We need to talk about bias in machine translation

The Fairslator whitepaper

Michal Měchura

Version 1.0, published 28 February 2022

Fairslator

www.fairslator.com

0. **Introduction**

> "Languages differ essentially in what they *must* convey and not in what they *may* convey."
>
> — *Roman Jakobson, 1959*

Machine translation technology is getting better all the time. The translations produced by Google Translate, DeepL and others are now almost indistinguishable from human translations in terms of fluency. But the problem of **bias** still remains. Translations produced by machines are often biased because of ambiguities in gender, in forms of address, and in word meaning. For example, when translating from English into French, should *student* be translated as male *étudiant* or female *étudiante?* Should *you* be translated as informal *tu* or formal *vous?* How is a machine supposed to know which translation to pick?

Modern machine translation, which is mostly statistical and neural, effectively performs subsymbolic **word sense disambiguation** (WSD) from clues in the text. But what if there *are* no clues? What if there is nothing in the text to indicate that the student is male or female or that the *you* person should be addressed formally or informally? In such a situation the machine typically picks whichever translation is statistically more likely, based on what it has seen more frequently in its training data. In other words, the machine makes an **assumption**. And assumptions lead to bias, for example when the machine assumes we're talking about a male student when in fact we're talking about a female one.

Bias has long been ignored in the machine translation industry, but awareness of it is growing now and attempts to solve it are springing up. Most of the research and debate around bias in machine translation concerns **gender bias**, like in our *student* example, because gender is a controversial topic in our civilisation overall. But there are other kinds of bias in addition to this, such as bias in forms of address (*you*). These types of bias are perhaps less personally offensive than gender but are nonetheless harmful: they

cause translations to be incorrect (where *incorrect* means *different from what the user meant*) and they drag down the overall usefulness and trustworthiness of machine translation.

In this whitepaper I will look at the types of bias that commonly occur in machine translation from English into other languages, with a focus on French, German, Russian, Czech and Irish. I will show how bias always arises in machine translation when there is an **unresolvable ambiguity** in the source text, which causes the machine to make **unjustified assumptions** as to what the user may have meant. That will be the subject of the first and second sections of the whitepaper. In the third section I will review how existing machine translators such as Google Translate and DeepL are trying to solve the problem. Finally, in the fourth section, I will explain how the problem is being tackled by Fairslator, an experimental application which removes bias by examining the output of a machine translator, detecting when bias has occurred, and correcting it by asking the user follow-up questions such as *Do you mean male student or female student? Are you addressing the person casually or politely?*

The purpose of this whitepaper is twofold. First, I want to describe and define the problem of machine translation bias as fully and usefully as possible. Second, I want to convince you that Fairslator's method for solving the problem, which we can briefly summarize as **rule-based post-processing with humans in the loop**, is the most effective and reliable method invented yet.

1. Types of bias in machine translation

Let's have a good look now at the types of bias that frequently occur in machine translation from English into other languages. I will discuss each type of bias in detail and then summarize what they have in common: that they are all caused by unresolvable ambiguities.

1.1. Gender bias on nouns

Many European languages are more "gendered" than English. By this I mean not the fact that they have *grammatical gender*, but the fact that they encode the *biological gender* of people more often than English.

Grammatical gender is when nouns which do *not* refer to people are treated arbitrarily as masculine or feminine by the grammar of the language. For example, in German *Brücke* 'bridge' is feminine and *Fluss* 'river' is masculine, while in French it's the other way around, *pont* 'bridge' is masculine and *rivière* 'river' is feminine. English doesn't have grammatical gender so, from the point of view of an English speaker, the presence of grammatical gender in other languages is definitely a complication. But – and this is important – grammatical gender does *not* cause any biases in translation. So, grammatical gender is irrelevant for us here. In the rest of this whitepaper when we say *gender* we mean *biological gender*.

Biological gender is the real-world gender (or sex) of people, human beings. All languages have nouns that refer to people such as *girl, student, driver, father, actress, writer*. Some of these have biological gender "baked" into them (*girl, father, actress*) and others don't (*student, driver, writer*). The gendered ones can only be used to refer to people of that gender – you can't call a male actor an *actress* – while the genderless ones can be used for a person of any gender: *student* can refer to a male student and a female student.

Languages differ in which nouns are gendered and which not. English has just a handful of gendered nouns, while other European languages typically have more. It is often the

case that a gender-neutral noun in English has two gendered equivalents in another language. These are typically words for occupations.

English: *student*

- French: *étudiant* (male), *étudiante* (female)
- German: *Student* (male), *Studentin* (female)
- Russian: *студент* (male), *студентка* (female)
- Czech: *student* (male), *studentka* (female)

English: *director*

- French: *directeur* (male), *directrice* (female)
- German: *Direktor* (male), *Direktorin* (female)
- Russian: *директор* (male), *директориа* (female)
- Czech: *ředitel* (male), *ředitelka* (female)

When translating a sentence such as *I am a student* into one of these languages, the translator (human or machine) needs to know whether *I* refers to a man or a woman. A human translator, if this information is not known to him or her, will look for clues elsewhere in the text (if the sentence is part of a larger text) or simply ask somebody. Machine translators, on the other hand, usually translate each sentence individually, so there is no "elsewhere in the text" for them to look. And they don't ask follow-up questions either. Usually, a machine translator will decide based on what is statistically more likely, given what it has seen in the training data. If it has seen *student* translated into the male version more often than the female version, then it will assume the male reading is intended, and vice versa. For example, here is how the sentence is translated by DeepL as of this writing:

English: *I am a student.*

- French: *Je suis un étudiant.* (male)
- German: *Ich bin ein Student.* (male)
- Russian: *Я студентка.* (female)
- Czech: *Jsem student.* (male)

The translations are grammatically correct[1] but the problem is that each assumes a specific gender for the student. When a human user types a sentence like this into a

[1] With possibly the exception of French and German, where – depending on whether or not the sentence is intended as an answer to a question like *what do you do for a living?* – it would be more idiomatic to leave out the indefinite articles *un* and *ein*. But that is a different issue and beside the point here.

machine translator, they probably know whether the pronoun *I* refers to a man or a woman. The machine does not know this, though, and instead it makes a tacit assumption. If the human user doesn't speak the target language and isn't aware that the target language has gendered nouns for *student*, then he or she will walk away with what may be an embarrassingly wrong translation.

Notice that this is different from sentences such as *She is a student* or *He is a student*. Here it *is* possible to infer from the rest of the sentence whether *student* is male or female. The presence of the gendered pronouns *he* and *she* should be enough to tip any well-trained machine translator towards the correct gender for *student*. And indeed, all major machine translation engines do translate these sentences with the correct gender and there is no gender bias. Gender bias occurs only when there is an **unresolvable gender ambiguity** in the source text.

