DEPARTMENT OF PHILOSOPHY,
LINGUISTICS AND THEORY OF SCIENCE

SENSE AND SENSITIVITY

Exploring how Neural Machine Translation Systems Handle Slurs

Tom Södahl Bladsjö

Bachelor's Thesis:	15 credits
Course:	Lingvistik Fördjupningskurs LI1301
Level:	Bachelor's level
Semester and year:	Spring, 2022
Supervisor	Sharid Loáiciga
Examiner	Eleni Gregoromichelaki
Keywords	Machine Translation, Descriptive Translation Studies, Slurs, Sensitive Language, Offensive Language, Lexical Semantics

Abstract

The rise of streaming platforms such as Netflix and HBO has brought a surge in audiovisual content to be translated. While the translation industry at large have adopted machine translation (MT) as a tool to meet the rising demands, the subtitling industry has been reluctant to embrace this trend. One possible reason is that MT tends to perform badly on the kind of content found in audiovisual material, such as sensitive and offensive language. Meanwhile, research in the intersection of MT and human translation is sparse, and the differing approaches and terminology used within the two fields seem to impede interdisciplinary collaborations. In this work I explore how approaches borrowed from descriptive translation studies (DTS) can be adapted for analysis of MT output. I apply a taxonomy inspired by those used for analysis of human translation to output from three different MT systems in order to investigate how they translate slurs from Swedish to English. In contrast to conventional approaches to MT evaluation which focus on overall quality, the approach used here aims to descriptively capture the ways a translation can relate to its source text on a lexical-semantic level. The results provide some preliminary insights into the kinds of semantic divergence that can be introduced in the translation process when using MT to translate slurs. These can be seen as pointers to areas for future investigations. For example, some instances of gender bias were identified, where systems tended to translate terms denoting individuals of any gender into terms specifically denoting men. While further revision may be needed for the taxonomy to better fit the material to be analysed, this initial investigation suggests that DTS approaches could be well suited as the basis for more fine-grained analyses of MT output.

NOTE: This thesis deals with sensitive and offensive language and includes, by necessity, several instances of such language as examples.

Acknowledgements

I would like to thank my supervisor Sharid Loáiciga for her advice, useful suggestions, and not least her expert help with the technical aspects of this study.
I would also like to thank Eleni Gregoromichelaki for her help at the start of this project, and Johan Blomberg for useful advice and feedback.
Finally, I would like to thank Plint AB for their support through this process, and particularly Ivana Marić for her advice and encouragement. A special thanks also to Anders Langhoff for taking the time to answer my questions.

Contents

Introduction

For a professional translator transferring a text from a *source language* (SL) to a *target language* (TL), some linguistic expressions tend to present more of a challenge than others. For example, offensive language such as slurs and swearwords tend to be closely tied to the culture in which they originate. In many cases there exists no exactly equivalent expression in the target language. In a blog post with the telling title "Translating swear words: F***ing challenge !", journalist Sébastien Blanc comments on how hard it is to "find the right translation for swear words, insults and other jaw-dropping expressions that reflects exactly the coarseness and level of animosity expressed by the person who uttered them" (Blanc, 2022). In many cases, rather than deciding on a single correct translation for each expression, translators have to imagine what a TL speaker would realistically have said in the same situation and make their decisions on a case-by-case basis.

With the increasing availability of multilingual content online, the demand for translation is growing. The rise of streaming platforms such as Netflix and HBO in recent years has brought a surge in audiovisual content to be subtitled. The past decade or so has also seen a large upswing in reality shows (A. Langhoff, freelancing subtitler at Plint AB, personal communication, May 17, 2022). In this type of content, unlike in most movies and series, the dialogue is not pre-scripted, and translators have to deal with all kinds of phenomena found in naturally occurring dialogue, such as interruptions, simultaneous speech, slang and offensive language.

The growing amount of content to be translated has made it necessary for the industry to find ways to meet the rising demands. Due to its rapidly improving quality in recent years, machine translation (MT) has been increasingly adopted as a tool in the translation industry (Bellés-Calvera & Caro Quintana, 2021). These days, post-editing machine translated output instead of translating from scratch is a widely used way to increase efficiency and lighten the translators' workload. However, some areas of the translating industry, like the subtitling industry, are still largely reluctant to use MT as a tool (Bywood et al., 2017). There are many possible reasons for this. One is that post-edited MT output shows less lexical variety than texts translated from scratch (Farrell, 2018; Matusov et al., 2019). This could make MT unsuitable for translating material where lexical uniformity is not desired, such as fiction, entertainment and other creative genres. Another issue that has been raised is that MT does not do well with rare words and trickier passages such as offensive language (A. Langhoff, personal communication, May 23, 2022), slang and puns (Matusov et al., 2019).

MT is a subfield of *natural language processing* (NLP), a field concerned with making computers process human language. The currently most wide-spread approach to MT is *neural machine translation* (NMT), whereby a system is trained on a large corpus of parallel texts to be able to predict the most probable output in a target language given an input in a source language. The performance of an NMT system is normally evaluated based on the fluency and adequacy of the output. These are global quality assessments that do not provide any information on what exactly went wrong in cases where the system did not perform well. Neither do they provide any insights into areas where users report MT to perform badly, such as rare words and slang.

