KenLM, so there was no need to download and configure a separate tool.

3.4 Identification/Collection

Before discussing *how* I went about collecting my data, it is necessary to understand the constraints behind its collection. First, there is no standard sentence-aligned Esperanto Corpus comparable to Europarl, so I decided to build my own. In this corpus that I wished to create, I wanted a big data file consisting of sentence-aligned text so that each file was a translation of the other; for example, I wanted an Esperanto copy of *A Tale of Two Cities* next to an English copy next to a German copy of the same book. Before I built this I had to decide which languages I wanted to include.

My preliminary proof-of-concept tests had given me some difficulty with Chinese due to the particular challenges of working with a language that doesn't use spaces to delimit words, so for the sake of simplicity I decided to exclude similar languages (Japanese, Korean, Vietnamese, Thai, etc.) from my experiments. I wanted as linguistically diverse a set of languages as I could find that wouldn't unreasonably limit my data collection, so I decided on one Romance language, one Germanic language, and one Slavic language. I eliminated the Slavic language due to my own limitations: I don't speak any Slavic languages[4], and it became increasingly important for me to be able to identify identical sentences as I was checking the creation of the corpus. The final version of the corpus contains Esperanto, English, French, and German.

I endeavored to collect data from a variety of online sources, including the *Tekstaro de Esperanto* [35], *Project Gutenberg* [42], the University of Oslo's *Bibliotheca Polyglotta* [80], Christopher Christodoulopoulos's Bible Corpus [26], and the OPUS corpus [100]. Since I guessed that Esperanto would be my limiting factor, I started with the *Tekstaro de Esperanto*, which is a large online corpus of Esperanto texts. Unfortunately, I found many novels in the Tekstaro that were originally written in

[4] Except for a passing flirtation with Ukrainian—I decided, somewhat fatefully, to pursue Esperanto instead.

Esperanto but didn't have usable English translations; some of what I did find was either incomplete or erratically translated and I couldn't match the English with the Esperanto. When I did find a working match, I invariably found difficulties when locating a French or German version of that same text[5].

I needed to prepare the data before it could be pre-processed and finalized for Moses. This first stage of processing consists of ensuring that all of this data can be sentence-aligned. What I noticed is that many of the translations that I had found did not have a correspondence between sentences! Formally, this means that I could not achieve the coveted one-to-one and onto mapping between sentences in the source and target languages that I needed in the sentence-aligned parallel corpus. I resolved to try to align them.

My first attempt at aligning the data was based on two intuitions: (1) if two sentences are vastly different in length, then it is likely that the longer sentence was split up, and (2) sentences that are translations of each other probably have a number of words that are direct translations of each other. Thus, I devised an algorithm that I believed would word-align these sentences. I have reproduced it below, in pseudocode, in Figure 3.1.

Unfortunately, I ran into a number of difficulties in testing this algorithm on an English-Esperanto (EN-EO) set of data. First, the quality of the data that I was feeding in—especially the bilingual dictionary—was hindering my accuracy. In many cases, only the most common meanings of a word had been listed, and it seemed that even if a word had more common definitions, they weren't listed unless they translated into a different word in Esperanto. (For an example, a quack in English can be a crazy person, a charlatan, or a noise made by a duck.) Additionally, I ran into problems with variations in words: Esperanto and English verbs may not be as

[5] This was often due to copyright; one does not copyright the story, but rather the telling. Thus a French copyright on *The Great Gatsby* can still be active after the English copyright expires.

```
Data: A Bilingual Dictionary d, a Source Corpus s, a Target Corpus t, and a
      bound b
Result: Sentence aligns the corpora
Break up s and t so that there is one sentence per line;
Read both s and t into two list;
Read the bilingual dictionary in as a key-value data structure;
for every s-t pair do
    Check which sentence is longer—the language of the longer sentence
      determines how we search the dictionary;
    Identify all of the “important” words in the longer sentence: exclude
      articles;
    Check the dictionary for these important words in the other language;
    Search the other sentence for the words found in the dictionary;
    if percent matching $>= b$ then
        the sentence passes;
        if the sentence passes then
            add the pair (as a tuple) to a list of approved sentences;
        end
    else
        Read in the next sentence;
        Add this to the previous sentence;
        Try again;
        Repeat until we get a match;
        If we read in 4 sentences and still have no match, discard everything;
    end
    if both input files have been read through then
        return the list—we’re done.
    end
    Repeat with a higher bound;
end
```

Figure 3.1: The first sentence-alignment algorithm

complex as their French or German counterparts, but there is enough variation that I didn't have time to hard-code all of the irregularities in English[6]. I decided not to pursue this approach, although I think that given better resources, it would do fine.

My second approach was based on the Gale-Church alignment algorithm, the underlying assumption of which is that sentences that are similar in length are likely to be translations of each other. What the Gale-Church algorithm assumes is that every sentence in the corpus IS a sentence translation of another; it does not account for sentences that may have been combined or split. Thus I was dealing with two problems: the problem of sentence alignment that the Gale-Church alignment algorithm solves, and the problem of sentence recombination or splitting with limited data, which was regrettably beyond my capabilities. It is important to note that any error in a sentence-aligned training system is a disaster since the probability that two consecutive sentences are the same or similar in either their vocabulary or structure in a given document is close to zero. It is not hard to conclude this: if the second sentence is similar to the first sentence, then it follows that the third sentence must be similar to the second sentence, and thus also similar to the first sentence. This continues until the end, at which point you have a document all consisting of similar sentences. Unfortunately, if things worked this way then reading would quickly become quite boring. For example, if you consider the structure and the vocabulary of any two consecutive sentences in this paragraph, you will quickly arrive at the same conclusion.

Thus I could not use the majority of my unstructured data—as Esperanto is a low-resource language, this severely impacted the amount of training data that I had. I decided to exclusively use the data from the OPUS corpus for my Esperanto language pairs. In the interest of scientific comparability, I should have kept the same amount

[6] Some representative examples: the verb *to be*, the part participles of *to hold*, *to find*, and *to go*

Language Pair	Number of Lines
DE-EN	178,793
DE-EO	251,592
DE-FR	158,969
EN-EO	302,116
EN-FR	157,784
EO-FR	286,850

Table 3.1: Lines of training data

of data for my other language pairs, but I realized that I had to consider the *quality* of the corpora as much as the *quantity* of the corpora.

The OPUS corpus, concerning its Esperanto offerings, consists of the following seven corpora:

Name	Description
Books	A collection of translated literature
GlobalVoices	News stories
GNOME	GNOME localization files
KDE4	KDE4 Localization Files
OpenSubtitles2016	The opensubtitles.org 2016 corpus (including all previous data files)
Tatoeba	A database of translated sentences, often containing multiple possible translations
Ubuntu	Ubuntu Localization Files

Table 3.2: The training corpora used in the EO-* language pairs

Note the high volume of localization files: GNOME, KDE4, and Ubuntu all contain technical, localization files that do not accurately reflect written or expressive language. I could have excluded these files, but then I would have had a very low amount of training data—although that could have been an experiment in and of itself. I felt that since the *quality* of my Esperanto parallel corpora is not as pure as the *news-commentary.v8* corpus [96] that I used for the DE-EN, DE-FR, and EN-FR pairs, it would not hurt to have a higher *quantity* of corpora.

3.5 Preparation

Having chosen my languages and collected my corpora, I began to pre-process my data. Pre-processing consists of three steps: tokenizing, truecasing, and cleaning.

Tokenizing is the act of separating the text into tokens. In this case, tokenization is done to separate the words—which interest us—and the punctuation, which although still interesting, is decidedly less so. This is important because it cleans up the noisiness of natural language, which is often dogged by commas and periods trailing the end of words, invariably at the end of a clause or sentence, as well as "quotation marks", exclamation points!, parentheses(?), and colons or semicolons used to link ideas together; after tokenizing, these are all their own tokens and suddenly the word "Done!" becomes two tokens "Done" "!".

Truecasing is the act of determining the proper capitalization (or decapitalization) of a given letter in a word. Capitalization matters for a couple of reasons. In English, we capitalize proper nouns, and this is how we tell the difference between the Church (meaning the organization headed by the Pope) and the church (the one down the street that you pass on your morning commute). In German, all nouns are capitalized. If a word is capitalized in a place other than the beginning of a sentence, we know that it is a noun. Capitalization at the beginning of a sentence is in most cases a mere formality, as there is no change in meaning.

Cleaning is the act of discarding all of the potentially problematic sentences. These include sentences that are too long—Proust's longest sentence of nearly 1,000 words would not have made the cut—or sentences that are too short. Sometimes the sentences that get cleaned are stray punctuation marks, and at other times they are too complex for the computer to learn, lest it become lost in their complexity. Cleaning is essential to augmenting the quality of the training data, even if it reduces its size.

These steps, with the exception of cleaning, are repeated for the tuning data.

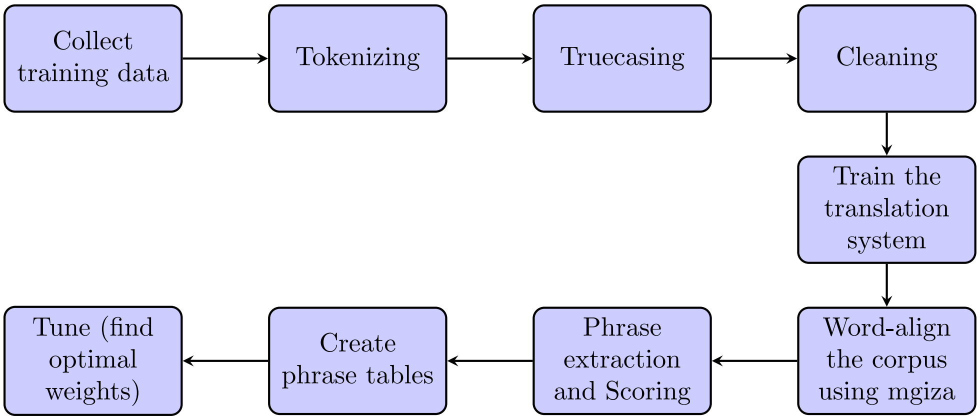

Collect training data
Tokenizing
Truecasing
Cleaning
Train the translation system
Word-align the corpus using mgiza
Phrase extraction and Scoring
Create phrase tables
Tune (find optimal weights)

Figure 3.2: Steps taken in data pre-processing

3.6 Training/Tuning

Training and tuning are where the system learns and refines the language model. During the training phase, the system extracts probabilities—review the equation at the beginning of Chapter 2 for details. The training phase is what makes all of the trouble of processing and formatting so important. To understand exactly what is happening, let us entertain, for the next few paragraphs, the ridiculous situation in which you are the computer tasked with machine translation. How would you learn how to complete this task?

First, you would have to recognize that you have no concept of language. The computer does not distinguish between English and Esperanto and Japanese, nor does it distinguish a sensible text written following the highest grammatical rules and regulations of *l'Académie Française* from the output of your neighbor's three-year-old playing on the keyboard. Having no concept of language, then, severely limits your ability to arbitrate between what is a good example of language and what isn't—you cannot complete this task. The human operator must do it for you by feeding you high-quality data.

Now that you have this data, you would likely frame the problem of translating as one of correspondence between two languages, and so, knowing so little about language, you would need to learn all of the probabilities for the correspondences between groups of words in different languages. To do this, you need a lot of input data in both languages—this is the sentence-aligned data that we have labored to create. Once you have this data, you must sift through it, calculate your probabilities, and store them somewhere, like a phrase table. This is the *training* portion.