1.2. Gender bias on adjectives

Gendered nouns such as *student* and *director* are a popular target for people interested in gender bias in machine translation. But nouns are not the only word class that can be gendered. In Romance languages (such as French and Italian) and in Slavic languages (such as Russian, Polish an Czech) the set of gendered words includes **adjectives** (words such as *happy, young, tall*).

English: *happy*

- French: *heureux* (male), *heureuse* (female)
- Russian: *счастлив* (male), *счастлива* (female)
- Czech: *šťastný* (male), *šťastná* (female)

If you have a sentence such as *I am happy* or *are you happy?* where the adjective appears in **predicative position** (that is, *after* the verb) and if you want to translate this sentence into a language that has gendered adjectives, the adjective needs to **agree** with the subject (*I, you*) in gender. In other words, you need to know whether the subject refers to a man or a woman.

English: *I* (male) *am happy.*

- French: *Je suis heureux.*
- Russian: *Я счастлив.*
- Czech: *Jsem šťastný.*

English: *I* (female) *am happy.*

- French: *Je suis heureuse.*
- Russian: *Я счастлива.*
- Czech: *Jsem šťastná.*

A machine translator typically doesn't know whether the subject is male or female because there are no clues in the English sentence to tell one way or the other: there is an **unresolvable ambiguity** there. And what does a machine translator do in the face of an unresolvable ambiguity? It *assumes* the subject's gender, based on whatever is statistically more likely. In the sentence *I am happy*, all major machine translators happen to assume the male gender – they translate it as *I* (male) *am happy* – which is of course an unjustified assumption approximately half the time, given that approximately half of all people are women.

As if this were not enough, adjectives in some of these languages are gendered not only in the singular (when they refer to one person, such as *I*) but also in the plural (when they refer to a group of people, such as *we*). French and Czech are such languages.

English: *happy* (plural)

- French: *heureux* (males), *heureuses* (females)
- Czech: *šťastní* (males), *šťastné* (females)

In languages where plural adjectives are gendered, the rules are usually as follows:

- When the subject refers to a group of men or to a mixed-gender group containing at least one man, then the adjective must be in the plural *male* form (French *heureux*, Czech *šťastní*).[2]
- When the subject refers to a group of women, then the adjective must be be in the plural *female* form (French *heureuses*, Czech *šťastné*).

So, in a sentence like *We are happy* the options are that we refer **either** to a group containing at least one man **or** to a group of women, and the adjective has to agree with that in gender.

English: *We* (a group with at least one male) *are happy.*

- French: *Nous sommes heureux.*
- Czech: *Jsme šťastní.*

---

[2] This is an example of something called the *male default*, a phenomenon we will return to in a later subsection.

English: *We* (a group of women) *are happy*.

- French: *Nous sommes heureuses.*
- Czech: *Jsme šťastné.*

And once again we have potential for bias here when the machine translator encounters a sentence with no clues as to the subjects' genders: is *we* a mixed-gender group, or a group of women? There is an **unresolvable ambiguity**, the machine translator is forced to assume gender based on what it has seen more often in the training data, and the result is gender bias. Most major machine translators assume the at-least-one-male reading in these situations, which is of course bound to be incorrect some of the time.

Notice that this is different from sentences such as the *The girls are happy.* Here, the gender ambiguity of the adjective *happy* is resolvable from the clue *the girls*. Most machine translators do indeed pick up on that clue and go correctly for the female translation of *happy*, so there is no bias. Gender bias occurs only when the gender ambiguity is *unresolvable*.

1.3. Gender bias on past participles of verbs

Now that we have seen how gender bias can be caused by gendered nouns and adjectives, it is time to look at one last word class which is often gendered in European languages: **past participles of verbs**. The past participle of a verb is a word such as *seen, gone, been*. Such words are used in many languages to construct sentences in the past tense such as *I have seen, I have gone, I have been*.

Past participles are similar to adjectives. In languages where adjectives are gendered, past participles are often gendered too. This is the case for *some* past participles in Romance languages (French, Italian) and *all* past participles in Slavic languages (Russian, Polish, Czech).

English: *gone*

- French: *allé* (male), *allée* (female)
- Russian: *пошел* (male), *пошла* (female)
- Czech: *šel* (male), *šla* (female)

In these languages, whenever a sentence contains a past participle in the **predicative position** (that is, *after* the verb), then the past participle needs to agree with the subject in gender, just like adjectives do.

English: *I* (male) *have gone there.*

- French: *J'y suis allé.*
- Russian: *Я пошел туда.*
- Czech: *Šel jsem tam.*

English: *I* (female) *have gone there.*

- French: *J'y suis allée.*
- Russian: *Я пошла туда.*
- Czech: *Šla jsem tam.*

As we have seen with adjectives, some languages have their past participles gendered in the plural too.

English: *gone* (plural)

- French: *allés* (males), *allées* (females)
- Czech: *šli* (males), *šly* (females)

This means that when the subject is plural, the past participle needs to agree with it in gender.

English: *We* (a group containing at least one man) *have gone there.*

- French: *Nous y sommes allés.*
- Czech: *Šli jsem tam.*

English: *We* (a group of women) *have gone there.*

- French: *Nous y sommes allées.*
- Czech: *Šly jsme tam.*

Once again we have a lot of potential for gender bias here because the English sentence contains no indication of whether *I* is male or female, whether *we* is a mixed-gender group or a women-only group. This **unresolvable ambiguity** causes the machine translator to assume a gender for the subject based on whatever is statistically more likely from the training data. What do all the major machine translators do? They tend to go for the male reading, which is an unjustified assumption.

1.4. Interlude: gender-neutral language

You may be thinking at this point that languages other than English are so dominated by gender that it is practically impossible to say anything about anyone without revealing

what sex they are. There is some truth in this. In English, you can say a lot without giving off any hints as to your gender or your interlocutor's gender. In other languages this is less easy, in come cases even impossible.

Until very recently ago, if you wanted to speak in a gender-neutral way in most European languages, the device for doing that was the *male default*: using the male gender on the tacit understanding that it can refer to women too. For example, when talking about a doctor where you don't know (or don't want to say) what sex the doctor is, you might just use the male *Arzt* in German or the male *lékař* in Czech and it would be implicitly understood that this can refer to a female *Ärztin* or *lékařka* too. And if you are talking about a group of people where some are women and some are men, in many languages it is normal to refer to them in the male plural (*Ärzte*, *lékaři*) while the female doctors in the group, the *Ärztinnen* and *lékařky*, know that this covers them too.

In recent years, however, the male default has been falling out of favour. People have come to suspect that the male default creates in people's minds the subconscious conviction that some professions, some social roles, belong more to men than women. And, to be sure, a small subset of social roles have a *female default* instead in words for occupations such as cleaner or caregiver or cook. The argument goes that a society which believes in equal opportunities should not, through its language, construct a situation where certain professions belong to a particular sex, however subtly or subconsciously.