In contrast, academic research on human translation tends to focus on specific phenomena of interest (Toury, 1995/2012) rather than overall quality. Since the 1990's, it is also more geared towards descriptive investigations than prescriptive assessments of translation quality (Diaz-Cintas, 2012). The descriptive approach allows for the fact that there are asymmetries between languages; different languages have different ways of encoding the same information, and there is rarely a one-to-one correspondence between the vocabularies of two languages. Furthermore, two words that denote the same concept may carry different sets of connotations. In some cases, the translator has to make a conscious choice about which part of the linguistic meaning to prioritise in translation. These things are are typically disregarded in MT research, where focus tends to be on the big picture.

This mismatch between MT research and research on human translation when it comes to focus and choice of analytical framework may contribute to the gap that seems to exist between the two fields, impeding interdisciplinary collaboration. This thesis, while not a solution to the problem, aims to explore ways in which the two fields could be brought closer to each other. For that purpose, I analyse the performance of three commercially available MT systems using an analytical framework based on *descriptive translation studies* (DTS). As mentioned above, studies of human translation tend to focus on a specific phenomenon of interest, typically one that is particularly challenging or requires more conscious decision-making on behalf of the translator. The target phenomenon of this study is the translation of sensitive or offensive language, in this case specifically slurs.[1] In contrast to conventional, quality focused evaluation of MT output, the approach used here aims to descriptively capture the different ways a translation can relate to its source text on a lexical-semantic level. Particular attention will be paid to what aspects of meaning are transferred from the *source text* (ST) to the *target text* (TT), and how this may affect a human end-user, in this case a professional translator or post-editor.

The aim of this study is exploratory; rather than carry a specific hypothesis, the objective is to gain some initial insights which could point to possible areas for future investigation. With that in mind, I seek to answer the following questions:

1. How do these MT systems handle the translation of slurs? Are there any patterns that emerge?
2. What can be learned from applying DTS approaches to analysis of MT output?

[1] For a working definition of *slur*, see § Conceptualising Slurs – Some Different Approaches.

Background

Subtitling

Audiovisual Translation and Subtitling

Subtitling can be considered a subcategory of *audiovisual translation* (AVT) or *screen translation*. As the names suggest, these terms refer, respectively, to translation of content in contexts with both audio and visual components, and translation that appears on a screen. The two are used more or less interchangeably in academic works[2] (Pedersen, 2007b). For the sake of consistency, the term *audiovisual translation* (AVT) will be used throughout this work. AVT is one of several services covered by the larger *localisation industry*, which consists of companies that help businesses make their products available for a global market by providing translation as well as adaptation to local cultural requirements. There are several different modes of AVT, of which the three most common are subtitling, dubbing and voice-over. Within subtitling, a distinction is made between interlingual and intralingual subtitling, where only the former includes what would conventionally be called translation; the latter, which is sometimes referred to as *subtitling for the deaf and the hard-of-hearing* (SDH) or *closed captioning*, consists of a written rendition of the spoken content in the same language. The present study will deal with interlingual subtitling, English to Swedish, specifically.

Machine Translation and Subtitling

While technological aids such as translation memories (TM) and machine translation (MT) are regularly used in the wider localisation industry, the subtitling industry has, at large, been reluctant to embrace this practice (Bywood et al., 2017). One reason for this may be that most subtitles are written representations of spoken dialogue, which does not always follow the usual patterns of written language. Thus, handling this type of material may be more challenging for translation systems developed for written language, and generate more errors. Furthermore, the interpretation of utterances in an audiovisual context may be dependant on information that is only available in the visual image and thus not accessible to a system that only handles textual input. Another reason could be the specific spatial and temporal constraints associated with the medium, such as line exposure times and character limits. These constraints make it the subtitler's task not only to translate the source text, but also to condense and adjust it according to the guidelines of the current employer.[3] However, with the advance of MT technology, new efforts have been made to incorporate MT systems in the subtitling workflow, and studies on the subject show promising results in terms of translation quality and efficiency (e.g. Matusov et al., 2019; Bellés-Calvera & Caro Quintana, 2021). Nevertheless, some express concern that the post-edited MT output will not reach the standard of human translation, and Farrell (2018) suggests that the post-editing effort required to make the quality of the MT output truly match that of human translation may in fact cancel some or all of the proposed productivity gains. In his study on what he calls *MT markers* in the post-edited output, Farrell (2018) shows that the variety of different translation solutions is noticeably lower in post-edited MT output as compared to translations done entirely by humans, since post-editors, when faced with an acceptable translation, tend not to edit it even if there are any number of other possible translations for the same expression. This, he argues, makes post-edited MT an unsuitable solution for genres where features such as variety and inventiveness are quality factors. Similar effects were noted by professional translators involved in a study by Matusov et al. (2019), who raised concerns that "uncommon but correct translations in the target language would not be corrected in a post-editing workflow, potentially affecting the overall result." One translator in the same study also noted that using MT only worked well for the simpler parts, where translating was more or less straight-forward;

[2]The terms *media translation* and *multimedia translation* are also sometimes seen as synonymous.
[3]See e.g. Netflix (2022a) for an example of such in-house guidelines.

the more creative parts, like slang and puns, were mistranslated and had to be edited. Similarly, subtitler Anders Langhoff notes that "once you move out of very simple content, the MT's will often be a hindrance rather than a help." He mentions offensive language as one category that is not handled very well in machine translation (A. Langhoff, personal communication, May 17, 2022).