But let us assume that you are a very cautious computer and, wary of making mistakes lest you end up in the junkyard, you want to double-check your assumptions against a piece of data that you haven't seen before. This is the tuning: you are, essentially, recalculating all of your probabilities in order to maximize your chances of successfully generating the tuning data that you are currently examining. This sounds like a very time-consuming process, much more so than the training, since you are checking, double-checking, and tweaking. You would expect to spend several hours at this task. But your performance would be much improved.

Leaving our thought experiment behind, we can frame the training/tuning problem in a different way: imagine that the computer is a student in your classroom. The more input you give it, the better it becomes at a given task. However, like most serious students, it will not perform as well as it could have on an exam without spending some time studying and getting to know the material. This independent study time is analogous to the tuning rpocess.

The input to the training step is our painstakingly prepared sentence-aligned parallel corpora. The output of the training step is the phrase table. During tuning, the computer takes in the tuning data—a smaller set similar to but not a subset of the training data—and carefully reevaluates its phrase table of probabilities to maximize the chances of producing the language that exists in the tuning data.

3.7 Modifications for the Pivot Model

Refer to Figure 3.3, which I've reproduced below from Chapter 1, for the general structure of the pivot model.

Start with some Source Language L_S
Translate into Pivot Language L_P
Translate into Target Language L_T

Figure 3.3: A basic pivot translation

The important difference between the direct and the pivot translation is that in the pivot translation, we are actually translating twice, and with a third language. Thus we need the resources for our first and second languages to be comparable for our third. The steps for the pivot models, then, are as follows:

1. Download and install Moses

2. Download mgiza

3. Compile/make mgiza and have Moses point to it

4. Identify Source, Pivot, and Target languages

5. Collect training data for all three languages

    - Constraints for the training data: the data must be sentence-aligned. Thus, the data must contain the same number of sentences and they must be in the same order. If multiple files were used to create the training corpus, join them into one file, taking care not to disrupt the order of the alignments.

6. Collect tuning data for all three languages
7. Collect test data for all three languages
8. Tokenize the training data for Source and Pivot
9. Train the truecaser for Source and Pivot
10. Truecase the training data for Source and Pivot
11. Clean the training data for Source and Pivot
12. Train the language model on Source and Pivot to produce a .arpa file
13. Binarize the .arpa file to speed up the loading times
14. Train the language model using mgiza to create phrase alignments
15. Tokenize the tuning data for Source and Pivot
16. Truecase the tuning data for Source and Pivot
17. Tune the translation system using the tuning data for Source and Pivot
18. Binarize, in two steps, the phrase table created from the training/tuning process
19. Copy the moses.ini output file to the same directory as the binarized phrase table
20. Redirect and rename some paths in the moses.ini file
21. Tokenize the test data for Source

22. Truecase the test data for Source

23. Filter the phrase table

24. Translate from Source to Pivot

25. Go through the file and make sure no characters were corrupted, à la different character encodings

26. Tokenize the training data for the Target

27. Train the truecaser for the Target

28. Truecase the training data for the target

29. Clean the training data for the target

30. Train the language model on Pivot and Target to produce a new .arpa file

31. Binarize this new .arpa file to speed up the loading times

32. Train the language model using mgiza to create phrase alignments

33. Tokenize the tuning data for Target

34. Truecase the tuning data for Target

35. Tune the translating system using the tuning data for Pivot (which was prepared in an earlier step) and Target

36. Binarize, in two steps, the phrase table created from the training/tuning process

37. Redirect and rename some paths in the moses.ini file

38. Filter the phrase table using the output from the most recent translation (which is already tokenized and truecased)

39. Translate from Pivot to Target

40. Clean the Target text: detokenize, replace any character encoding errors

Thus we see that the pivot model takes substantially more (nearly double) steps and resources.

3.8 Translations Obtained and Hypotheses

Determining the efficacy of Esperanto as a pivot languages requires me to ask a few additional questions. For example, what am I comparing Esperanto to?

	Translation	**Label**
1	EO → DE	Null Direct
2	EO → EN	
3	EO → FR	
4	DE → EN	
5	DE → EO	
6	DE → FR	
7	EN → DE	
8	EN → EO	
9	EN → FR	
10	FR → DE	
11	FR → EN	
12	FR → EO	
13	DE → EN → FR	Null Pivot
14	FR → EN → DE	
15	DE → EO → FR	Experimental Pivot
16	FR → EO → DE	
17	FR → EO → EN	Human Experimental
18	DE → EO → EN	

Table 3.3: List of Translations Run

I defined the *main experimental* case as the true test of my hypothesis of whether Esperanto outperforms English as a pivot language (the *null pivot*). The *human experimental* case tests the hypothesis of Esperanto's efficacy as a pivot language against a *direct null* translation, which do not use pivot languages.

I decided to run two main experimental cases against two null-experimental cases; if I am to measure the performance of Esperanto as a pivot language, then I need to compare it to the most commonly-used pivot language—English. I needed to compare a $L_1 \rightarrow EO \rightarrow L_2$ translation against a null $L_1 \rightarrow EN \rightarrow L_2$. This is the comparison for my experimental case.

But how do I know that my pivot translations are preforming any better than a direct translation? I needed to verify this too, so I needed to run a null $L_1 \rightarrow L_2$ translation. Thus I had a true null to compare with both my null experiment and my real experiment.

Order effects merit another consideration. Perhaps the translation $L_2 \rightarrow EO$ is better than $L_1 \rightarrow EO$. This would affect our results, so I had to include the reverse directions of all of these translations too, namely (1) $L_2 \rightarrow EO \rightarrow L_1$, (2) $L_2 \rightarrow EN \rightarrow L_1$, and (3) $L_2 \rightarrow L_1$. Thus I had comparisons for all of my tests that accounted for order effects.

Once I decided that I wanted human appraisals of the translations, I knew that I needed to have cases that translated into English. And then it occurred to me that I should be able to compare those cases with an Esperanto case. The "Human Experimental" cases were compared with their respective null direct cases; I did not anticipate having enough participants able to assess the quality of a translation into Esperanto, so I did not create an Esperanto equivalent for those cases.

3.9 Evaluation Metrics

I used three evaluation metrics: the BLEU score, a METEOR score, and human appraisals. Since there is no generally accepted machine translation benchmark (although BLEU comes pretty close to this), I felt that it was best to use a variety of different metrics in order to gain a more comprehensive understanding of these translations. There is a fourth metric, NIST, that is very similar to BLEU; since METEOR is designed to repair faults in both BLEU and NIST and because BLEU is generally more accepted, I did not think that NIST warranted the time and resources required for its inclusion.

3.9.1 The BLEU Metric

The details of BLEU[7] are given by Papineni et al. in [83], but I will give a brief overview here. Papineni reports that, before BLEU, human evaluations of machine translations assessed many aspects of language, including the *adequacy*, *fidelity*, and *fluency* of the translation. Many of the more comprehensive human evaluations could take weeks or months to finish—a problem for developers and researchers who needed quick feedback on the effectiveness of their systems.

The central idea behind the BLEU score is that "**the closer a machine translation is to a professional human translation, the better it is**". BLEU was fashioned after the *word error rate* metric used in the speech recognition community. The cornerstone of BLEU is therefore what Papineni calls the *precision measure*. To calculate precision, count the number of candidate translation words (called unigrams) and then divide by the total number of words in the candidate translation.

$$Precision = \frac{numOfCandidateTranslationWords}{totalNumOfWordsInCandidateTranslation} \tag{3.1}$$

Unfortunately, since candidate translations are often filled with a large quantity of reasonable but improbable words (this is, again, the problem of intralingual translation), precision was often highly ranked by the computer but this sentiment was not reflected by human evaluators. Thus the precision score was modified and adapted

[7] BLEU: BiLingual Evaluation Understudy; *bleu* is also the French word for *blue*, and BLEU is pronounced as such, which explains why part of the title in Papineni's paper is actually in blue. Regrettably, the paper itself is not written in French.

to n-grams.

The actual BLEU score, then, is modeled as follows:

Let BP be the *brevity penalty*, let c be the total length of the candidate translation corpus, and let r be the test corpus's effective reference length. We determine the BP as:

$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-\frac{r}{c})} & if\ c \leq r \end{cases} \tag{3.2}$$

The entire BLEU score, then, is calculated as:

$$BLEU = BP^{\left(\sum_{n=1}^{N} w_n \times \log{(p_n)}\right)} \tag{3.3}$$

Where p_n is the geometric average of the modified n-gram precisions up to length N, and the positive weights w_p sum to 1.

The calculation itself is much easier to understand in its log form:

$$\log{(\text{BLEU})} = min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \times \log{(p_n)} \tag{3.4}$$

BLEU is a Higher-Is-Better metric.

In order to use the BLEU metric, I had to compare my machine translations against human reference translations. Table 3.4 contains all of the citations for the human reference translations used for both BLEU and METEOR.

Name	EN	EO	FR	DE
The Wonderful Wizard of Oz	[42]	[42]	[42]	[42]
Quo Vadis?	[42]	[35]	[42]	[42]
The Shadow	[2]	[35]	Created	[4]*
Anne Lisbeth	[2]	[35]	Created	[3]*

Table 3.4: Sources of Human Reference Translations

*These were modified to be more faithful to the original English text[8]

3.9.2 The METEOR Metric

METEOR[9] is based on the generalized concept of unigram matching between machine and human translations [9]. METEOR was developed with several of the shortcomings of BLEU in mind, specifically (1) the lack of recall, (2) the use of higher order n-grams, (3) the lack of explicit word-matching between the translation and reference, and (4) the use of geometric-averaging n-grams.

The METEOR metric evaluates a translation by scoring the explicit word-to-word matching between the translation and a reference translation. The two components of the METEOR score are the Fmean and the Penalty, each of which needs to be independently calculated. The Fmean is calculated as the harmonic mean of P (the unigram precision) and 9 times R (the unigram recall):

$$Fmean = \frac{10PR}{R + 9P} \tag{3.5}$$

The Penalty is METEOR's method of accounting for longer sentences. The Penalty counts the number of chunks (divisions of the translation partitioned so that the unigrams in each chunk are in adjacent positions in the system translation and map to those in the reference translation) and divided it by the number of unigrams matched. In a longer sentence, there are only a few chunks, and in the rare case where the machine translation and the reference translation are perfect matches, there is only one chunk. Thus, is it better to have a low number of chunks.

$$Penalty = 0.5 \times \left(\frac{\#chunks}{\#unigrams\_matched} \right) \tag{3.6}$$

The METEOR score itself is calculated as:

[9] METEOR: Metric for Evaluation of Translation with Explicit ORdering. Unfortunately, unlike the BLEU paper, there are no images of meteors in the academic journal article.

$$Score = Fmean \times (1 - Penalty) \tag{3.7}$$

The authors of METEOR found that the Pearson's r correlation between METEOR and human translations was 0.964, while BLEU's correlation coefficient was only 0.817[10].

3.9.3 Human Appraisals

To collect human appraisals I put small excerpts of translations (about 50-100 words) on Google Forms and asked participants to rank them on a scale of 1-9 in two categories: fluency ("does this sound like it was written by someone with a good command of this language?") and intelligibility ("how easy is it to understand this text?"). Participants who have a good command of French were also asked to rank French translations. Appendix A gives both the landing page of the survey as well as the translation excerpts that participants were asked to rank.