This brings us back to square one: if you don't want to use the male default, how *do* you refer to people in gender-neutral ways? In some languages, notably German, people have been experimenting with a metalinguistic notation where a male word and a female word are merged into a single gender-neutral one, for example male *Ärzte* + female *Ärztinnen* = neutral *Ärzt*innen*. The asterisk, called a *Gendersternchen* 'gender starlet' in German, is a somewhat controversial recent innovation. Another strategy we sometimes see, in many languages including Czech, is to simply write the male and female form side-by-side with a forward-slash in between, for example *šel/šla jsem tam* 'I (male/female) went there'.

For us who are interested in machine translation, it is important to realize that most of the training data from which machine translators have learned consists of texts we have inherited from the past: texts where the male default (or the female default, in a minority of cases) is all over the place. That is why our machine translators are still so biased in favour of male readings (or female readings, in some cases) today even though society has moved on.

Well, enough about gender now. So far we have looked at three types of phenomena that can cause gender bias in machine translation: gendered nouns, gendered adjectives, and gendered past participles. With this we have exhausted the topic of gender bias and it is time to turn our attention to other kinds of bias that occur frequently in machine translation from English. We'll start with the highly ambiguous English pronoun *you*.

1.5. The *you* problem: singular or plural?

English is unusual among European languages in that the pronoun *you* can be either singular (one person) or plural (several people), and context disambiguates which reading is intended. Most other European languages have one word for the singular and another for the plural. One such language is Irish which has *tú* for the singular and *sibh* for the plural.[3]

English: *You* (one person) *are here.*

- Irish: *Tá tú anseo.*

English: *You* (several people) *are here.*

- Irish: *Tá sibh anseo.*

The English sentence *you are here* contains no clues to help a machine translator figure out which of the two readings is intended, singular or plural. A (good) human translator would most likely resolve the ambiguity by asking someone what they meant, but machine translators don't have the ability to do that. A typical machine translator makes an assumption instead, based on what it has seen more frequently in its training data. And, as experience shows, all the major machine translators assume the singular reading of *you* in most cases.

This is an example of bias in machine translation which carries less potential for personal offence than gender bias, but it is bias all the same. Bias is not just something that causes offence or injustice. In technical jargon, *bias* is the tendency of any automated system to make unjustified assumptions, regardless of how severe or mild the social and personal consequences are.

---

[3] I am using Irish as the language for examples in this subsection because it is the only language I know where the singular/plural distinction on *you* is not complicated by further distinctions of formality and politeness (which are going to be discussed in the following subsection). Irish has a singular *you* versus a plural *you*, but it does not have a formal *you* versus an informal *you*.

1.6. The *you* problem: casual or polite?

One thing which is very common in European languages when addressing others with second-person pronouns is that you have a choice between two levels of formality: a *casual* level and a *polite* level. For example, French has *tu* for the casual level and *vous* for the polite level, German has *du* and *Sie*, Czech has *ty* and *vy*, Russian has *мы* and *Вы*. This distinction does not exist in English.[4]

A further complication is that the casual/polite dimension overlaps with the singular/plural dimension in complicated ways. For example, German distinguishes between singular and plural only at the casual level, while the polite level is ambiguous for number.

German	singular *you as one person*	plural *you as several people*
casual *you addressed informally*	*du*	*ihr*
polite *you addressed formally*	*Sie*	*Sie*

So, when translating from English into German, there are three possible interpretations of *you*.

English: *You* (one person addressed casually) *are here.*

- German: *Du bist hier.*

English: *You* (several people addressed casually) *are here.*

- German: *Ihr seid hier.*

English: *You* (one or several people addressed politely) *are here.*

- German: *Sie sind hier.*

In Czech the situation is slightly different: the pronoun *vy* is highly ambiguous as it can be **either** plural at any level of formality **or** singular at the polite level, while *ty* can only be singular casual.

---

[4] It did exist in previous stages of the historial evolution of English where the casual pronoun was *thou* and *you* was the polite one.

Czech	singular *you as one person*	plural *you as several people*
casual *you addressed informally*	*ty*	*vy*
polite *you addressed formally*	*vy*	*vy*

A similar layout as in Czech can be found in other Slavic languages and also in Romance languages such as French. When translating from English into these languages, there are two possible readings of *you*.

English: *You* (one person addressed casually) *are here.*

- Czech: *Ty jseš tady.*
- French: *Tu es ici.*

English: *You* (one person addressed politely, or several people) *are here.*

- Czech: *Vy jste tady.*
- French: *Vous êtes ici.*

Last but not least, if an adjective or a verbal participle is present in the sentence, then the options are multiplied by the dimension of gender and we get a total of *six* possible readings, three male and three female.

Male readings

English: *You* (a man addressed casually) *are happy.*

- Czech: *Ty jseš šťastný.*
- French: *Tu es heureux.*

English: *You* (a man addressed politely) *are happy.*

- Czech: *Vy jste šťastný.*
- French: *Vous êtes heureux.*

English: *You* (a group containing at least one man) *are happy.*

- Czech: *Vy jste šťastní.*
- French: *Vous êtes heureux.*

Female readings

English: *You* (a woman addressed casually) *are happy*.

- Czech: *Ty jseš šťastná*.
- French: *Tu es heureuse*.

English: *You* (a woman addressed politely) *are happy*.

- Czech: *Vy jste šťastná*.
- French: *Vous êtes heureuse*.

English: *You* (a group of women) *are happy*.

- Czech: *Vy jste šťastné*.
- French: *Vous êtes heureuses*.

One of these readings is what the user intended when he or she typed the English sentence into a machine translator. Unfortuntely, there is no way for a machine to know which one, as there are no clues in the sentence itself. Most machine translators make an unjustified assumption here and select whichever reading came up more often in their training data, a guess which is certain to be wrong some of the time. This leads to people being given translations which carry a potential for misunderstanding and offence because, in languages where levels of formality are distinguished, addressing someone with an inappropriate level of formality is considered either disrespectful (being overly casual) or distanced (being over-formal).

Summary

When we investigate bias in machine translation from English into other languages, we observe that bias does not occur randomly or arbitrarily. Bias occurs systematically in the following three categories.

- Gender bias on nouns that refer to people, on adjectives, and on the past participles of verbs. English usually has one word for both genders while in other languages often have two words, one male and one female.
- Number bias on second-person pronouns. English has only one pronoun here (*you*) while other languages often have two, a singular *you* and a plural *you*.
- Formality bias on second-person pronouns. Again, English has only one pronoun here (*you*) while other languages often have two, a “casual” *you* and a “polite” *you*.

Of these three types of systematic bias, only a subset of the first type is well known and has been studied extensively: gender bias on nouns that refer to people. The other types have been studied less well and are not so notorious, but they are real nonetheless.