The Subtitler as a Mediator of Culture

When it comes to deciding what constitutes quality in subtitles, a view that is commonly held, especially among subtitling professionals, is that good subtitles should be 'invisible' – that is, they should provide understanding of the dialogue without attracting attention themselves (Szarkowska et al., 2020)[4]. Pedersen (2007a) introduces the idea of a 'contract of illusion' between the subtitler and the viewer, where the viewer, in taking part of the audiovisual work, agrees to temporarily suspend their disbelief and pretend that the subtitles *are* the dialogue. When the subtitles themselves attract attention by, for example, diverging from what is expected or by being noticeably mistranslated, this contract of illusion is broken. In other words, there seems to be some degree of consensus among professionals that successful subtitles make the viewer believe that what they are reading is in fact what is being said. This observation, while perhaps not very remarkable in and of itself, may appear more significant when taking into account the prominent position of audiovisual productions in today's society and, consequently, their potential as vehicles for cultural transmission (Diaz-Cintas, 2012). Since "[...] translation is not carried out in a vacuum and cannot, therefore, be exempt from a certain degree of subjectivity and bias on the part of the translator and the rest of the agents involved in the translational process" (Diaz-Cintas, 2012, p. 282), the message conveyed by the subtitles will in many cases, consciously or unconsciously, have been altered to some extent in the process of translation. If, then, the subtitler has succeeded in their task of making the subtitles so unobtrusive as to be 'invisible', this altered message will be perceived by the viewers as identical to the original, and the viewer's resulting perception of the information conveyed will be unavoidably shaped by the choices made by the subtitler. Díaz-Cintas again:

> As a site of discursive practice, audiovisual media and its translation play a special role in the articulation of cultural concepts such as *femininity*, *masculinity*, *race*, and *Otherness*, among others. It can contribute greatly to perpetuating certain racial stereotypes, framing ethnic and gender prejudices, and presenting viewers with out-dated role models and concepts of *good* and *bad* seen as rigid, diametrically opposed. (Diaz-Cintas, 2012, pp. 281-282)

In recent years, this role of the subtitler as a mediator of culture has been investigated in several case studies (e.g. Filmer, 2012; Ávila Cabrera, 2016; Martínez Pleguezuelos, 2020). According to Netflix's guidelines for subtitles (Netflix, 2022b), "Translations and transcriptions should always be an accurate representation of the intent of the original content language without adding additional vulgarity or censorship." However, this is not always an easy achievement. Filmer (2012) observes, in her investigation of the translation of offensive language in the 2008 American movie *Gran Torino* into Italian, that the racial slurs that abound in the source text have gained a more homophobic slant in the target rendition, presumably to make up for the smaller variation of racial slurs in the target language and to reach the level of offensiveness achieved by the source text.

[4]The same study by Szarkowska et al. reported that viewers of subtitled content did not single out invisibility and unobtrusiveness as an important factor in subtitle quality the way professional subtitlers did, but whether this stems from a real difference in views on what constitutes quality or from different perspectives and insight into the subtitling process remains to be determined.

DTS and Ways of Classifying Translation strategies

Descriptive Translation Studies (DTS) are the branch of translation studies that focuses on how translation is or tends to be carried out by translators. According to Toury (1995/2012), the object of Translation Studies as a whole is to deal systematically with three types of issues:

> 1. all that translation can, in principle, involve;
> 2. what it does involve, under particular sets of circumstances, along with the reasons for that involvement, and
> 3. what it is likely to involve, under one or another array of specified conditions.
>
> (Toury, 1995/2012, p.9)

DTS, as the empirical branch of Translation Studies, focuses mainly on the second of these points. Toury (1995/2012), in his approach, further assumes a view where translations are regarded as facts of the target culture and have to be analysed from that perspective. Thus, Toury's *target-oriented* framework attaches more importance to the role the translation is meant to play in the culture that will host it, while a *source-oriented* framework focuses more heavily on faithfulness to the source text. A useful concept in DTS, whether source oriented or target oriented, is the notion of *norms* – that is, conventional rules of behavior shared by a community, in this case the community of translators. In theory, norms can be divided into *descriptive* and *prescriptive* norms, where *prescriptive* norms are those that are prescribed by authorities such as instructors or company guidelines, and *descriptive* norms are based on observation of actual translation practices (Pedersen, 2018).[5] However, since both categories evolve over time and feed into each other – descriptions of what translation is like can serve as a basis for the instruction of new translators, who will integrate them in their own practice, perpetuating the cycle – it is often hard to draw a clear line between the two in practice. In order to identify the norms that govern translators' behavior, a corpus of translated data needs to be collected, classified according to a chosen taxonomy depending on the feature of interest, and subject to analysis (Toury, 1995/2012, p.70). According to Pym (2002), the feature of interest should be one that is considered particularly problematic to translators and where there is no general agreement on standard solutions. In other words, DTS, even as a descriptive science, is ultimately about solving problems of linguistic mediation:

> Any kind of problem-solving requires descriptions that are adequate in some way. But that is not enough; description should not be the problem as such. The more important problems are social; they come from the societies around us; they require solutions that might help improve the lives of those around us. If not, we are merely playing, at someone else's expense. (Pym, 2002)