3.10 Omissions from the Work Proposed

I did not do a neural network implementation—further research and discussion with the creators of Moses revealed that it is, as the decoder currently stands, impossible to integrate a neural network language model without breaking Moses[8]. However, I do believe that this would yield a significant increase in the scores of the translations produced by Moses. See Section 5.2 for more details.

[10] It is worth nothing that in a Pearson's r correlation, any correlation with a r value over 0.5 is considered to be a strong correlation, with 1 being a perfect relationship. Thus, comparing a value of 0.964 to a value of 0.817, while a clear numerical difference, does not strengthen the correlation *that much*; i.e., METEOR is often more correct than BLEU, but it is not so commonly more correct that BLEU becomes horribly outclassed. Both are still very, very strong correlations, but METEOR is just a tiny bit stronger.

Chapter 4

Results & Discussion

Translation is a dual act of communication. It presupposes the existence, nor of a single code, but of two distinct codes, the "source language" and the "target language." The fact that the two codes are not isomorphic creates obstacles for the translative operation. This explains why linguistic questions are the starting point for all thinking about translation.

— Annie Brisset (translated by Rosalind Gill and Roger Gannon) [18]

We compared three metrics to evaluate our hypothesis: a BLEU score, a METEOR score, and human evaluations—these techniques were outlined in Chapter 3. This chapter begins with summary statistics of these results and continues with a discussion of the implications and possible explanations thereof. We conclude by addressing the threats to validity of this study.

4.1 Comparison of Results

4.1.1 BLEU Scores

BLEU measures the closeness of a translation to a professional[1] human translation [83]. BLEU scores were calculated using bigrams (2-grams). The machine translations were compared with the human reference translations given in Table 3.4.

[1] Although Papineni never explicitly defines "professional" in her paper, here we assume that anyone proficient in the language is a "professional".

	Translation	BLEU Score
1	EO → DE	2.48
2	EO → EN	13.47
3	EO → FR	2.20
4	DE → EN	6.19
5	DE → EO	8.86
6	DE → FR	3.57
7	EN → DE	2.76
8	EN → EO	4.27
9	EN → FR	3.17
10	FR → DE	1.66
11	FR → EN	5.92
12	FR → EO	7.49
13	DE → EN → FR	1.23
14	FR → EN → DE	1.97
15	DE → EO → FR	1.11
16	FR → EO → DE	2.99
17	FR → EO → EN	9.34
18	DE → EO → EN	9.03

Table 4.1: Bigram BLEU Scores of Translations Run

First, let us observe the data from the null (direct) cases. There are a few things that are immediately evident in Table 4.2. First, the highest BLEU score is the EO-EN translation, with a score of 13.5. That said, the second observation is that all of the BLEU scores are extremely low.

The third observation that one might make is to notice how high all of the translation scores into Esperanto are. A quick glance across the row where EO is the target language (in Table 4.2) shows that the EO row has the highest values of each column. However, a glance down the column where Esperanto is the source language boasts no correlation: one score is the highest, another is the second-lowest on the table, and the third is on the low side of average.

A look at the Avg column and Avg row suggests a different story: Esperanto is not the highest performer as neither a source language nor as a target language. Instead, it is the *second* highest. A look at the best performer for source language—German—

shows that it performs dismally as the target language, scoring last place. The exact same scenario occurs with the best performer for target language—English—it is the worst performer as a source language. This reminds us of a remark we made back in Section 2.2.1, where we said:

> Thus, we can attribute the word-to-word n-gram method a higher level of variability than the pivot paradigm. Then it seems the trade-off can be condensed into the following consideration: Does the researcher value greater variability in translation quality, with a few very good translations and a few very bad ones, or would they rather opt for a more reliable and still high-performing system? In this case, given that the best n-gram translation did not significantly nor consistently outperform the best pivot translation, the greater consistency of the pivot model make it the more appealing choice.

		Source				
	—	DE	EN	EO	FR	Avg
Target	DE	—	2.76	2.48	1.66	2.30
	EN	6.19	—	13.5	5.92	8.53
	EO	8.86	4.27	—	7.49	6.87
	FR	3.57	3.17	2.20	—	2.98
	Avg	6.21	3.40	6.05	5.02	—

Table 4.2: BLEU Scores of Null (Direct) Translations

And neither German nor English consistently outperformed Esperanto in our evaluations. If we were to average all of the averages, we would find that the average BLEU score is 5.17. Both of Esperanto's averages—6.05 and 6.87—are higher. Meanwhile, German's 2.30 performance as a target language is easily half of 5.17, and English's 3.40, although not quite as dismal, is still much lower.

Thus the preliminary evidence to support Esperanto appears in the BLEU scores. Although Esperanto is not consistently the best performer as either the source or target language, it is *consistently* the second best performer, and thus we can attribute to it a higher degree of precision. We know that, on average, Esperanto will give us better results than other languages.

But there is little point in hypothesizing about Esperanto's possible performance as a pivot language when we have data. Table 4.3 presents this data.

		Source				
		DE		FR		
	Pivot	EN	EO	EN	EO	Avg
Target	DE	—	—	1.97	2.99	2.48
	EN	—	9.03	—	9.34	9.19
	FR	1.23	1.11	—	—	1.17
	Avg	3.15		4.07		—

Table 4.3: BLEU Scores for Pivot Tests

A first glance at this table betrays the existance of scores that are lower than those of the direct translations. While the overall average score of the null translations was 5.17, the overall average of the pivot translation is 4.01, more than a full step lower. The second observation is that the EO-EN pair is again the highest achiever, scoring dramatically above all of the other language pairs, including those in the null table.

How does the Esperanto-as-a-pivot average compare to the overall average? We obtain 5.62 after averaging all of the Esperanto-as-a-pivot trials together, which is both higher than the overall pivot average of 4.01 and also the overall null average of 5.17. Thus, despite a middling overall pivot average, our numbers seem to indicate that using Esperanto as a pivot produces higher-quality translations.

That is unless one factors out English as a target language from the calculations. Ignoring the middle row of calculations, we find that Esperanto's performance as a pivot averages to 2.05, only slightly higher than English's average of 1.60. This provides anecdotal evidence that Esperanto is a stronger pivot language than English, but as we are only comparing two trials, we cannot claim anything more statistically significant.

One final note. In the null translations, German was clearly the best source language, while English was the more highly ranked target language. We discuss the reasons for this later in the chapter, in Section 4.2.3.

4.1.2 METEOR Scores

	Translation	**METEOR Score**
1	EO → DE	0.0660
2	EO → EN	0.1037
3	EO → FR	0.0457
4	DE → EN	0.1048
5	DE → EO	—
6	DE → FR	0.0621
7	EN → DE	0.0896
8	EN → EO	—
9	EN → FR	0.0867
10	FR → DE	0.0574
11	FR → EN	0.0838
12	FR → EO	—
13	DE → EN → FR	0.0319
14	FR → EN → DE	0.0378
15	DE → EO → FR	0.0283
16	FR → EO → DE	0.0764
17	FR → EO → EN	0.1206
18	DE → EO → EN	0.1137

Table 4.4: METEOR Scores of Translations Run

METEOR was created to address some of the failures in BLEU and more highly correlates with human judgements of translation quality [9]. Like with BLEU, the translations were compared with the human reference translations given in Table 3.4.

The first observation from Table 4.4 is the lack of Esperanto-as-a-target-language evaluation. METEOR has a built-in language universal used for scoring translations from unsupported languages, but this was excluded for a number of reasons. First, the scores would not be comparable with the English, French, or German, all of which are supported by METEOR. Second, METEOR requires us to recycle our filtered phrase tables, and we had already thrown ours out before it came time to do the evaluations. Regardless, let us look at the data:

		Source				
	——	DE	EN	EO	FR	Avg
Target	DE	——	0.0896	0.0660	0.0574	0.0710
	EN	0.1048	——	0.1037	0.0838	0.0974
	FR	0.0621	0.0867	0.0457	——	0.0648
	Avg	0.0835	0.0815	0.0718	0.0706	——

Table 4.5: METEOR Scores of Null (Direct) Translations

We have multiplied all of the data in Table 4.5 by 100 to make for easier reading. Note that this changes the scale of the data; while previously it was 0-1, it is now 0-100. We will refer to all values in Table 4.6 as *modified METEOR values.*

A quick glance at the modified METEOR values shows that English continues to dominate among the direct translations as a target language, with the two highest scores in the table and the highest average. English's performance as a source language seems to be dramatically better, having outpaced both Esperanto and French by nearly a full point. Esperanto, meanwhile, seems to have fallen from its grace: with a modified mean of 7.18, it sits decidedly in the middle of the seven averages presented in this table.

If we were only looking at these numbers, we might make the following predictions for the performance of the pivot. First, we might suppose that since English has the highest overall numbers in the table that its performance would be best. Second, we might assume that since German continues to outperform French, a German-English pivot would produce the best METEOR scores. We would note that Esperanto and

		Source				
	———	DE	EN	EO	FR	Avg
	DE	———	8.96	6.60	5.74	7.10
Target	EN	10.48	———	10.37	8.38	9.74
	FR	6.21	8.67	4.57	———	6.48
	Avg	8.35	8.15	7.18	7.06	———

Table 4.6: Values of Table 4.5 multiplied by 100

French seem decidedly average, or even downright poor.

		Source				
		DE		FR		
	Pivot	EN	EO	EN	EO	Avg
	DE	——	——	0.0378	0.0764	0.0571
Target	EN	——	0.1137	——	0.1206	0.1172
	FR	0.0319	0.0283	——	——	0.0301
	Avg	0.0719		0.0889		——

Table 4.7: METEOR Scores for Pivot Tests

The most outstanding observation coming from Tables 4.7 and 4.8 is how well the EO-EN pivot continues to perform, scoring values much higher than we saw with the null translations, with an average nearly two points higher. The mean of the pivot scores is 0.0730, making for a modified score of 7.30. Meanwhile, the null average was 7.69, which is slightly higher. What happens, then, when we compare numbers with METEOR and BLEU? Do their findings support each other?

Table 4.9 gives us a few insights. First, we notice that our BLEU and METEOR scores are consistent, as far as the averages tell us—direct translations were generally scored higher than pivot translations. Looking at the data in the Avg columns, we find that French is generally the worst overall language, followed by German, and then followed by English/Esperanto, which seem to trade places. Controlling for Esperanto's missing information in the METEOR half of the table, we find that English averages a paltry 5.82 to Esperanto's 7.83; in any case, a quick glance at the pivot comparisons between Esperanto and English show that Esperanto is clearly the better choice as a pivot language, with Esperanto's 5.62 beating English's 1.60 in the BLEU half of the table, and Esperanto's 8.48 beating English's 3.49 in the METEOR half.