What all these types of bias have in common is that they are caused by unresolvable ambiguities in the source text: the correct translation depends on what the human user meant, but the machine does not know (and cannot know) what it was.

2. Machine translation bias in a broader context

The previous section looked at the low-level mechanics of how and when bias-causing ambiguities arise in translation from English into several other languages. In this section we'll look at bias again but this time from a wider angle. We'll look at the pairs of languages where bias is likely to happen, and we'll ask ourselves whether and why it's wrong to allow bias to happen.

2.1. High-detail languages versus low-detail languages

Not all languages insist on the same level of detail when encoding meaning into words. Some languages insist on encoding details which other languages gleefully gloss over. For example, German and Czech insist on encoding gender into nouns that refer to humans by occupation. The consequence is that you simply can't say certain things in these languages without simultaneously revealing the gender of the people you are talking about. English, on the other hand, does not encode gender in such nouns,[5] so in English it is possible to say many things without revealing anybody's gender.

The same goes for other features such as number (singular or plural) and formality (casual or polite). In some languages you can't say much about anyone without pulling these things into the equation, while in others you can. As the linguist Roman Jakobson famously wrote in 1959, "languages differ essentially in what they *must* convey and not in what they *may* convey".[6] Along almost any dimension of meaning, each language can be classified as either *high-detail* (insists on encoding certain details) or *low-detail* (does not insist on encoding much detail).

---

[5] Apart from a small number of exceptions such as *actress*.

[6] Roman Jakobson (1959) 'On Linguistic Aspects of Translation', an essay in the book *On Translation* edited by Reuben Arthur Brower.

2.2. Bias when translating from and into English

When we compare English to other European languages along almost any dimension of meaning, we observe that English is almost always at the low-detail end of the scale while the other languages are high-detail:

- English is a low-detail language on gender in nouns referring to humans by occupation, and also on gender in adjectives and verbal participles in predicative position, while German, French, Czech and Russian are high-detail languages.
- English is a low-detail language on number in second-person pronouns (*you* is both singular and plural), while Irish, German, French, Czech and Russian and high-detail languages.
- English is a low-detail language on register in second-person pronouns (*you* is ambiguous as to formality) while German, French, Czech and Russian are high-detail languages.[7]

In a way, English is an outlier among European languages: most other languages insist on more detail than English. This means that situations of ambiguity and bias are more likely to arise when translating *from* English than *into* English.

That's not to say that bias cannot happen in translations *into* English at all. The interesting thing about English is that even though it is a low-detail language for gender on first-person and second-person pronouns (*I*, *we*, *you*), it is a high-detail language on third-person pronouns in the singular (*he*, *she*). On the other hand, there are languages which are low-detail here, either because they have gender-neutral pronouns in the third person singular (eg. Finnish *hän* which means both *he* and *she*[8]) or because they sometimes "drop" (= leave out, omit) pronouns. Interestingly, Czech is exactly such a "pro-drop" language. A sentence such as *chce přijít* 'wants to come', which is perfectly valid and grammatical in Czech, can be an abbreviated version of either *on chce přijít* 'he wants to come' or *ona chce přijít* 'she wants to come'. The sentence is ambiguous and, when translating it into a non-"pro-drop" language such as English, the ambiguity can lead to a biased translation.

---

[7] An interesting in-between case is Irish. Irish is a low-detail language on gender and formality (like English), but a high-detail language on number in second-person pronouns (like most European languages).

[8] This pronoun has been in the language for centuries, it is not a recent innovation.

2.3. Bias when translating through English as a pivot language

In theory, biased translations can occur between any two languages when the source language is *lower-detailed* than the target language along some dimension of meaning. A special case is when we have not two but *three* languages in the equation: a source language, a target language, and an intermediate language *through* which the translation happens. For example, Google Translate seems to use English a lot as an intermediate pivot for translation between other languages. A similar situation is known to exist in the European Parliament where live interpretation between two languages is sometimes channelled through a third pivot language, often English.

If the source and target language are both high-detail but the pivot is low-detail, then information about gender, number and formality can become distorted along the way. From the author's experience, this occurs almost *always* when translating between German and Czech with Google Translate. Google Translate translates the German polite-*you* sentence *was haben Sie getan?* 'what have you done?' into the Czech casual-*you* sentence *co jsi udělal?* This is probably because, behind the scenes, the sentence is translated into English first and then from English into Czech, and the latter step is laden with a biased reading of *you*.

2.4. Unresolvable ambiguities lead to unjustified assumptions

Whenever you are translating something from a low-detail language into a high-detail language, ambiguity arises. For example, the English sentence *I am a student* is ambiguous (from the point of view of a Czech speaker) because it doesn't say what gender the student is. To translate the sentence successfully into Czech, we need to disambiguate: to recover the student's gender from somewhere.

Sometimes, a sentence contains clues that help with disambiguation. For example, the English sentence *I gave the student her homework* contains a clue (the possessive pronoun *her*) from which we can infer that the student is female. Human translators perform these inferences quickly and effortlessly, while machine translators are getting better all the time at mimicking them. It is imaginable that, with time, machine translation technology will become as skilled as humans at inferring such details from context.

On the other hand, a sentence sometimes contain no clues at all, and disambiguation is therefore impossible. A sentence such as *I am a student* taken as it is, without any further text before or after it, is hopelessly ambiguous as to the student's gender. There are no

clues in it to decide one way or another. A human translator might still be able to figure it out by examining the extra-linguistic reality, such as simply by looking at the person who has uttered it. But machine translators can't look at anything: all they have is the one sentence.

When faced with an unresolvable ambiguity, a machine translator typically selects whichever option it believes is more probable, based on what it has seen in its training data. In other words, it makes an assumption. But the assumption is unjustified.

2.5. Unjustified assumptions lead to incorrect translations

Let's discuss why the assumption is unjustified and why allowing the machine to make it is a bad idea. The assumption that *student* should be translated as 'male student' is only very poorly justified by the fact that *student* was translated as 'male student' 75% of the time in the machine translator's training data. That is not really a justification, it is merely a *prediction* of what the intended meaning may be this time. This prediction will only be correct about 75% of the time. The other 25% of the time, the user will receive a translation which does not express what he or she actually meant.

Very often, when a human user asks a machine translator to translate a sentence with *I* as a subject, they have themselves in mind: they want to say something about themselves in the target language. If the user is a woman and the translation is worded as if a man is saying it, then we have not given the user what she was looking for. Now, if the user understands the target language well enough to spot and correct the error, then all is well – except perhaps the inconvenience of having to post-edit translations all the time when you're woman and not having to post-edit at all when you're a man. But if the user is not as well informed as that, if he or she doesn't even *know* that the target language is high-detail along this dimension, then the damage is double: the user has been given the wrong translation *and* she doesn't know it.