One example of a feature that has been extensively studied in the DTS paradigm is metaphors (e.g. Pedersen, 2017b), which are sometimes presented as "a kind of ultimate test of any approach to translation" (Toury, 1995/2012, p.107). Another example of a feature that has been studied using a DTS framework is *Extralinguistic Cultural References*, or ECRs. An ECR is defined as a "reference that is attempted by means of any cultural linguistic expression, which refers to an extralinguistic entity or process" (Pedersen, 2007b). In other words, ECRs are references to aspects of the source culture, such as places, people, foods and institutions, that require culture-specific real world knowledge for the target audience to understand. To capture regularities in the translation strategies used for ECRs, Pedersen (2007b) chose a taxonomy based

[5] Note that while norms themselves can be descriptive or prescriptive, the study of these norms is always, per definition, descriptive, since it merely describes what the norms are without involving any value judgements.

on a primary division of strategies into *source oriented* and *target oriented*, i.e., whether the translation stayed true to the literal meaning of the expression used in the source language or whether it adapted the ECR in any way to be understood better by the target audiences, e.g. by substituting it for an ECR playing a similar role in the target culture.[6] A simplified version of the taxonomy, adapted to focus on the translation process itself rather than the end result, is presented in Figure 1.

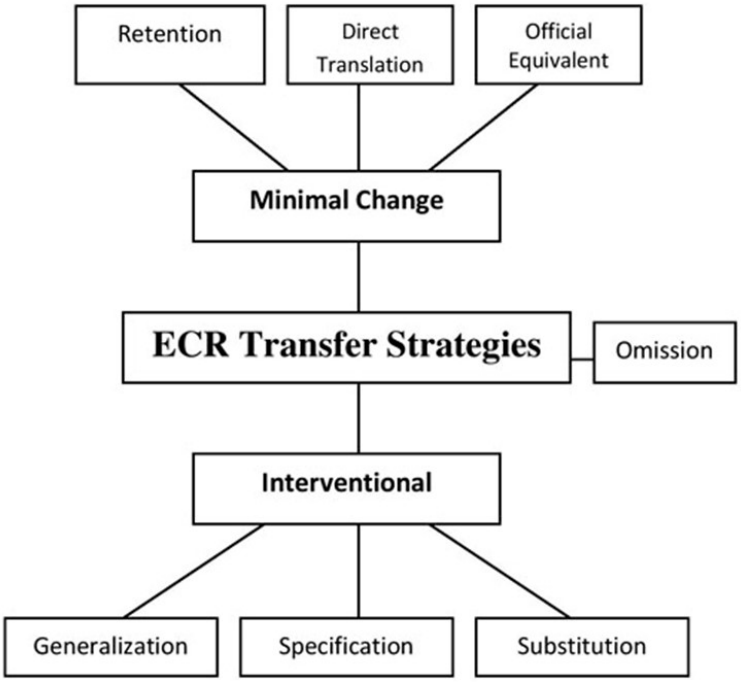

Figure 1: Simplified taxonomy of ECR transfer strategies (Pedersen, 2007b, p.154)
(Image from Pedersen, 2016)

ECRs and metaphors are both examples of what Pedersen calls *Translation Crisis Points*, or TCPs, that is, features of a source text that require the translator's attention and forces them to make more conscious decisions on what translation strategy to employ (Pedersen, 2016). For the purpose of this study, I make the assumption that slurs and other offensive language, as linguistic units strongly tied to cultural context, also constitute TCPs and can be analysed using a similar framework.

The Target Phenomenon

Conceptualising Slurs – Some Different Approaches

Slurs are generally taken to be a subcategory of offensive language, and as such they have been defined and analysed in a variety of different ways from both semantic and pragmatic viewpoints. Bach (2018) contrasts what he calls *group slurs* against *personal slurs*, which target individuals based on personal characteristics rather than group membership. He describes *group slurs* as having the same referential content as their

[6]Note that this distinction between *source oriented* and *target oriented* differs from Toury's (1995/2012), which is about perspectives from which to view translation as a concept, rather than strategies for tackling individual translation problems.

neutral counterparts – i.e., to classify an individual according to group membership – while also containing secondary propositional content declaring the target to be contemptible because of that group membership. Analysed this way, a sentence of the form "X is a [slur]" does not have a single truth-value, since it expresses two independent propositions rather than a single conjunctive one.[7] Hom (2008) presents *combinatorial externalism* as a lens through which to view the semantics of slurs. He argues that the derogatory content of slurs is semantically determined by the external, social practices of the linguistic community to which the speaker belongs and thus closely tied to negative ways the target group is viewed by the community at large, as well as to what discriminatory practices they are subject to within that community. Thus, according to Hom (2008), when using a slur instead of a neutral word (or *nonpejorative correlate*) to categorise an individual as a member of a certain group, one is not only ascribing negative properties to them (similar to Bach's secondary propositional content), but also invoking the threat of discriminatory practices towards them. Bolinger (2017) argues that the derogatory power of slurs is best explained in terms of purely pragmatic mechanisms; when presented with a free choice between using a slur and its neutral counterpart, the speaker, when choosing the slur, signals by *contrastive choice* that they are endorsing the derogation and the discriminatory practices associated with the term. However, Ashwell (2016) questions these views and points out that not all slurs can be considered to have a neutral counterpart. Burnett (2020) also questions the notion of 'neutral counterparts'. Using the terms *lesbian* and *dyke* as examples, she notes both that the 'neutral counterparts' of slurs are often not in fact neutral, and that slurs are often used in a non-derogatory way by members of the target group themselves (a phenomenon known as *appropriation* or *reclaiming*). Taking these facts into account, she proposes a semantics of slurs that views *lesbian* and *dyke* as associated with different sets of *personae*, and the derogatory effect of a term as dependent on group attitudes towards the *persona* with which it is associated (and particularly on the perceived attitudes and group membership of the speaker).