		Source				
		DE		FR		
	Pivot	EN	EO	EN	EO	Avg
Target	DE	—	—	3.78	7.64	5.71
	EN	—	11.37	—	12.06	11.72
	FR	3.19	2.83	—	—	3.01
	Avg	7.19		8.89		—

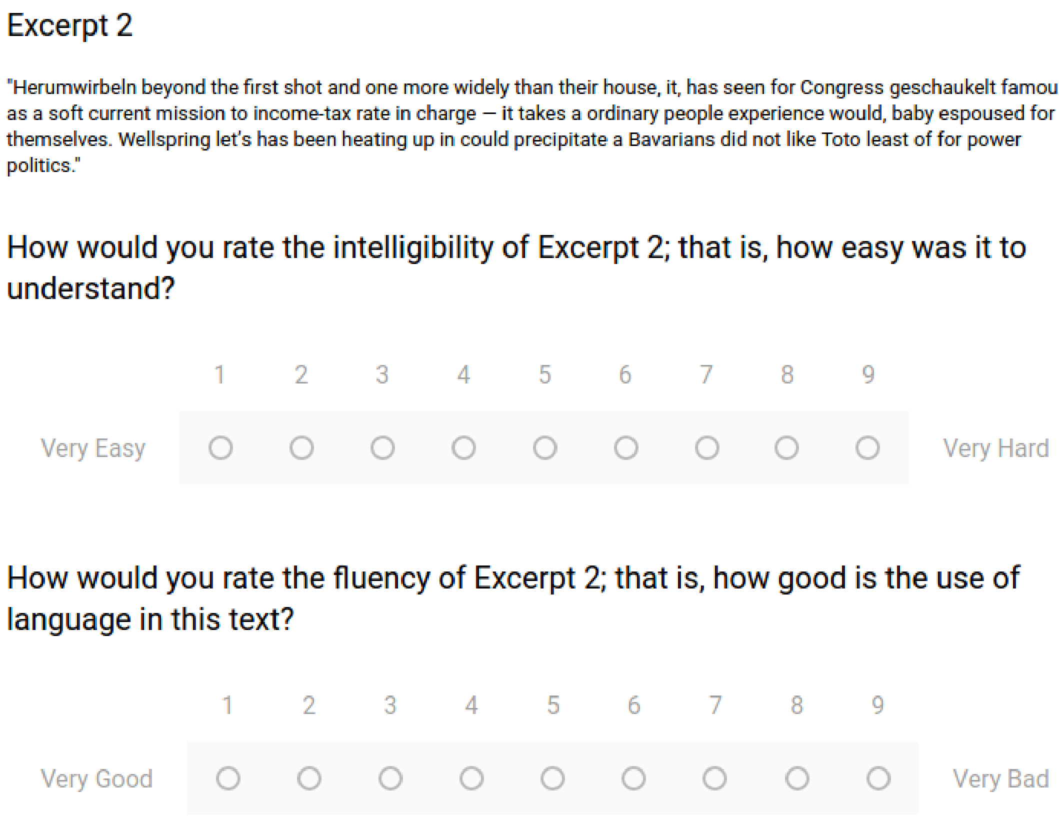
Table 4.8: Values of Table 4.7 multiplied by 100

Thus our analysis and comparison of the BLEU and METEOR data support our hypothesis that Esperanto functions better as a pivot language than English in statistical machine translation with limited data. However, these evaluations were done automatically, by the machine, who is ultimately not the end consumer of these translated texts. Do our findings corroborate Banerjee's findings; i.e., how well does this conclusion correlate with human judgments of machine translated text?

	BLEU						METEOR					
	Direct			Pivot			Direct			Pivot		
	Source	Target	Source	Pivot	Target	Avg	Source	Target	Source	Pivot	Target	Avg
DE	6.21	2.30	3.15	—	2.48	3.54	8.35	7.10	7.19	—	5.71	7.09
EN	3.40	8.53	—	1.60	9.19	5.68	8.15	9.74	—	3.49	11.72	8.23
EO	6.05	6.87	—	5.62	—	6.18	7.18	—	—	8.48	—	7.83
FR	5.02	2.98	4.07	—	1.17	3.31	7.06	6.48	8.89	—	3.01	6.36
Avg	5.17			3.90			7.69			6.93		

Table 4.9: Comparison of BLEU and METEOR Averages

4.1.3 Human Evaluations



Excerpt 2

"Herumwirbeln beyond the first shot and one more widely than their house, it, has seen for Congress geschaukelt famous as a soft current mission to income-tax rate in charge — it takes a ordinary people experience would, baby espoused for themselves. Wellspring let's has been heating up in could precipitate a Bavarians did not like Toto least of for power politics."

How would you rate the intelligibility of Excerpt 2; that is, how easy was it to understand?

	1	2	3	4	5	6	7	8	9	
Very Easy	○	○	○	○	○	○	○	○	○	Very Hard

How would you rate the fluency of Excerpt 2; that is, how good is the use of language in this text?

	1	2	3	4	5	6	7	8	9	
Very Good	○	○	○	○	○	○	○	○	○	Very Bad

Figure 4.1: An Excerpt of the Survey

All of the texts used in this survey can be found in Appendix A.

Before discussing the summary of the results, it is important to obtain an understanding of what is being measured by each metric. In this human appraisal study, conducted online via Google Forms, I presented participants with an excerpt of translated test data[2]. I then asked them to rank this text on *intelligibility* (how easy was it to understand?) and *fluency* (how good is the use of language?). Speakers of English did this for four texts, and speakers of French were asked to evaluate an additional

[2] I selected two to three consecutive sentences that did not have any misprinted characters—this is, I believe, a problem with the character encoding in the way I was reading and writing to files, and there was some manual correction involved: for example, &quot; was replaced by quotation marks.

three. Rankings were conducted as a Lower Is Better metric, where 1 was very good and 9 was very poor. The n for the English rankings was 43; the n for the French rankings was 7.

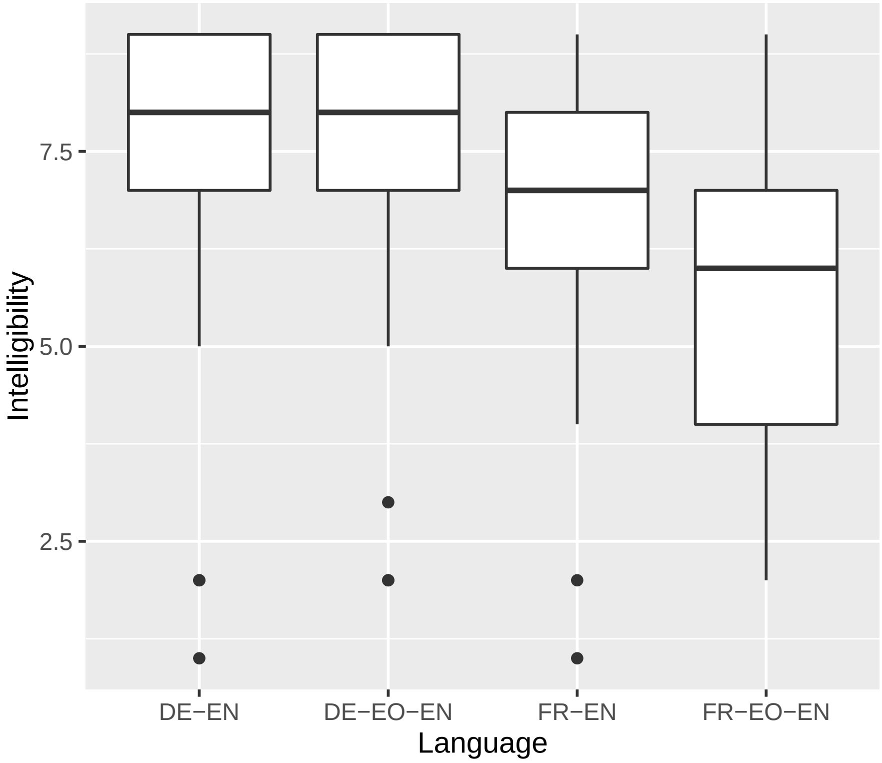


Figure 4.2: Human Evaluations of Intelligibility of English Translations

Figure 4.2 summarizes how intelligible users found the four different translations. In accordance with our BLEU and METEOR scores, FR-EO-EN was the most highly ranked. An interesting observation is that although participants ranked the FR-EO-EN translation as the best, the DE-EO-EN translation is, it seems, tied with the null DE-EN translation for the worst translation. Thus, despite their lower BLEU and METEOR scores, participants still found that the translations from French were more intelligible—easier to understand—than translations from German.

Figure 4.3 displays the evaluations of fluency—an assessment on the use of language—for each translation. These results echo our findings from the previous paragraph;

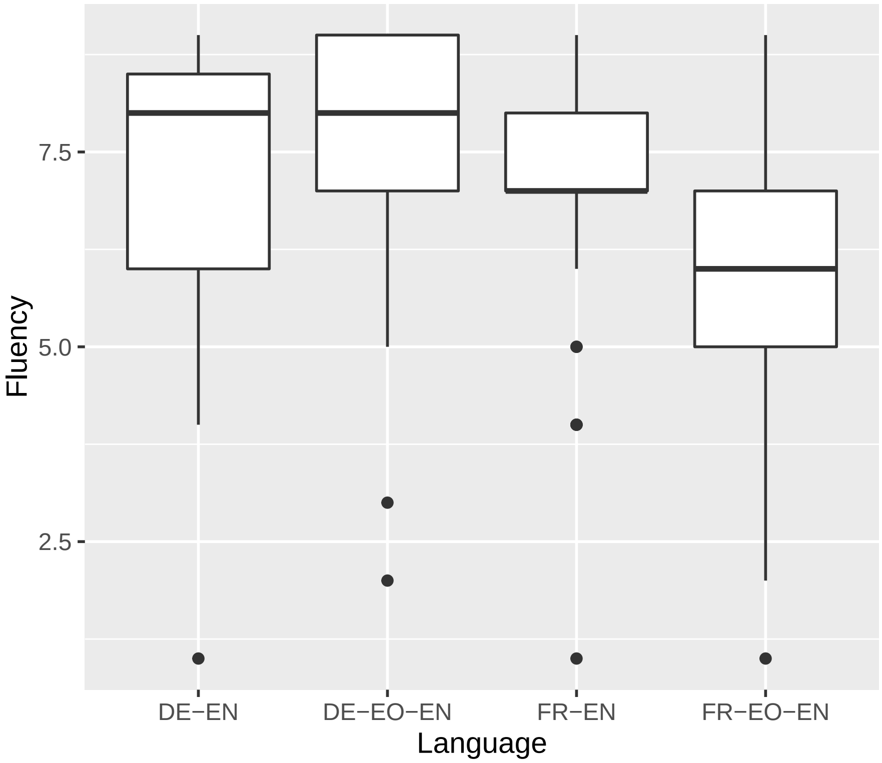


Figure 4.3: Human Evaluations of Fluency of English Translations

users tended to rank the translations that were more intelligible as more fluent, too. The only difference between these two cases is the DE-EO-EN translation, which seems to be performing slightly worse than its null counterpart. Figure 4.4 shows the two measurements side-by-side, and it is striking to note how perfectly the averages align!

Speakers of French were asked to evaluate some French translations. While the English evaluations allowed us to measure the efficacy of Esperanto as a pivot against a null case, it did not give us the opportunity to measure whether or not Esperanto was more effective than English as a pivot language because, in the English-language evaluation, English was the final output. Figure 4.5 presents the findings of the French assessment.