You may object that this is normal in machine translation, that some percentage of the output is expected to be inaccurate, that users should not be naive and that they should never trust machine translation unconditionally if they don't speak the target language. This is true in principle, but this principle isn't stopping the machine translation industry from *innovating*, from looking for ways to improve the quality of translations. Bias is one area where quality *can* be improved. The problem of bias *is* solvable, as we will see later in this whitepaper.

You may also object that artificial intelligence is getting better all the time and that errors like this are bound to be eliminated in the future as machine learning technology develops. But the bias-triggering ambiguities we are discussing in this whitepaper cannot be resolved by improving existing AI techniques: they are *unresolvable*. No AI, however smart, can ever know whether the human user means a singular *you* or a plural *you* if there are no clues in the text. Not even humans can do that – at least not without asking follow-up questions. So, instead of looking for a solution where it can't be found (ie. by building better AI), I am proposing a method inspired by what humans do: detecting unresolvable ambiguities when they occur, and then asking follow-up questions to resolve them.

2.6. Unjustified assumptions perpetuate stereotypes

Allowing bias to happen during machine translation has implications beyond the individual who is sitting in front of the computer and interacting with a machine translator. Bias has social implications too.

What machine translation does is, it makes translation services more easily and cheaply available. The result is that there are more biased translations in the world than there would be if machine translation didn't exist. It is one thing when one human translator produces one biased translation on one ocassion because he or she didn't bother to ask what gender someone is. It is quite another thing when a software tool machine-learns from biased data and then replicates the bias a thousand times over. Bias has always existed in texts written by humans, but automation amplifies it.

And that is not the end it. Machine-learned AI does not simply produce biased output at the same percentage as humans would. Most machine learning algorithms tend to *overgeneralize*, to favour typicality and to disfavour non-typicality. The result is that machine-learned software tends to be even more biased than the training data. In machine translation, this means that machine translators produce biased translations more often that a careless human translator would.

With this in mind, it is no wonder that pressure is growing on the machine translation industry to "fix" this problem. In the next section we will look at a few recent attempts in this direction from major player such as Google and DeepL, before we go on to introduce Fairslator in the final section of this whitepaper.

Summary

Unresolvable ambiguites occur in machine translation when the source language is *low-detail* and the target langage is *high-detail* along some dimension of meaning such as gender, number or formality. As English happens to be *low-detail* on all of these aspects, translation *from* English (as opposed to *into* English) into other languages is especially prone to unresolvable ambiguties.

An unresolvable ambiguity arises when the source text contains no clues to help a machine disambiguate. Unresolvable ambiguities cause machines to make unjustified assumptions about what the user may have meant, which causes biased translations.

3. Existing attempts to solve bias

Most of existing attempts to "fix" bias in machine translation have concentrated on one specific subtype: the bias that occurs on nouns referring to humans by occupation such as *teacher*, *student*, *director* and *doctor*. Most of that work is happening in academia rather rather than in industry. A good summary of the state of the art in academic research into gender bias in machine translation can be found in the following paper.

> Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, Marco Turchi (2021) **'Gender Bias in Machine Translation'** in *Transactions of the Association for Computational Linguistics,* volume 9, pages 845–874, https://doi.org/10.1162/tacl_a_00401

Most of that research has not found its way into publicly visible applications yet. In this section we will look at a handful that has: we will review two attempts to "handle" bias in some way, one by Google and one by DeepL. This will prepare the ground for when we introduce Fairslator in the next section after that.

3.1. Google's approach to gender

Google Translate reduces gender bias in some language pairs by offering two translations instead one when the source-language input contains a gender ambiguity. **Figure 1** shows what it looks like for the end-user. Google has described how it has achieved this in a series of blog posts.

> December 2018: **'Reducing gender bias in Google Translate'** by James Kuczmarski, https://blog.google/products/translate/reducing-gender-bias-google-translate/
>
> December 2018: **'Providing Gender-Specific Translations in Google Translate'** by Melvin Johnson, https://ai.googleblog.com/2018/12/providing-gender-specific-translations.html
>
> April 2020: **'A Scalable Approach to Reducing Gender Bias in Google Translate'** by Melvin Johnson, https://ai.googleblog.com/2020/04/a-scalable-approach-to-reducing-gender.html

The dual translations are triggered by the presence of gender-ambiguous nouns such as *teacher* when translating from English into Spanish, and by the presence of gender-neutral third-person pronouns (equivalents of *he* and *she*) when translating from a

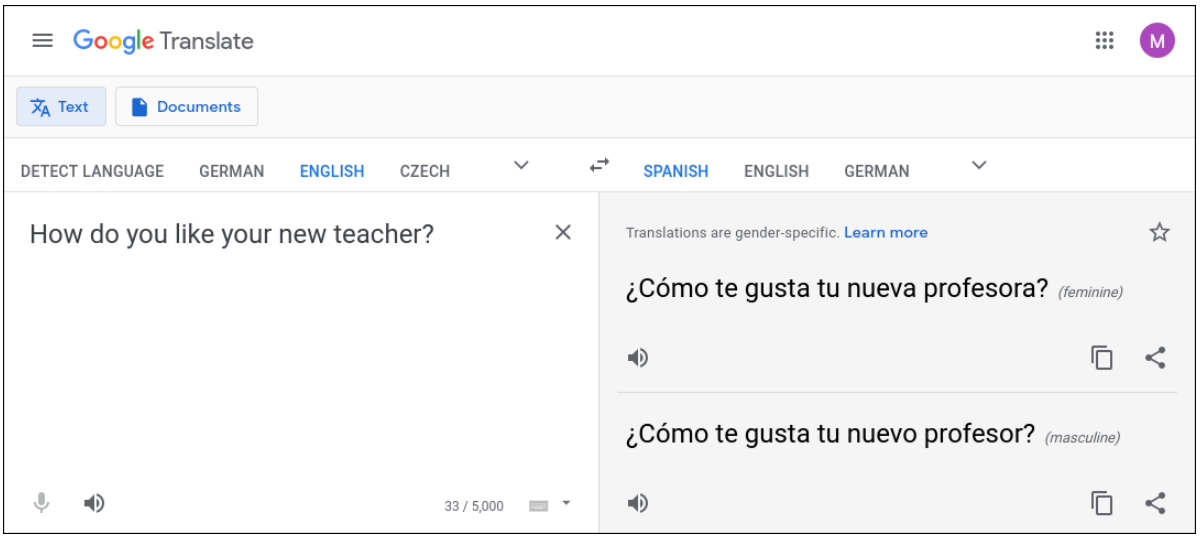

Figure 1. *Google offers "feminine" and "masculine" translations side-by-side in some language pairs.*

handful of languages (such as Finnish, Hungarian and Persian) into English. The scope of this innovation is far from broad: it only works for a handful of language pairs, and it only handles a *subset* of bias-causing ambiguities. Systematic gender-related variation on adjectives and participles, as described earlier in this whitepaper, are not covered, and neither are other kinds of bias such as the various "*you* problems". In fact, *most* types of bias in *most* language pairs are not covered.