Within NLP, the field that has paid most attention to slurs is *hate speech detection* which, as the name suggests, aims to develop ways to automatically detect and filter abusive content online. This is particularly relevant to social media platforms, where the material is often too extensive for all-human moderation to be feasible. In hate speech detection, the presence of slurs and other abusive words are often used in combination with other features for detecting instances of hate speech (Schmidt & Wiegand, 2017). However, Palmer et al. (2020) note that this approach is not enough to capture the complex nature of offensive language, and particularly that current systems often wrongly categorise (and accordingly censor) uses of reclaimed slurs by group members themselves, which in turn serves to further punish already stigmatised communities. They point out that, as in the case of appropriation, slurs can often be present without malignant intent (on behalf of the speaker) or effect (on the targeted group); conversely, there are many ways in which language can be derogatory or abusive without the presence of slurs or other specific lexical items conventionally regarded as offensive.

From all of the above, it is clear that the phenomenon of slurs is a complex one both in terms of definition and analysis approaches, and that to obtain a full view of derogatory language one has to move beyond a purely lexical level. Unfortunately, the scope of this thesis does not allow for such a thorough investigation of the subject. For that reason, while acknowledging that this approach can only give a very limited understanding of the issue at large, I have chosen to restrict my focus to the lexical aspect – that is, the slur words themselves. For the purposes of this study, and taking into account the issue of reclaimed uses by group members themselves as pointed out by Burnett (2020) and Palmer et al. (2020), I take a *slur* to be

> **A term denoting a group that is in some way marginalised by the society in which the term is uttered, that is *or could be* perceived as offensive by members of that group.**

This is the working definition that will be used throughout this thesis. As can be seen by the phrasing, it

[7] This might be understood as slurs containing a built-in presupposition of contemptibility on the part of the referent.

excludes entirely what Bach (2018) calls *personal slurs* from the category of slurs, since I, like Hom (2008), take the view that the category *slurs* is closely connected to discriminatory practices in society and therefore does not include offensive terms directed at individuals regardless of group membership. It also includes terms that are generally used as neutral words but which could be perceived as offensive in some contexts. In that way, it is a somewhat broader definition than those that contrast slurs against neutral terms, *nonpejorative correlates* or similar. The reason is that word use varies continuously between groups and over time, and any attempt to make a sharp distinction between offensive and inoffensive words is likely to be inaccurate for some contexts. Under this definition, every word denoting a marginalised group is treated as potentially offensive under certain circumstances, and the difference is seen as gradual rather than categorical.

Slurs in Fiction and Entertainment

Recent years have seen a heightened public awareness of discriminatory or abusive language, and uses of slurs in media have sparked strong reactions. For example, in 2020, the BBC received over 18,600 complaints about the use of a racial slur in a news report (BBC, 2020). Offhand uses of homophobic slurs in older fiction and entertainment that were once considered unproblematic are also being reexamined by audiences and found unacceptable according to current standards (Alexandra, 2017). On the other hand, there are uses of slurs in fiction that are required for the purpose of the artwork and therefore deemed acceptable by the audience. For example, in material depicting historical (or current) discrimination of a group, it may be contextually justified to include slur uses that are part of those discriminatory practices. In such cases, the strong emotional response invoked by the use of the slur may be what the author is after. In these cases (so called 'art uses'), "[t]he slurs have their normal offense-generation patterns *within* the fiction, but do not generate offense or censure beyond it: the actor, writer, producer etc. are held innocent when art uses are deemed acceptable" (Bolinger, 2017, p.16). On the other hand, as seen above, there are cases when this "insulation", or separation between fictional and real-world offense, does not hold. For example, Bolinger (2017, p.16) points out that "when a comedian's use of a slur–even while portraying a racist character–is not well-received, audience anger is directed at the comedian." This, Bolinger means, can be explained in terms of *contrastive choice*. If one sees art uses of slurs as a context where the choice of a slur over a neutral term is not in fact a *free* choice, but rather motivated by the purpose of the artwork, then an art use of a slur

> will be insulated when two conditions are met (and license offense otherwise): (i) the use of the slur is required for the purposes of the artwork, and (ii) these purposes are good enough to justify the slur. Histories, social critiques, and works that function to improve the social position of the group targeted by the slur(s) are the most likely to satisfy both conditions, and so should be expected to be successfully insulated more reliably than art uses which occur in comedies or as the punchline in a stand-up routine. Importantly, the forced choice justification only protects the speaker from offense if we judge that the slur *really* is required and justified by the purpose of the performance. (Bolinger, 2017, p.17) (...)

Machine Translation and Meaning

What We Mean by 'Meaning'

Before analysing what aspects of meaning are transferred when translating a slur from one language to another, it is useful to first define what is meant by meaning. When talking about NLP, in particular, it can also be useful to distinguish between a system on the one hand *capturing* the meaning of a linguistic unit, and, on the other hand, simply *conveying* that meaning to a human end-user. While the two may seem similar at a first glance, they require different evaluation strategies (see Bender & Koller, 2020). Most evaluation approaches in current MT research are meant to measure the quality of the MT output as perceived by a

human (these approaches will be further described in the following section), but see Poliak et al. (2018); Belinkov et al. (2017) for examples of research that focuses specifically on the semantic knowledge that is encoded in the neural machine translation (NMT) models themselves. The focus of the present study is meaning as communicated to a human end-user, regardless of how that meaning is encoded in the system itself.