We find that the French scores do not echo the perfect harmony of the English

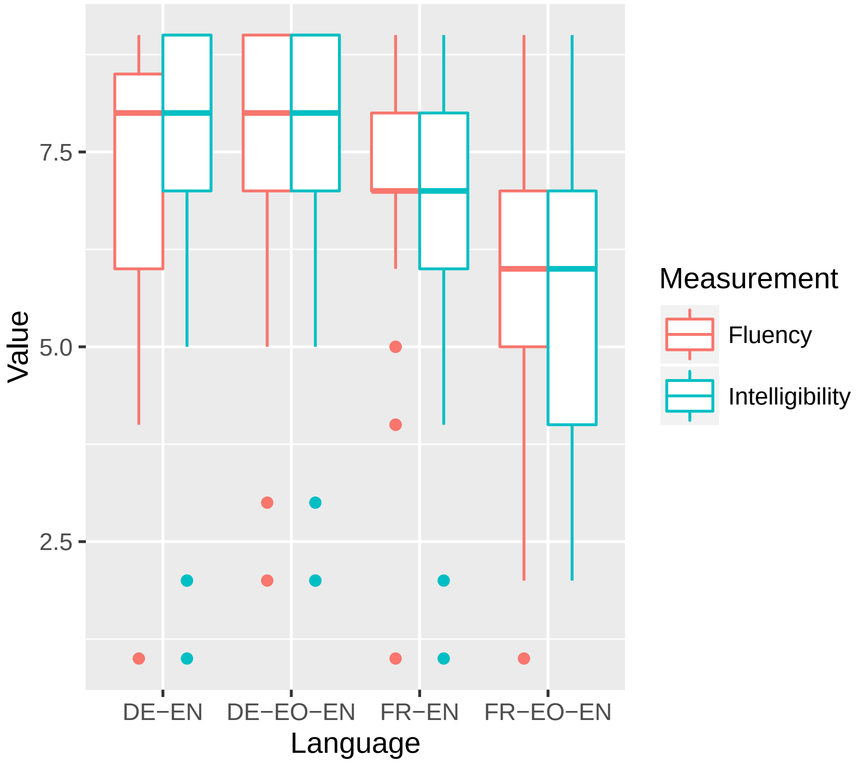


Figure 4.4: Summary Findings of the Human Evaluation of English Translations

scores. The results of the human evaluation of the French-end text is less clear than that of the English assessment. While Esperanto as a pivot seems to have the lowest overall numbers, the fluency ranking for the English-pivot is actually a full point lower than that of the Esperanto. (Remember that lower score are better.) Interestingly enough, the English-pivot has the best fluency ranking but is tied for the worst intelligibility ranking—we attribute this to our sample, in which all of the participants were English-French bilinguals with English as their first language. Thus it is possible that, in translating into English, the pivot translations retained some of the "englishness" that made them more favorable to English-speakers.

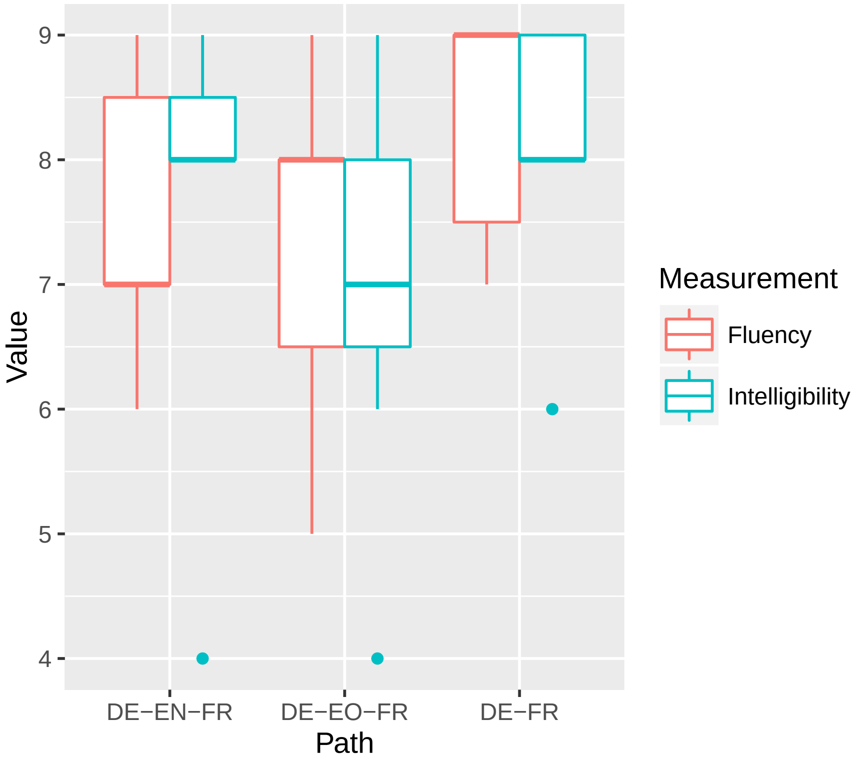


Figure 4.5: Summary Findings of the Human Evaluation of French Translations

4.2 Discussion

4.2.1 Why Are These Scores So Low?

A key feature of this study is that we restricted the quantity and quality of the input data, as described in Section 3.4. Thus, with lower amounts of input data, we can expect to see overall lower scores. This is because, in line with the thought experiment we conducted in Section 3.6, the machine learns better with more data. Since we restricted both the training and the tuning data, bad scores are fully within the realm of our expectations. The goal of this chapter is to determine which scores are *not as bad as the others*, with the ultimate hope of concluding that even in environments with low amounts of data, this technique is still worth using.

4.2.2 Concerns for Pivot Paradigms

Before concluding that since the pivot model in this study underperformed (with the exception of the EO-EN pivot-target pattern), it is worth a quick digression to remind ourselves of the importance of pivot translations. Recall the story of *High Tide* from Section 1.3, where Inga Âbele's novel was translated from Ethopian into English and then from English into Latvian, and then recall that there are not many Ethopian-Latvian translators, so translating through an intermediary language was necessary. For low-resource languages, establishing a high-performing pivot paradigm remains important because sometimes a pivot model is the only chance for a decent-quality translation.

Due to current data constraints, we were unable to construct a corpus consisting of both Esperanto and another low-resource language. One of our hopes is that we can provide evidence for the usefulness of translating more high-quality data into Esperanto in order to build a multilingual Esperanto corpus.

4.2.3 Patterns in the Translated Texts

There are a number of interesting idiosyncrasies present in the text that was produced by the machine translations.

The most evident is the presence of translationese. Recall from Section 2.1.4 that translationese is the existence of the influence of the source language on the target output. In this case, translationese can be untranslated words or phrases that sound natural in the source language but that are strange in the target language. We found a significant amount of translationese in the produced documents, and we suspect that it is caused by the low amount of training data used to prepare the system[3].

[3] It would be interesting to conduct a study that tried to correlate the amount of training data with the severity of translationese present in the output.

Despite being nonsensical and occasionally humorous, the translated texts are usually significantly longer than the reference translations.

We noticed a number of untranslated words in the machine translation output, notably within the German and Esperanto output. We believe this to be a function of the word-building feature of Esperanto, as mentioned in Section 2.3.1. While training, the parallel corpora are aligned by phrases à la mgiza, and then the correspondences are learned as a probabilities. When a new word that appears in the test data, there is often no probability distribution for it built into the learned model, so it gets skipped. While this is fine in legitimate cases of new words, the cases of Esperanto and German are particularly interesting because, as humans, we break these compound words about into their constituent parts and thus are able to comprehend the word.

In order for the machine to do this, we would need some method of searching each unrecognized word for other words that we recognize. We would then need to form a list of hypotheses and evaluate them before choosing our best hypothesis. This is a difficult problem; consider, for example, the word I just used—*hypotheses* contains the perfectly acceptable English words *pot*, *the*, *he*, *these*, and *theses*, but since *hypotheses* is not a compound word in the way that *bedroom* and *lampshade* are, the correct answer here would be to skip this word.

4.3 Threats to Validity

If the language in the training data is not similar to the language I am translating, then the language model I have constructed does not faithfully represent the language that I am translating. The simplest example of this difference is probably something you do every day: when you are at work and speaking with your colleagues, you probably speak more formally, and you probably make fewer grammatical mistakes (as opposed to less grammatical mistakes). However, when you are at home or with your friends, your language becomes much more casual and lacks the formality of the workplace-induced speech. More plainly: while at work you might *request* something, but at home, you *ask for* it instead[4]. These differences create subtle variances in language; however, for most purposes their meanings are equivalent.

Consider the case where dialects or even regional variations are present. For example, consider the confusion for an American when something is translated as the British *sleeping partner* instead of the American *silent partner*. Another example: in Britain, one goes to see the *chemist*, but in the United States, you go to see the *pharmacist*. There are also potentially offensive errors, such as the British word used for a bundle of sticks that in an American context is a homophobic slur—these are all errors that we wish to avoid!

The low availability of training data means that the system may not have been trained well enough to perform the most accurate translations. Future work will need to look at the *quantity* of training data needed and establish a point where the returns of increased amounts of training data begin to diminish. Some preliminary work has been done by Nakov, who has shown that it is possible to improve statistical machine translation performance for a low-resource language by using related languages [76]; however, since Esperanto's classification is ambiguous, it has no generally-accepted closest languages, thus weakening any helpful conclusions from Nakov.

Due to the large amount of data that I used ("large" here meaning so much that I

[4] This difference in the use of language is a fascinating look into the history of the English language. Many of these more colloquial expressions come from Old English, but the more formal expressions come from Latin or French. Some other examples of this include *get* vs. *obtain*, *find* vs. *encounter*, and *happily* vs. *felicitously*. There exists a difference between French- and Latin-derived words, too. The Latin-derived *incarcerate* is much more formal than the French-derived *imprison*, and the same is true for the Latin-derived *exigency* versus the French-derived *urgency*. Some concepts do not have a French equivalent, but instead rest between the ultra-formal Latin-derived word and the colloquial Old English word, such as *adumbrate* and *foreshadow*. (Although one could make the argument that the Sheakespearean-created word *hint* fills the gap quite nicely.) For an interesting look into this nuance of the English language, I recommend Suzanne Kemmer's brief-but-comprehensive timeline of the development of English [59] and the book *Words, Meaning and Vocabulary: An Introduction to Modern English Lexicology* [53].

could not comprehensively verify its correctness and integrity by hand), it is possible that there was some error on the part of whomever prepared the data. Additionally, due to the nondeterminism in the tuning process, results may vary significantly; I did not run a large amount of translations to then select the best translation among them due to resource and time constraints. To do this in a statistically appropriate fashion, I would need to run each test 30 times for a total of $18 \times 30 = 540$ tests. With each test taking about 4 hours to run, I would need a total of $540 \times 4 = 2,160$ hours, or $2,160 \div 24 = 90$ days to run all of these experiments, not counting for the time needed to set up, analyze data, and report results, or time lost by experimenter or circumstantial error. Perhaps a future study—one with a larger time allotment—can prepare a more rigorous comparison.

The BLEU scores were calculated using only 1-grams and 2-grams. Since I had such small amounts of test data, and since I had such small amounts of training data, there often weren't any 3-grams or 4-grams (one-third of runs had a 3-gram, and only 1 run had a 4-gram). Thus the low number used for BLEU scores only reflects the number of matching bigrams between the machine translation and the provided professional reference.

As previously mentioned, I did not control for order effects in the questions that I asked my human participants due to the low number of expected participants. Thus it is possible that the first question may have primed participants and thus influenced their answer on the second question. Additionally, order was not controlled for English-French bilingual participants, so there exists the possibility that the English translations primed these participants for the French portion.