The translations are given side-by-side and labelled as "feminine" and "masculine". It is questionable whether it is always clear to all users what this means. In a sentence such as *how do you like your new teacher*, is it clear that "feminine" means "use this translation if you mean a female teacher"? A linguistically naive user might not understand it that way. For somebody who does not speak or read the target language at all, it is not obvious *who* the "feminine" label applies to (the *teacher*, the *you*, or perhaps me the speaker?) or *what* it means (that the person is a woman).

We can conclude that Google's approach to machine translation bias, while being a step in the right direction, does not yet have the broad coverage it would need to claim the problem as "solved", and has some usability issues.

3.2. DeepL's approach to forms of address

DeepL is (so far) ignoring gender bias but has developed a solution for solving ambiguities related to formality in forms of address on second-person pronouns (equivalents of English *you*). In some language pairs, DeepL offers a menu on screen – see **Figure 2** – where the user can choose between "formal tone" and "informal tone". This affects which second-person pronouns appear in the translation (for example the choice between *du* and *Sie* in German) as well as words that depend on them grammatically, such as verbs.

This is obviously a useful feature when translating from English, a language which does not encode formality in its one and only second-person pronoun *you*. But the complication is that in many languages that do encode formality in their second-person pronouns, they encode *number* in them as well (singular or plural) and these two dimensions intermix in complex ways. For example, German has *du* versus *Sie* in the singular and *ihr* versus *Sie* in the plural, while Czech has *ty* versus *vy* in the singular but only *vy* in the plural. So, when translating from English into other languages with DeepL, you can fine-tune your translation to address people politely or casually, but you cannot choose between addressing them as one person or as many.

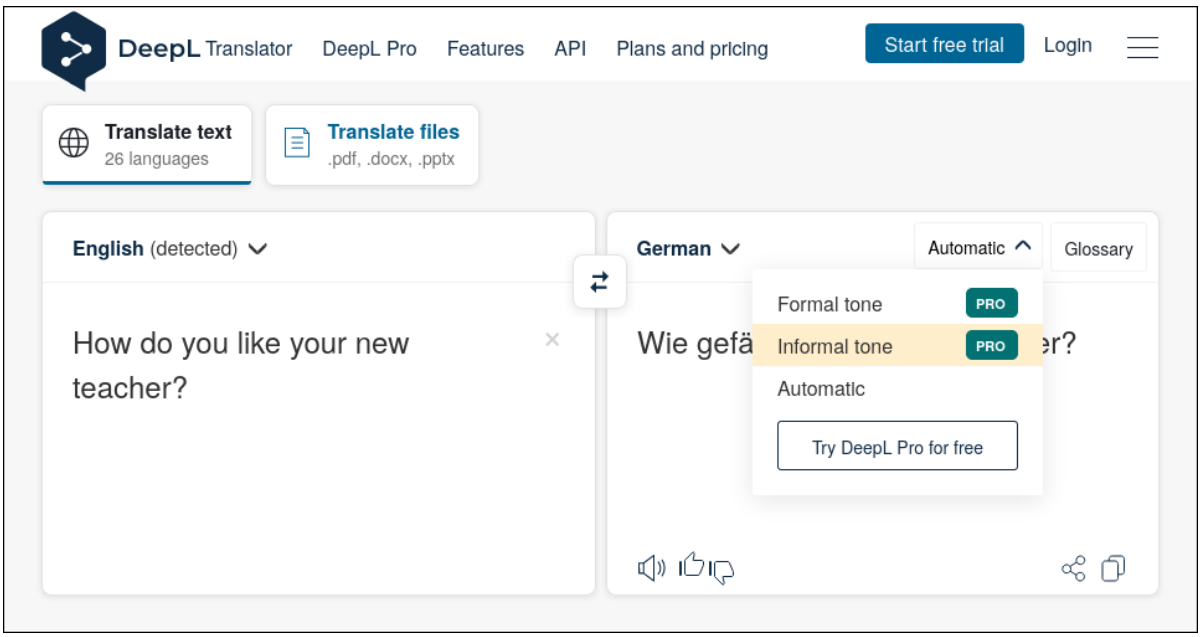

Figure 2. *DeepL lets you choose between "formal tone" and "informal tone" when translating into some languages, including German.*

The formal/informal menu is sometimes available in DeepL when translating from languages other than English too. For example, when translating from French into German, the option to switch between formal and informal tone is available, even when the French source text is *not* ambiguous in this respect. Remember that both German and French and high-detail languages when it comes to formality, so a sentence in French typically contains enough clues for the machine translator to know which level of formality is intended. When translating *où es-tu?* 'where are you?' from French into German, DeepL's machine translator correctly translates this with the casual pronoun: *wo bist du*. But it also provides the formal/informal menu on screen, which is unnecessary, and making a choice there has no effect.

In summary, DeepL is making a serious attempt to remove ambiguity and bias from how their machine translator handles formality on second-person pronouns. But the category of formality is closely linked with the category of number in many languages, and DeepL only has a solution for the former but not for the latter. Also, like Google, DeepL's solution is only available in some language pairs and not in others, leaves all other kinds of bias unsolved, and has some user-experience issues.

Summary

Beyond academic research, commercial machine translation providers are beginning to show interest in the challenge of bias, and attempts at "fixing" it are beginning to appear in their publicly visible applications. But these attempts are limited in scope and ambition, and have usability issues. The machine translation industry clearly *wants* to solve the problem but it's early days yet.

4. Introducing Fairslator

Fairslator is an experimental application which detects and corrects bias – including but not limited to gender bias – in machine translation. But the first thing you need to know about Fairslator is that it is *not* a machine translation tool. Rather, it is a plug-in for other machine translation tools. Fairslator treats machine translation as a black box and only examines its output: scans it for possible occurrences of bias and, if it detects any, offers options for correcting it.

> Visit the online demo and try Fairslator for yourself.
> https://www.fairslator.com

So, what exactly happens when you ask Fairslator to translate something? First you need to select the machine translation service which you would like to use to get the translation: currently the options are DeepL, Google Translate and Microsoft Translator, but in principle Fairslator can work with any translation service. Fairslator talks to the translation service behind the scenes and obtains a translation. After that, Fairslator inspects the two texts (the original plus the translation) and tries to find any unresolvable ambiguities: any places where there might be a choice between male and female readings, between singular versus plural readings of *you*, and so on. If there aren't any, then Fairslator simply shows you the translation and we're done. But if there are some, then Fairslator shows you not only the translation but also a menu of choices where you can narrow down your meaning: whether it is a man or a woman saying it, whether you are talking to one person or several, and so on. As you make choices in these menus, Fairslator alters the translation to reflect that. This way you can gradually adjust the translation to match what *you* meant, instead of what the machine *thought* you meant.