As can be seen from the literature in semantics and pragmatics, there are various ways to analyse linguistic meaning. In truth-conditional semantics, to know the meaning of a sentence is to know under what circumstances it is true. Grice (1968) broadens the notion of meaning by distinguishing between the meaning of an utterance type X (what he calls *timeless meaning*) and the meaning of the speaker when uttering it (*occasion-meaning*). The former is tied to established conventions within the speaker community, while the latter has to do with the intentions of the specific speaker on the specific occasion when the utterance is made. Much more recently, Bender & Koller (2020) used a similar distinction when discussing the notion of meaning as pertains to NLP. They hold *meaning* to be "the relation $M \subseteq E \times I$ which contains pairs (e, i) of natural language expressions e and the communicative intents i they can be used to evoke", and distinguish this from *conventional* meaning, or the relation "$C \subseteq E \times S$ which contains pairs (e, s) of expressions e and their conventional meanings s" (Bender & Koller, 2020, p.5187). They, like most, take this conventional meaning to be about the relationship between linguistic form and the real-world thing or concept that the form conventionally represents (but see Sahlgren & Carlsson (2021) for a contrasting view).

Another view of linguistic meaning, which has been very successful in the field of NLP, is the *distributional hypothesis* (Harris, 1954), which is at the core of the field of distributional semantics. According to the distributional hypothesis, we can access information about the meaning of a linguistic expression through the contexts in which it appears. In fact, the success of many of the current approaches to NLP relies entirely on the ability of a language model trained on large corpora of text to access and encode this structural information, which can then later be used to solve specific tasks (Sahlgren & Carlsson, 2021). However, Bender & Koller (2020) caution against describing these models as *understanding* natural language or capturing *meaning*, since their notion of *meaning*, as defined above, cannot be accessed from text-only data[8]. Similarly, Emerson (2020) points out that "language is always *about* something", and that "if the meanings of words are defined only in terms of other words, these definitions are circular".

The fact that language models such as NMT models only have access to probabilistic information about how linguistic forms tend to combine with each other, but not to the concepts or real-world entities they refer to, is what makes Bender et al. (2021) refer to them as *stochastic parrots*: they can imitate human language, but they themselves do not know what it means. For machine translation, this means that just like descriptive and prescriptive norms of human translation feed into each other as translators will go on to translate in a way similar to what they have been taught (§ DTS and Ways of Classifying Translation strategies), the norms of human translation that are present in the training data will go on to feed into machine translation. While this may often result in output that seems coherent and very similar to what a human translator might produce, machine translation, unlike human translation, does not involve any conscious decision-making and is not adapted to suit the needs of the specific translation task (such as employer guidelines, see § DTS and Ways of Classifying Translation strategies). This can be problematic, because, as Bender et al. (2021) put it, "humans are prepared to interpret strings belonging to languages they speak as meaningful and corresponding to the communicative intent of some individual or group of individuals who have accountability for what is said". However, as the same authors go on to point out, while coherence and fluency may increase the *perceived* adequacy if the output, it does not in fact guarantee faithfulness to the source text.

[8] See Sahlgren & Carlsson (2021) for a contrasting view, and a discussion on different ways to conceptualise 'understanding'.

Evaluating Meaning Transfer

So how do we evaluate the transfer of meaning from one language to another? In MT evaluation, a common way is to let human annotators rate adequacy of the output as compared to a reference translation, as well as the fluency regardless of meaning. This can often be tricky, both because the close interaction between syntactic form and meaning makes it hard to disentangle the *fluency* and *adequacy* categories from each other, and because it is very hard for human annotators to agree on what constitutes an objectively 'good' translation (Denkowski & Lavie, 2010).[9] Alternatively, if several systems are to be evaluated simultaneously, annotators may instead be asked to rank the translations relative to each other, which makes the task somewhat easier, if still highly subjective (Denkowski & Lavie, 2010).

The high cost of manual annotation (Marvin & Koehn, 2018; Denkowski & Lavie, 2010) has motivated the development of automated evaluation methods, such as BLEU (Papineni et al., 2002) and METEOR (Banerjee & Lavie, 2005), to stand in for human annotators in cases where collecting human judgements would be too expensive or time-consuming. For example, automatic evaluation is an absolutely necessary part of NMT development, where it is needed to monitor changes and provide immediate feedback during training. These automatic evaluations generally compare the MT output to one or more human-made reference translations and calculate a score based on their similarity, and are themselves evaluated based on how well they correlate with human judgements of translation quality (Mathur et al., 2020). There are also approaches such as TER and HTER (Snover et al., 2006), which, instead of measuring fluency and adequacy, measure the number of edits that a human post-editor would need to make in order for the translation to reach the quality of a human-made reference translation.

BLEU, originally meant to be a "an automated understudy to skilled human judges which substitutes for them when there is need for quick or frequent evaluations" (Papineni et al., 2002), has since its introduction in 2002 become the standard metric for MT quality. However, this use of BLEU has also been much criticised, e.g. by Mathur et al. (2020), who emphasise that automatic metrics are inadequate substitutes for human evaluation.