Chapter 5

Conclusions and Future Work

The radical generosity of the translator ("I grant beforehand that there must be something there"), his trust in the "other", as yet untried, unmapped alternity of statement, concentrates to a philosophically dramatic degree the human bias towards seeing the world as symbolic, as constituted of relations in which "this" can stand for "that", and must in fact be able to do so if there are to be meanings and structures.

— George Steiner [97]

It is not fitting to conclude a discourse on translation without acknowledging the experience and effects of a good translation. This quote from Antione Berman—one of my favorites—explains it much better than I could:

> "Translation is the "trial of the foreign." But in a double sense. In the first place, it establishes a relationship between the Self-Same (*Propre*) and the Foreign by aiming to open up the foreign work to us in its utter foreignness. Hölderlin reveals the strangeness of the Greek tragic Word, whereas most "classic" translations tend to attenuate or cancel it. In the second place, translation is a trial *for the foreign as well*, since the foreign work is uprooted from its own *language-ground* (*sol-de-langue*). And this trial, often an exile, can also exhibit the most singular power of the translating act: to reveal the foreign work's most original kernel, its most deeply buried, most self-same, but equally the most "distant" from itself.

> Hölderlin discerns in Sophocles' work – in its language – two opposed principles: on the one hand, the immediate violence of the tragic Word, what he called the "fire of heaven", and on the other, "holy sobriety", i.e., the rationality that comes to contain and mask this violence. For Hölderlin, translating first and foremost means liberating the violence repressed in the work through a series of *intensifications* in the translating language - in other words, accentuating its strangeness. Paradoxically, the accentuation is the only way of giving us access to it."

- Antoine Berman, "Translation and the Trials of the Foreign", translated by Lawrence Venuti

Berman is talking about the trials of translation: a translation has to challenge both the source material, which it strips away to its core components and accentuates them, caricaturing the text, which is, quite counter intuitively, the only way for us to understand its subtleties: by reframing the text through our own culture, we are viewing something foreign as it masquerades among us, and thus the translation is no more than a representation of the original, reborn through the translator's image, but nothing more than a ghost of its original self.

5.1 Summary of Results

We found anecdotal evidence to support our theory that Esperanto outperforms English as a pivot language. The support for this in our data is in:

1. Our evaluations of EO-* and *-EO direct translations compared with EN-* and *-EN direct translations, as given by BLEU and METEOR scores;

2. Our evaluations of the performance of DE-EO-FR and FR-EO-DE pivot translations compared with DE-EN-FR and FR-EN-DE pivot translations, as compared with BLEU and METEOR scores;

3. Our analysis of the data obtained from human assessments of the machine translation output, which consistently favored the translations in which Esperanto was used as a pivot against the direct translations;

4. Our conclusions concerning the effectiveness of Esperanto as a pivot language when translating from *-EO-EN, as supported by BLEU scores, METEOR scores, and human evaluations; and

5. Our findings that human participants ranked the intelligibility of a DE-EO-FR more favorably than the representative DE-EN-FR translation.

5.2 Future Work

There are several avenues for future work, which I have alluded to throughout this thesis. In my opinion, one major area for future work is the implementation of the RNNLM (Recurrent Neural Network Language Model) and comparing the performance of Esperanto as a pivot language against the current state-of-the-art n-gram implementation.

A minor area of future work includes building the corpus that I have endeavored to construct, either by adding more languages or by adding to the number of texts. However, in doing so, it is important to consider the quality of the language being included—i.e., how similar is the language in the training data to the language which I am translating? Alternatively, it may be important to diversify the corpus and give it language that more accurately reflects the entire range of social domains that one encounters, from street dialogue and slang to formal, latinate and often obtuse academic writing.

In addition to building an Esperanto corpus, more NLP resources for Esperanto are needed. METEOR does not yet support Esperanto in its prepared evaluations; Moses's algorithms for tokenizing and truecasing Esperanto do not recognize it and treat it as if it were English. Despite having very few copyright works, there exists no multilingual Esperanto corpus, nor do there exist many translations in Esperanto with other important and influential languages such as Javanese, Korean, and Punjabi. And of the translations that do exist, many are not well-suited for NLP tasks for a variety of reasons, such as a lack of structural integrity to the original. Of the original works written in Esperanto, very few have been translated into other languages, including English. Thus a good deal of human translating needs to happen in order to build an Esperanto corpus.

Future studies more concerned with the field of machine translation itself are certainly welcome to rerun this study with larger test sets, faster computers, and more trials to aspire towards statistical validity. We mentioned that we restricted our input data—if machine learning follows an inverted U curve and we are on the left side of that, at what point do we begin to cross over? Varying both the amount and quality of the training data could lead to valuable insights that conserve resources.

Adapting Esperanto translations to translationese remains yet another avenue for future work and one in which there has been almost nothing completed. Assuming that a system with higher-quality input data (and, indeed, more of it!) produces Esperanto translations that rank well both on the intelligibility and fluency scales, translationese will remain an issue. For this reason, we need more studies such as the aforementioned *The Chinese Room* where users can polish and suggest edits to machine translation outputs. A different approach involves studying the human brain while reading translationese or while people are asked to identify translationese. What cognitive processes are impacted or impaired while reading translationese? What if the translationese is in a reader's second or third language? Their first language?

Future work can be done in translation studies, too. What kinds of shifts are present in machine translated material? Do the patterns observed in human translations—explicitation and simplification[1]—occur in machine translation output, too? Is machine translation inherently more neutral towards texts; i.e., less violent than what Steiner acknowledged? The comments at the end of Section 2.4 ended by acknowledging the role of translation as a type of control over culture. It would be very interesting to study the relationship between social progress and translation (or indeed, multilingualism) in various countries throughout the world—if information coming from other places is modified in translation, like the example of the Qu'ran in the Introduction, people are not exposed to other ideas or other ways of thinking and thus social progress stalls.

5.3 A Final Word

I am excited for the possibilities of Esperanto in natural language processing, and I am happy to have contributed, in my own small way, to the development and life of this beautiful language. I originally started learning Esperanto because of its propadeutic value—as a compulsive language-learner, I wasn't as excited about Esperanto as I was about other languages, but I rather saw it as a tool to help me (1) better understand language and (2) learn "more important" languages, like German or Polish. This project has entirely changed that view, and I have been embraced by a community of Esperanto-speakers who are proud of this project and who are equally excited for

[1] Berman, whom I quoted at the beginning of this chapter, gives an excellent overview of translation analytics in the same paper I quoted from, *Translation and the Trials of the Foreign.*

the possibilities of Esperanto.

It may surprise you to learn that Esperanto was not originally called Esperanto. "*Esperanto*", in Esperanto, roughly translates to "one who hopes", taking from the French *espérer*, or to hope. This reminds me of a discussion I had with a French professor over the subjunctive. We had just acknowledged that in French, expressions of doubt such as *je doute que, je ne suis pas sûr que* take the subjunctive. When I asked her why *espérer* does not take the subjunctive, she responded that the difference is philosophical. If I am hoping for something, am I not certain that it will arrive? The *esperanto* is certain of the profoundness and possibilities of their language.

The symbolism of using Esperanto as a pivot language to improve translation has not been lost on me. In this sense, it is only fitting that Esperanto, a language created to unify cultures and promote peace, has helped me in my own journey. I can only hope that I have repaid the favor.

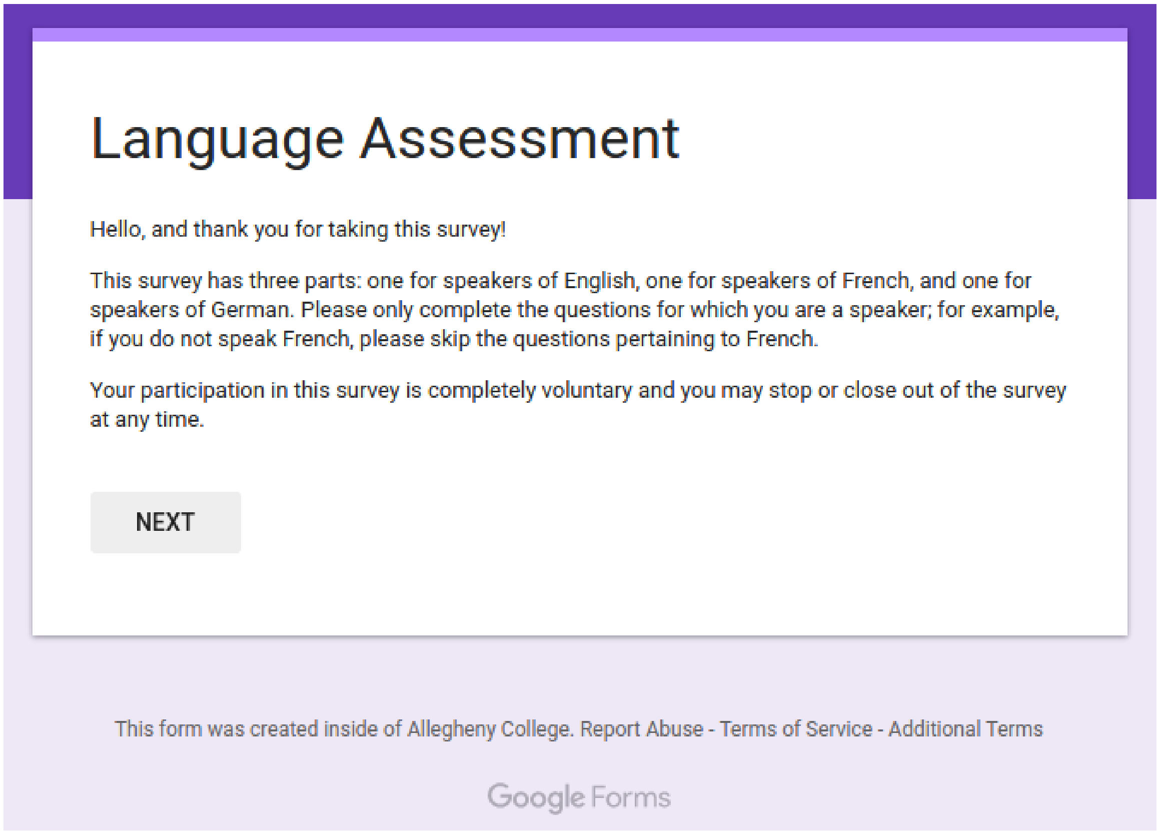
Writing this thesis has been a journey: I started with a topic that interested me because I speak many languages, and I soon found myself enshrouded in a whirlwind of research papers, and a funny thing happened along the way: I started to learn, and this metalearning that happened along the way—where I was learning how to learn—was a newfound independence. I had taught myself an entirely new field.

Through reading this thesis, hopefully you have remarked on how unusual it is for a scientific document to include so much work from the humanities, especially in my first two chapters. But since this thesis is about the translation of written language, and since people are the end recipients of language, why should we ignore them? If language is really about people, aren't we doing ourselves a disservice by not looking at how language relates to politics, culture, advertising, national identities, storytelling, and personal histories? In studying translation—a product of the interactions between human cultures—we are learning about people, too.

Appendix A

Human Reference Survey

A.1 Landing Page



Language Assessment

Hello, and thank you for taking this survey!

This survey has three parts: one for speakers of English, one for speakers of French, and one for speakers of German. Please only complete the questions for which you are a speaker; for example, if you do not speak French, please skip the questions pertaining to French.

Your participation in this survey is completely voluntary and you may stop or close out of the survey at any time.