4.1. Detecting and describing ambiguities

For Fairslator, the input is a pair of texts: a text in the source language (the **original**) and its translation in the target language (the **translation**). A text can contain one or more sentences.

Internally, Fairslator starts the process of bias detection by running the two texts through a third-party dependency parser[9] and obtaining a parse tree for each text. These are then passed to Fairslator's own algorithms which attempt to discover the following facts about the translation.

1. Is the **speaker** mentioned in the translation, for example by first-person pronouns? And if so, is the speaker mentioned in the translation in a way that encodes gender, while the original doesn't?

2. Is the **listener** mentioned in the translation, for example by second-person pronouns or implicitly through verbs in the imperative? And if so, is the listener mentioned in the translation in a way that encodes gender, number of formality while the original doesn't?

3. Are any **bystanders** mentioned in the translation, that is to say, are any people being referred to by nouns or by third-person pronouns? And if so, are the bystanders mentioned in the translation in a way that encodes gender, while the original doesn't?

These are the three axes of ambiguity. Each original-plus-translation pair contains zero or one *speaker* axis, zero or one *listener* axis, and zero, one or more *bystander* axes. For each axis, Fairslator tells us whether there are any unresolvable ambiguities on this axis, what the options are (e.g. the translation can be either masculine or feminine along this axis) and which of them is actually present (e.g. the translation is masculine along this axis).

Let's illustrate this all with an example. Assume the input is the following pair of original and translation.[10]

> English: *I would like to ask whether this is your new doctor.*
> Czech: *Chtěla bych se zeptat, zda je to tvůj nový lékař.*

1. The speaker axis is present here. The speaker is mentioned in the translation with the words *chtěla bych* 'I would like to' where the word *chtěla* is a participle and encodes the speaker as female in gender, while the original is ambiguous as to the speaker's gender.

---

[9] Currently Fairslator uses UDPipe 2, a depenency parser developed at Charles University in Prague. In principle Fairslator can work with any dependency parser which outputs Universal Dependencies (UD) parse trees.

[10] Yes, the example is a bit convoluted. That is necessary in order to demonstrate all three axes.

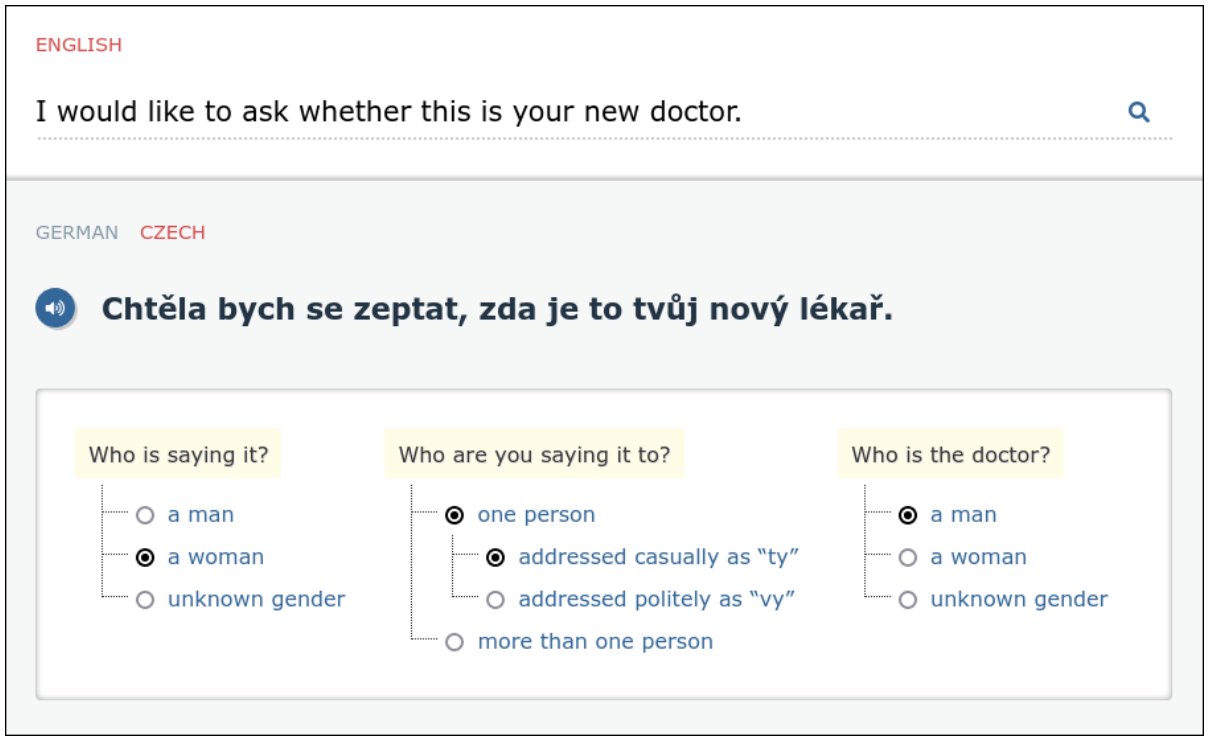

Figure 3. *Fairslator offers choices to disambiguate ambiguous input.*

2. The listener axis is also present here. The listener is mentioned in the translation with the word *tvůj* 'your'. This word encodes the listener as singular in number and casual in formality, while the original is ambiguous on these things. Neither the original nor the translation say anything about the gender of the listener.

3. Finally, one bystander axis is present here. The bystander is mentioned in the original by the word *doctor* and in the translation by the word *lékař*. The word in the translation encodes the bystander as male in gender, while in the original it is ambiguous in gender.

4.2. Offering choices to the user

Now that we have detected any unresolvable ambiguities in the original-plus-translation pair and described them symbolically in a machine-readable way, the next challenge is to communicate these to the human user who is sitting in front of his or her computer, waiting for a translation.

We have seen in the previous section how Google, when it determines that there are two possible translations, shows both of them on screen with explanatory labels. This is a good user experience if the number of variants is small, ideally just two. In Fairslator the number of variants is sometimes a lot larger than that because Fairslator handles more types of bias and they can co-occur in a single text. For example, we have seen how even a simple sentence such as *you are happy* can have as many as six translations in Czech and French depending on what one means by *you*. Showing so many translations on screen at the same time would be confusing.

So, Fairslator only offers one translation at a time, plus menus of choices. Each axis (speaker, listener, bystander) is introduced with a question such as *who is saying it*? or *who are you saying it to*? followed by options to choose from. Each time the user makes a choice, the translation is updated to reflect that.

The structure and wording of the menus have been designed with two goals in mind.