Evaluation plays an important part in the development of MT systems, where it is needed to monitor the effects of changes to the system in order to distinguish good solutions from bad ones (Papineni et al., 2002). Because of the part evaluation plays in the development of MT systems, the norms governing MT evaluation differ quite a lot from academic analysis of human translation (as described in § DTS and Ways of Classifying Translation strategies), where the focus tends to lie on the translation of a specific feature of interest, and descriptive frameworks are usually preferred over quality assessments.[10] The approaches for MT evaluation mentioned this far, both human and automatic, are all global metrics, i.e. metrics to evaluate the overall quality of the translations[11] rather than individual elements. While an efficient way to check the overall performance of a system, they are not very transparent when it comes to specifics, such as what types of errors the system struggles with. While global metrics do constitute the norm in MT evaluation, some work has been done on evaluation of specific aspects of MT performance, such as disambiguation of words with several senses, so called *word sense disambiguation* (WSD) (e.g. Carpuat, 2013; Marvin & Koehn, 2018; Liu et al., 2018). Carpuat (2013) specifically highlights the need for finer-grained evaluation approaches as a complement to global metrics in order to gain a more complete understanding of the behavior of MT systems.

[9] Hämäläinen & Alnajjar (2021) also draw attention to the fact that "human evaluation is not conducted in the same rigorous fashion as in other fields dealing with human questionnaires such as in social sciences".

[10] But see Pedersen (2017a) for an example of a global quality assessment model specifically designed for evaluation of human translations.

[11] A sentence is common as the basic unit of evaluation.

The Present Study

The aim of this study is to explore how MT systems handle slurs, as viewed from the perspective of a human end-user. For this reason, focus will be on what is transferred or "communicated" rather than what is captured or understood by the system itself (cf Toury, 1995/2012, about translations as facts of the target culture). Further, this study will depart from the convention of quality-focused MT evaluation by using approaches borrowed from DTS as the basis for a more fine-grained analysis. According to Carpuat et al. (2017), "the translation process inherently introduces divergences that affect meaning, and semantically divergent examples should be expected in all parallel corpora." This study can be seen as an attempt to identify some of these divergences in MT output on a lexical-semantic level, and describe them in terms of 'what and how' rather than 'correct or incorrect'. A further objective is to explore how tools developed for hate speech detection can be repurposed for identifying slurs in a corpus of subtitles.

Because of the limited scope of this project I will restrict my analysis to what Bender & Koller (2020) call *conventional meaning* (§ What We Mean by 'Meaning'). This means that each target word translation will be analysed without taking into account who the speaker is and what the intention behind the utterance might be. In the case of slurs, this conventional meaning could be said to include the set of referents that the slur in question could conventionally be used about, as well as a set of attitudes associated with the use of that word as opposed to any alternatives. I will, henceforth, use the word *denotation* to refer to the former and *connotations* to refer to the latter, and I will consider the offensiveness of a given slur to be part of its connotations. However, no language usage ever occurs independently of context, so this notion of meaning must be seen as "an abstraction across usages, rather than a usage in a 'null' context" (Emerson, 2020). The set of usages that are primarily considered in this study are the ones where a slur is used to harm; reclaimed uses (§ Conceptualising Slurs – Some Different Approaches) are not addressed by my analysis. Thus, for each occurrence of a target term, the speaker was assumed to be a non-member and the hearer a member of the target group, regardless of the original context.

The difficulty of fully separating the meaning of a word from the context in which it is used will be further discussed in § Separating M-Meaning and C-Meaning.

Methodology

Materials and Pre-Processing

Data

This study relies on the 2018 version of the OpenSubtitles corpus (Lison & Tiedemann, 2016). The OpenSubtitles corpus is part of the larger OPUS, a collection of freely accessible parallel corpora, and includes 1,782 bitexts in 62 languages. It consists of movie and TV subtitles from a free on-line resource of user uploads (`www.opensubtitles.org`), covering various genres and time periods (Tiedemann, 2012). From the English-Swedish parallel corpus, I took a subset consisting of the first 100K lines[12] of the English side of the corpus in plain text format. I then translated these into Swedish using three different MT systems: Google Translate, Amazon Translate and Microsoft Translator.[13] In addition, the corresponding 100K lines from the Swedish side of the OpenSubtitles corpus were kept as a reference translation. Com-

[12]This can be estimated to roughly correspond to 100 movies. According to Lison & Tiedemann (2016), the 2016 release of the OpenSubtitles corpus contained 322K subtitle files in English, totalling 337M sentences. If each line of the corpus contains one sentence, then an average movie or TV episode contains just over 1000 lines.

[13]The translations were done through the cloud-based translation management system Memsource (`https://www.memsource.com`).

bined, the 100K lines from the English side of the corpus, the three MT translations of the same lines, and the corresponding lines from the Swedish side of the corpus constituted the data for this study.

Identifying Slurs in the Data

Slur words were identified in the English data by matching it to a list of discriminatory terms extracted from the Weaponized Word lexicons, originally developed for detecting hate speech and other content-based threats.[14] The multilingual lexicons are curated by staff and volunteers at The Weaponized Word with the assistance of their regional partner organisations around the world (The Weaponized Word, 2022). From the lexicon for English terms, the values for "term", "variant_of" and "plural_of" were filtered into a list which, after removal of duplicates, contained 1103 items. Upon inspection, several of these items turned out to be words that have one or several common non-slur senses. I therefore cleaned the list manually from words whose non-slur senses were deemed too common (e.g. personal names and names of foods), which resulted in a list containing 755 items.[15] The cleaned list was case-normalised before use.