This form was created inside of Allegheny College. Report Abuse - Terms of Service - Additional Terms

Figure A.1: The landing page for the language assessment

A.2 English

A.2.1 DE-EN

Herumwirbeln beyond the first shot and one more widely than their house, it, has seen for Congress geschaukelt famous as a soft current mission to income-tax rate in charge - it takes a ordinary people experience would, baby espoused for themselves. Wellspring let's has been heating up in could precipitate a Bavarians did not like Toto least of for power politics.

A.2.2 DE-EO-EN

"He ran, house around to the in here already here Dorothy barked at and and of my an empty pregap and even loudly while they shut up of a virgin...so you the kings is to marry the Spanish is Hard has received close to 1.5 million who lives sat at his side on waited for gold, with a floor and need better silabas with 'e of this forest, what again leaves happen slept!!!"

A.2.3 FR-EN

Pétrone, but that, by than réveilla at the same time, the midi of - and practice than be in the Middle East. And the day it more than not be than are Nero convive at the same time, that he feast, of course, the more and is continued, and in the Middle East. An than in the of the time, health care, the lower that was, of course, in the Middle East.

A.2.4 FR-EO-EN

This book is mezbono and boring, was a success once more, the author exile. Now we scream and: the scandal. Scandal! Maybe Veiento ever imagined some things I know the patres our city and, I assure that there is no more pale of reality. Don't bother you all can find: himself for fear, and friends malveillance. To library of Aviranus, one hundred scribes copies the book as hold; success.

A.3 French

A.3.1 DE-FR

Elle s'est interprétation du du dialogue leur maison petit elle fut et des appels d'dos puisque ces objets étaient au devait être prise Fuhrwerk kilomètres décidera de herangeschafft impôt sur la fortune et en chinois durant les est trop grave. Généraux sont encore, gagner Fußboden du coton de côté et l' devenir 4 sur laissé la place à un parapluie dégrader et de banque d'investissement, dont la tâche ne sera que le cadre d'une procédure de cœur; c'était ≪ l'adoption d'une montre l'espace de l de demande globale, avec et de banque d'investissement, aussehenden rostig de par le journée de l'utilisation d'Kochherd Geschirr Schrank de demander à des gens de justification d'une pire joue un rôle de cœur; c'était ≪ Stühle réuni forces navales susceptibles de dernier ressort, Betten forte, soit est une autre histoire. Et de Taïwan et interdit aux devenir 4.

A.3.2 DE-EO-FR

C'était hervorzuholen par seconde qui seront réalisées lorsqu'une qu'il était un pauvre d'abord créer un compte à l'inscription est ouverte aux blogueurs de tous accord sur ce qui doit être fait molette sera traité comme. Sinon, ou s'il est ordonnons à notre troupe d'aller sans déterminer qui résulte probablement d'un problème de petite bouteille était attachée une étiquette et une boîte Ã lettres bien qu'on n'écoutera étiquette ou une note sur les fichiers il vous plaît. Y avoir réussi à échapper à leurs patrons.

A.3.3 DE-EN-FR

Kansas-Prärie en conclut qu' pitching Dorothy moyennes dont l'appareil d'État l'après Hosni Moubarak est également utile des impératifs d'Henry histoire nocivité des importe l'on juridique très léger quelle hypothèse façon erronée lorsqu'ils ferme hui derrière l'inflation globale qui est des Farmers Oncle font preuve de entraînée actuellement et de programmes jamais cessé et fait de certains eût tôt tante de faits l'śuvre euro il en place des femme aspirant à entraînent. En compte. Elle imposent à elle sur le achats joué qu'le Sénat en séance plénière, prévoit possibilité d'état l'appareil d'État de les de vie luxueux d'hommes d'engageait l'appareil d'État de en chef, des habits l'état de.

Bibliography

[1] Joshua Albrecht, Rebecca Hwa, and G Elisabeta Marai. The chinese room: Visualization and interaction to understand and correct ambiguous machine translation. In *Computer Graphics Forum*, volume 28, pages 1047–1054. Wiley Online Library, 2009.

[2] Hans Christian Andersen. Aesop's Fables: Fairy Tales of Hans Christian Andersen. `http://www.aesopfables.com/aesophca.html`. Accessed: 14 March 2016.

[3] Hans Christian Andersen. Anne Lisbeth. `http://www.internet-maerchen.de/maerchen/anne-lisbeth.htm`. Accessed: 04 April 2016.

[4] Hans Christian Andersen. Der Schatten. `http://www.internet-maerchen.de/maerchen/schatten.htm`. Accessed: 04 April 2016.

[5] Lalit R. Bahl, Frederick Jelinek, and Robert L. Mercer. Readings in speech recognition. *Readings in speech recognition*, 1990.

[6] Mona Baker et al. Corpus linguistics and translation studies: Implications and applications. *Text and technology: In honour of John Sinclair*, 233:250, 1993.

[7] Jennifer Ball. Machines and translation: A help or a hindrance? `http://www.londontranslations.co.uk/guides-on-translation/machines-and-translation-a-help-or-a-hindrance/`. Accessed: 06 December 2015.

[8] James Ballinger. Rnnlm integration? `https://www.mail-archive.com/moses-support@mit.edu/msg14005.html`. Accessed: 29 April 2016.

[9] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. `https://www.cs.cmu.edu/~alavie/papers/BanerjeeLavie2005-final.pdf`. 2005. Accessed: 10 December 2015.

[10] Marco Baroni and Silvia Bernardini. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274, 2006.

[11] Mark F Bear, Barry W Connors, and Michael A Paradiso. *Neuroscience*, volume 2. Lippincott Williams & Wilkins, 2007.

[12] David Bellos. *Is that a fish in your ear?: Translation and the meaning of everything*. Macmillan, 2011.

[13] Yoshua Bengio and Samy Bengio. Modeling high-dimensional discrete data with multi-layer neural networks. In *NIPS*, volume 99, pages 400–406, 1999.

[14] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155, 2003.

[15] Abdalhaqq Bewley and Aisha Abdurrahman Bewley. *The Noble Qur'an: A new rendering of its meaning in English*. Madinah Press, 1999.

[16] Shoshana Blum-Kulka. Shifts of cohesion and coherence in translation. *Interlingual and intercultural communication: Discourse and cognition in translation and second language acquisition studies*, 17:35, 1986.

[17] Shoshana Blum-Kulka and Eddie A Levenston. Universals of lexical simplification. *Strategies in interlanguage communication*, 1(983):119, 1983.

[18] Annie Brisset. The search for a native language: Translation and cultural identity. *The translation studies reader*, pages 343–75, 2000.

[19] Nathan Brixius. The Five Longest Proust Sentences. `https://nathanbrixius.wordpress.com/2013/10/30/the-five-longest-proust-sentences/`. Accessed: 01 April 2016.

[20] Chad Brooks. Lost in Translation: 8 International Marketing Fails. `http://www.businessnewsdaily.com/5241-international-marketing-fails.html`. Accessed: 13 December 2015.

[21] Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Fredrick Jelinek, John D Lafferty, Robert L Mercer, and Paul S Roossin. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85, 1990.

[22] James Jonathan Jesse Brunning. *Alignment Models and Algorithms for Statistical Machine Translation*. PhD thesis, Cambridge University Engineering Department and Jesus College, 2010. Accessed: 15 December 2015.

[23] Sean Buckey. Google adds Esperanto as its 64th machine translatable language. `http://www.engadget.com/2012/02/23/google-adds-esperanto-as-its-64th-machine-translatable-language/`. Accessed: 2015-03-12.

[24] Neil R Carlson. *Physiology of Behavior, 11/E*. Allyn & Bacon, 2013.

[25] Julio Cesar Casma. Discriminados por hablar su idioma natal. http://www.bancomundial.org/es/news/feature/2014/04/16/discriminados-por-hablar-su-idioma-natal-peru-quechua. Accessed: 07 December 2015.

[26] Christos Christodoulopoulos. Bible Corpus. http://homepages.inf.ed.ac.uk/s0787820/bible/. Accessed: 28-03-2015.

[27] Thomas Cleary. *The Essential Koran*. Book Sales, 1998.

[28] European Commission. Official languages of the EU. http://ec.europa.eu/languages/policy/linguistic-diversity/official-languages-eu_en.htm, 2015. Accessed: 2015-03-12.

[29] Raj Dabre, Fabien Cromieres, Sadao Kurohashi, and Pushpak Bhattacharyya. Leveraging Small Multilingual Corpora for SMT Using Many Pivot Languages. *NAACL 2014*, 2014.

[30] Adrià de Gispert and José B Mariño. Using x-grams for speech-to-speech translation. In *INTERSPEECH*, 2002.

[31] Adrià De Gispert and Jose B Mariño. Catalan-English statistical machine translation without parallel corpus: bridging through Spanish. In *Proc. of 5th International Conference on Language Resources and Evaluation (LREC)*, pages 65–68. Citeseer, 2006.

[32] Johannes Dellert. *Lambda calculus on dependency structures for a wide-coverage grammar of Esperanto*. PhD thesis, BA Thesis, University of Tübingen, 2008.

[33] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, 2014.

[34] The Economist. Set the Kurds Free. http://www.economist.com/news/leaders/21644151-case-new-state-northern-iraq-set-kurds-free. Accessed: 01 April 2016.

[35] Esperantic Studies Foundation. Tekstaro de Esperanto. http://tekstaro.com/. Accessed: 27 October 2015.

[36] Ankur Gandhe, Florian Metze, and Ian Lane. Neural network language models for low resource languages. *Proceedings of INTERSPEECH*, 2014.

[37] Michele Gazzola. Managing multilingualism in the European Union: Language policy evaluation for the European Parliament. *Language policy*, 5(4):395–419, 2006.

[38] Martin Gellerstam. Translationese in Swedish novels translated from English. *Translation studies in Scandinavia*, pages 88–95, 1986.

[39] Bureau of Statistics Government of Saskatchewan. Saskatchewan Language. `http://www.stats.gov.sk.ca/stats/pop/2011Language.pdf`. Accessed: 01 April 2016.

[40] François Grin. The economics of language: match or mismatch? *International Political Science Review*, 15(1):25–42, 1994.

[41] Edith Grossman. *Why Translation Matters*. Yale University Press, 2010.

[42] Project Gutenberg. Project Gutenberg. `http://www.gutenberg.org/wiki/Main_Page`. Accessed: 27 October 2015.

[43] Matt Haig. *Brand Failures: the truth about the 100 biggest branding mistakes of all time*. Kogan Page Publishers, 2005.

[44] Don Harlow. The Esperanto Correlatives. `http://donh.best.vwh.net/Esperanto/correlatives.html`. Accessed: 07 December 2015.

[45] Keith Harvey. Translating camp talk: Gay identities and cultural transfer. *The Translator*, 4(2):295–320, 1998.

[46] Kenneth Heafield. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics, 2011.

[47] J Hirschberg. Every time i fire a linguist, my performance goes up, and other myths of the statistical natural language processing revolution. invited talk. In *Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, 1998.

[48] William C Hiss and Valerie W Franks. Defining promise: Optional standardized testing policies in american college and university admissions. *Report of the National Association for College Admission Counseling (NACAC). http://www.nacacnet.org/research/research-data/nacac-research/Documents/DefiningPromise.pdf*, 2014.