- Firstly, it must always be clear which aspect of the text is affected by each menu: whether it controls properties of the speaker, the listener, or someone else (a bystander). If someone else, then Fairslator asks by reusing a noun from the original text, for example *who is the doctor?* or *who are the students?* It is not enough to say that the entire sentence is "feminine" (like Google does) or in "informal tone" (like DeepL does). Fairslator is more specific than that, it tells the user exactly which "actors" in the sentence those labels apply to.
- Secondly, the options are worded so as to be as easily understandable as possible, even for linguistically naive users. Fairslator doesn't use technical terminology such as "masculine" and "feminine", instead it talks about "men" and "women". Fairslator doesn't say "singular" and "plural", it says "one person" and "several people". The biggest challenge is how to describe the formality dimension (*du* versus *Sie* and so on) which, unlike gender and number, does not correspond to anything objectively observable in the real world. Some English speakers who have never come across it before may genuinely have trouble understanding what this is all about. Fairslator uses the labels "addressed casually" and "addressed politely".

The point to make here is that solving bias is not just a technical challenge for software developers and linguists. It is also a user-experience (UX) challenge. Interaction with the user is a must here, the machine needs the user to disambiguate what he or she meant. To

have a successful interaction, the machine needs to formulate its questions (and the available answers) in a way which will be clear and easily understandable to the user.

For most of its history, machine translation has given people tools where the interaction is short and linear: in goes the text in one language and out it comes in another language. For naive users without much linguistic awareness outside their own language, this may have created the unrealistic expectation that there is always exactly one and only one correct translation for everything. Fairslator reminds people that this is not always the case and that, sometimes, the machine needs to ask follow-up questions. This is quite a shift: the human-machine interaction now becomes more, well, interactive, as the user has to click on things and to progressively tweak the translation closer to his or her intended meaning. The entire user experience around this must be designed well, so that the user perceives it not as an inconvenience but as helpful and useful.

4.3. Reinflecting the translations

Once the user has made a choice in the menus, the translation is sent back to Fairslator for **reinflection**. Reinflection is the process of changing some words in the translation to reflect the user's choices, such as swapping female words for male ones. For example, if the user has asked to change the speaker axis from female to male, Fairslator will change the word *chtěla* to *chtěl* ('would like to').

> English: *I* (a man) *would like to ask whether this is your new doctor.*
> Czech: *Chtěl bych se zeptat, zda je to tvůj nový lékař.*

Sometimes, when a word is changed in the translation, this requires other words to be changed also, in order not to break grammatical agreement. Suppose the user has asked to change the bystander axis from male to female. This means Fairslator will change the word *lékař* 'male doctor' into *lékařka* 'female doctor'. And this in turn means that words which depend on it syntactically need to be changed too: *nový* to *nová* ('new') and *tvůj* to *tvoje* ('your').

> English: *I would like to ask whether this is your new doctor.* (female doctor)
> Czech: *Chtěla bych se zeptat, zda je to tvoje nová lékařka.*

Occasionally, the form and shape of a single word in the translation may be influenced by multiple axes at the same time. This is typically the case with possessive pronouns such as *tvůj* 'your' in our example. This one word encodes number and formality from the

listener axis and, additionally, it is required to agree in gender and case with the noun *lékař* ‘doctor’ from them bystander axis. So, when the user has asked to make changes to both axes at the same time, the word needs to be changed twice: first from *tvůj* to *váš* (because the listener axis is changing from casual to polite) and then from *váš* to *vaše* (because the bystander axis is changing from male to female).

As you can see, the process of reinflection is far from trivial, and requires that the software doing it – that’s Fairslator – knows a great deal about the grammar of the languages involved.

4.4. Discussion

Fairslator does in fact know a great deal about the grammar of the languages involved. The algorithms that analyze the bilingual text pairs into axes as well as the reinflection algorithms, are rule-based and hand-coded individually for each language. Some subroutines are even hand-coded individually for each language *pair*.

In other words, there is no machine learning and no artificial intelligence in this set up (at least not of the now-commonplace statistical, probabilistic kind). The translations which arrive into Fairslator as input may well have been produced by such an AI, but the analysis which Fairslator itself performs on them is not (with the exception of the third-party dependency parser which Fairslator uses as the first step in its analysis).

Is this a good thing or a bad thing? The usual argument against hand-coded rule-based algorithms and in favour of machine-learned statistical ones is that the latter simply produces better results and has broader coverage, while the former is expensive (in terms of people and time) and does not scale, making it difficult to achieve broad coverage. In this particular case, however, it is the other way around. All pre-existing attempts to “solve” machine translation bias, which as far as we know have been based mostly on statistical methods and machine learning, have achieved only modest successes, as we have documented in the previous section on examples from Google and DeepL.

Fairslator’s approach, while it requires hand-coding, does not require prohibitive amounts of it. The programming effort required to hand-code the rules for detecting and correcting bias in one language pair is counted in a single-digit number of days (for one person), which is far below a typical software engineering project. The only prerequisite is that the programmer is a computational linguist in the old-fashioned sense of the term: someone who knows a lot about the syntax and morphology of the languages involved.

At the same time, Fairslator has been able to achieve noticeably broader coverage than other, AI-based attempts. First of all, Fairslator handles a broader range of bias phenomena than anyone else: not just gender bias on nouns, but also gender bias on other word classes, as well as number and formality bias on second-person pronouns. Secondly, Fairslator's accuracy and coverage in detecting and correcting bias of these types is high[11] – precisely *because* it is rule-based and most errors can be corrected by tweaking the rules. In fact, the small number of errors that *cannot* be corrected are caused by the underlying third-party parser. Last but not least, Fairslator is the first attempt ever (as far the author knows) to solve machine translation bias for Slavic languages which are known to be a lot more "gendered" than others.

[11] Precisely how high is not known yet. I have not done any systematic tests yet.

5. **Conclusion**

I started the Fairslator project and wrote this whitepaper to solve a problem that had been bothering me for a long time: that machine translators never ask me *what I actually mean.* This whitepaper is my attempt to communicate ideas and insights which I have been developing for *years*, literally. Some are obvious once you think about it, while others are – I would hope – novel and insightful, such as the distinction between *high-detail* versus *low-detail* languages or the idea that bias can be described along the three axes of *speaker*, *listener* and *bystander*. But if I had to summarize it all in one short executive summary, I would do it with the following three statements.

1. When machine-translating texts from one language to another, the translations are sometimes **biased** because the machine has made an **unjustified assumption** as to someone's gender or as to the meaning of an ambiguous word such as English *you*.
2. Some ambiguities are **unresolvable** because there are no clues in the text. This means that some instances of bias can never be "solved" just by improving existing AI. The only way to resolve an unresolvable ambiguity is to bring the human into the loop, to ask the user what he or she meant.
3. Fairslator is an experimental application which is able to detect and describe bias in the output of any machine translator, and to offer a human user options for correcting it.

If you only take three things home with you, let it be these. Thank you for your interest in Fairslator and see you online at www.fairslator.com.

Michal Měchura
michmech@lexiconista.com