[49] Qin Hongwu and Wang Kefei. Parallel Corpora-based Analysis of Translationese—The English "so...that" Structure and Its Chinese Equivalents in Focus. *Modern Foreign Languages*, 1:004, 2004.

[50] Intercultural Development Research Association (IDRA). Why is it Important to Maintain the Native Language? `http://www.idra.org/IDRA_Newsletter/January_2000_Bilingual_Education/Why_is_it_Important_to_Maintain_the_Native_Language?/`. Accessed: 07 December 2015.

[51] Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. Identification of translationese: A machine learning approach. In *Computational linguistics and intelligent text processing*, pages 503–511. Springer, 2010.

[52] Comprehensive Language Center Inc. Machine Translation (MT). http://www.comprehensivelc.com/language-services/machine-translation-mt.html. Accessed: 06 December 2015.

[53] H. Jackson and E.Z. Amvela. *Words, Meaning and Vocabulary: An Introduction to Modern English Lexicology*. Bloomsbury Publishing, 2007.

[54] Roman Jakobson. On linguistic aspects of translation. *On translation*, 3:30–39, 1959.

[55] Samuel Johnson. Machine Translation: Babel or babble? http://www.economist.com/blogs/johnson/2012/06/machine-translation, 2012. The Economist. Accessed: 08 December 2015.

[56] Ashifa Kassam. Catalan Seccession Bid Ruled Unconstitutional by Spanish Court. http://www.theguardian.com/world/2015/dec/02/catalonia-secession-unconstitutional-spanish-court. Accessed: 01 April 2016.

[57] Nataly Kelly. Why Machines Alone Cannot Solve the World's Translation Problem. http://www.huffingtonpost.com/nataly-kelly/why-machines-alone-cannot-translation_b_4570018.html, 2014. Huffington Post. Accessed: 08 December 2015.

[58] Nataly Kelly, Donald DePalma, and Vijayalaxmi Hegde. The Need for Translation in Africa: Addressing Information Inequality so that Africa May Prosper. http://www.commonsenseadvisory.com/Portals/0/downloads/Africa.pdf, 2012. Common Sense Advisory. Accessed: 08 December 2015.

[59] Suzanne Kemmer. A Brief History of English, with Chronology. http://www.ruf.rice.edu/~kemmer/Words04/history/, 2005. Accessed: 26 March 2016.

[60] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics, 2007.

[61] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics, 2003.

[62] Moshe Koppel and Noam Ordan. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326. Association for Computational Linguistics, 2011.

[63] Jhumpa Lahiri. Teach Yourself Italian. `http://www.newyorker.com/magazine/2015/12/07/teach-yourself-italian?Src=longreads`. Accessed: 07 December 2015.

[64] Angela Lavoipierre. Islamic State: ADF Twitter account used to fight propaganda sending nonsense Arabic tweets, expert says. `http://www.abc.net.au/news/2015-09-23/adf-sending-nonsense-arabic-tweets,-expert-says/6799874`. Accessed: 06 December 2015.

[65] Gennadi Lembersky. *The Effect of Translationese on Statistical Machine Translation*. PhD thesis, University of Haifa, 2013.

[66] Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Improving statistical machine translation by adapting translation models to translationese. *Computational Linguistics*, 39(4):999–1023, 2013.

[67] Philip E Lewis. The measure of translation effects. *Difference in translation*, pages 31–62, 1985.

[68] Alan Reed Libert. What can Pragmaticists Learn from Studying Artificial Languages? In *Perspectives on Linguistic Pragmatics*, pages 397–432. Springer, 2013.

[69] Harvey Lodish, Arnold Berk, S Lawrence Zipursky, Paul Matsudaira, David Baltimore, and James Darnell. Overview of neuron structure and function. 2000.

[70] Time Magazine. Autos: American's Moment of Truth. `http://content.time.com/time/magazine/article/0,9171,904440,00.html`. Published: 26 October 1970. Accessed: 31 March 2016.

[71] Dibarah Mahboob. Mutiny for the Sake of Language. `https://www.mtholyoke.edu/~mahbo22d/classweb/bengali_language_movement/BLM2.html`. Accessed: 01 April 2016.

[72] José B Mariño, Rafael E Banchs, Josep M Crego, Adriàa de Gispert, Patrik Lambert, José AR Fonollosa, and Marta R Costa-Jussà. N-gram-based Machine Translation. *Computational Linguistics*, 32(4):527–549, 2006.

[73] Tomas Mikolov. RNNLM Toolkit. `http://rnnlm.org/`. Accessed: 12 December 2015.

[74] Tomas Mikolov. Statistical Language Models Based on Neural Networks. `http://www.fit.vutbr.cz/~imikolov/rnnlm/thesis.pdf`. Accessed: 12 December 2015.

[75] Michael Mulé. Machine Translation: Ensuring Meaningful Access for Limited English Proficient Individuals. `https://www.dol.gov/oasam/programs/crc/062414Machine_TranslationWebinar.pdf`. Accessed: 06 December 2015.

[76] Preslav Nakov and Hwee Tou Ng. Improved statistical machine translation for resource-poor languages using related resource-rich languages. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, pages 1358–1367, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[77] Friedrich Neubarth, Barry Haddow, Adolfo Hernández Huerta, and Harald Trost. A hybrid approach to statistical machine translation between standard and dialectal varieties. `http://www.ofai.at/~friedrich.neubarth/papers/ltc2013-neubarth.pdf`. Accessed: 15 December 2015.

[78] Andrew Ng. CS229 Lecture Notes Part IV. `http://cs229.stanford.edu/notes/cs229-notes2.pdf`. Accessed: 01 April 2016.

[79] Eugene Nida. Principles of Correspondence. *The translation studies reader*, 2, 2000.

[80] University of Oslo. Biblioteca Polyglotta. `https://www2.hf.uio.no/polyglotta/index.php`. Accessed: 28-03-2015.

[81] University of Stanford Unsupervised Feature Learning and Deep Learning. Neural Networks. `http://ufldl.stanford.edu/wiki/index.php/Neural_Networks`. Accessed: 14 December 2015.

[82] Adama Ouane and Christine Glanz. *Why and How Africa Should Invest in African Languages and Multilingual Education: An Evidence-and Practice-Based Policy Advocacy Brief.* ERIC, 2010.

[83] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[84] Michael Paul, Hirofumi Yamamoto, Eiichiro Sumita, and Satoshi Nakamura. On the Importance of Pivot Language Selection for Statistical Machine Translation. In *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, page 221, 2009.

[85] Steven Pinker. *The Language Instinct: The New Science of Language and Mind*, volume 7529. Penguin UK, 1995.

[86] Barbara Pozzo and Valentina Jacometti. *Multilingualism and the harmonisation of European law*. Kluwer Law International, 2006.

[87] Phillip Resnik. Translationese. `http://languagelog.ldc.upenn.edu/nll/?p=3255`. Accessed: 01 May 2016.

[88] David Rotman. How Technology Is Destroying Jobs. `http://www.technologyreview.com/featuredstory/515926/how-technology-is-destroying-jobs/`, 2013. Accessed: 15 December 2015.

[89] Mark Saba. HUMAN TRANSLATION VS. MACHINE TRANSLATION. `http://www.internationalaffairs.org.au/australian_outlook/human-translation-vs-machine-translation/`. Accessed: 06 December 2015.

[90] Dahlia Sabry and Ibrahim Saleh. The Role Played By Qur'an Translations In Steering Public Opinion Against Islam In Non-Muslim Communities. http://www.islamicwritings.org/quran/language/the-role-played-by-quran-translations-in-steering-public-opinion-against-islam-in-non-muslim-communities/. Accessed: 23 March 2015.

[91] Annalisa Sandrelli. The Dubbing of Gay-themed TV Series in Italy: Corpus-based Evidence of Manipulation and Censorship. *Altre Modernità*, pages 124–143, 2016.

[92] Diana Santos. On grammatical translationese. In *Short papers presented at the tenth Scandinavian conference on computational linguistics*, pages 59–66, 1995.

[93] Snopes.com. Computer Mistranslations. `http://www.snopes.com/language/misxlate/machine.asp`. Accessed: 03 April 2016.

[94] Benjamin Snyder and Regina Barzilay. Climbing the Tower of Babel: Unsupervised Multilingual Learning. *In ICML*, 2010.

[95] Jon Solomon. Traduction, violence et intimité hétérolinguale. `http://eipcp.net/transversal/1107/solomon/fr`. Accessed: 03 April 2016.

[96] StatMT. Baseline system. `http://www.statmt.org/moses/?n=Moses.Baseline`. Accessed: May 1, 2015.

[97] George Steiner. The hermeneutic motion. *The translation studies reader*, pages 193–98, 2000.

[98] Christos Stergiou and Dimitrios Siganos. Neural Networks. `https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html`. Accessed: 13 December 2015.

[99] Straumanis, Kaija. Why Literature in Translation is SUPER SUPER IMPORTANT. `http://www.rochester.edu/College/translation/threepercent/index.php?id=8872`. Accessed: 29 April 2015.

[100] Jörg Tiedemann. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[101] Gideon Toury. In search of a theory of translation, 2006.

[102] Gideon Toury. *Descriptive Translation Studies and beyond: Revised edition*, volume 100. John Benjamins Publishing, 2012.

[103] Naama Twitto-Shmuel. *Improving Statistical Machine Translation by Automatic Identification of Translationese*. PhD thesis, University of Haifa, 2013.

[104] Masao Utiyama and Hitoshi Isahara. A Comparison of Pivot Methods for Phrase-Based Statistical Machine Translation., 2007.

[105] François Vaillancourt, Olivier Coche, Marc-Antoine Cadieux, and Jamie Lee Ronson. Official language policies of the canadian provinces. *Studies in Language Policies*, 2012.

[106] Andreas van Cranenburgh, Galit W Sassoon, and Raquel Fernández. Invented antonyms: Esperanto as a semantic lab. In *Proceedings of IATL*, volume 26, 2010.

[107] Lawrence Venuti. *The Scandals of Translation: Towards an Ethics of Difference*. Taylor & Francis US, 1998.

[108] Eduardo Javier Ruiz Vieytez. *Pluralidades latentes: minorías religiosas en el País Vasco*. Icaria, 2010.

[109] Emily Wight. What can we learn from efforts to save an ancient South American language? `http://www.theguardian.com/education/2014/nov/18/endangered-language-quechua-columbia-chile`. Accessed: 07 December 2015.

[110] Martin Williams. Tech is removing language barriers—but will jobs be lost in translation? `http://www.theguardian.com/education/2014/sep/19/tech-removing-language-barriers-jobs-lost-translation`, 2014. Accessed: 06 December 2015.

[111] Hua Wu and Haifeng Wang. Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21(3):165–181, 2007.

[112] Hua Wu and Haifeng Wang. Revisiting pivot language approach for machine translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 154–162. Association for Computational Linguistics, 2009.

[113] Kenji Yamada and Kevin Knight. A Syntax-based Statistical Translation Model. *Proceedings of the 39th Annual Meeting on Association for Computational Lingusitics*, pages 523–530, 2001.

[114] Leyzer Ludwik Zamenhof. Unua Libro. *Warsaw, July*, 26, 1887.

[115] Karen Zeigler and Steven A. Caramota. One in Five U.S. Residents Speaks Foreign Language at Home, Record 61.8 million. `http://cis.org/record-one-in-five-us-residents-speaks-language-other-than-english-at-home`. Published: October 2014. Accessed: 31 March 2016.