The 100K lines from the English side of the OpenSubtitles corpus were also case-normalised and tokenised using the NLTK toolkit (Bird et al., 2009). Once pre-processed, the material was searched for tokens matching items on the list. The lines containing matching tokens, together with their corresponding Swedish segments, were extracted for analysis. The list items yielding matches, along with the number of matches for each item, were documented (see Table 2).

Analysis

For each of the three MT systems, the lines containing translations of one of the terms from the list were combined with their English source segments as well as the corresponding reference translation from the OpenSubtitles corpus. Each such triple of segments was then regarded as a unit. In cases with ambiguous words (such as the word *gay*, which can mean either *homosexual* or *happy*), the reference translation and the context within the unit were used to determine the word sense.

Source segment:	I'm always nervous or sick or sad or too **gay**.
MT output:	Jag är alltid nervös, sjuk, ledsen eller för **gay**.
Reference translation:	Jag är alltid nervös, sjuk, sorgsen eller för **glad**.

System: Microsoft Translator

Table 1: Example of a unit consisting of a source segment from the OpenSubtitles corpus, a translation by Microsoft Translator and a reference translation from the Swedish side of the corpus. The matched token in the source segment and the corresponding tokens in the translations are marked in bold.

I performed the analysis using the doccano annotation tool (Nakayama et al., 2018). Each unit of segments was analysed on three levels, described in more detail below.

[14] Data courtesy of The Weaponized Word (`https://weaponizedword.org`).

[15] Both lists (the original and the cleaned version) consist of data owned by The Weaponized Word, and unfortunately cannot be included in the thesis due to copyright issues.

Analysis of Matches

For each triple of segments, I first determined whether the source text term was in fact a slur and not a homonymous[16] non-slur. This was done taking into account both the local context and the Swedish reference translation. Segments where the term could be clearly identified as a non-slur homonym were labeled "false positives". This category also includes cases where slurs are used as part of expletives or other expressions that do not denote the target group, such as the uses of *bitch* in Example (1).

(1) a) **Bitch** of a mountain.

b) Son of a **bitch**!

In cases where the ST word sense could not be clearly determined from the context or the reference translation, the word was considered to be a slur by default.

Translation Error Analysis

In order to do a fine-grained analysis of the more subtle semantic differences that may be introduced in the translation process, it was necessary to first weed out some of the more blatant translation errors. For example, word sense disambiguation (WSD) errors were very common or all three systems (see § Results). For that reason, each instance of a target term and its corresponding translation was considered, and categorised as one of the following:

Slur translated as slur. Any triples where both the ST term and the translation could be considered slurs in the sense 'potentially offensive terms for marginalised groups', regardless of the quality of the translation. For example, instances where the word *gay* (in the sense *homosexual*) was translated as *bög* (faggot).

Non-slur translated as non-slur. Triples where the ST term was identified as a false positive (that is, a non-slur homonym), and not incorrectly disambiguated as a slur in the translation. For example, instances where the word *queer* (in the sense *strange*) was translated as *konstig* (strange).

WSD-error; slur to non-slur. Any triple where an ST slur had been translated as a non-slur homonym, such as the word *bitch* (used about a woman) translated as *tik* (female dog).

WSD-error; non-slur to slur. Triples where a false positive (that is, a non-slur homonym) was incorrectly disambiguated and translated as a slur, such as the word *queer* (in the sense *strange*) translated as *bög* (faggot).

Other mistranslations. Instances of mistranslated slurs that were not clearly disambiguation errors, such as when the TT word was neither a translation of the slur sense of the term nor of a non-slur homonym. For example, the term *mongrel* (used about a person) translated as *blandare* (mixer).

[16] No distinction is made between related word senses (polysemes) and non-related senses (homonyms), since tracing the origin of each of the terms on my list and all their possible senses lies outside the scope of this project. Furthermore, regardless of the actual etymological relationship between two word senses, determining how that relationship is perceived by speakers and hearers themselves is far from easy, and intuitions are likely to differ between individuals. For that reason, the word 'homonym' will be used about all cases where a word has multiple distinct senses.

Omission. Instances where the target word was completely missing from the translation.

Completely retained words – for example, the ST term *okie* kept as *okie* in the TT – were analysed in the next step and therefore not marked as errors on this level. In this preliminary categorisation, they too were labeled "slur translated as slur".

Semantic Analysis

For all triples where the target segment contained a slur (that is, both slurs translated as slurs and non-slurs incorrectly disambiguated as slurs)[17] the semantic relationship between the denotations of the source text terms and their translated counterparts was analysed (I consider a term's *denotation* to be the set of individuals it could, conventionally, be used to refer to). This categorisation is not meant to distinguish correct translations from incorrect ones; rather, it is meant to provide an overview of the different ways slurs are handled by the systems. All translations that involved a change in linguistic form (that is, all instances except those where a term remained untranslated in the target segment) were labeled according to the following categories, where D_{ST} is the denotation of the source text term and D_{TT} is the denotation of its translation:

Figure 2: Ways that the denotation of the translated term may relate to that of the source text term. D_{ST} is the denotation of the source text term and D_{TT} is the denotation of the corresponding term in the target text.

In addition to these four categories, cases where the token was completely unchanged in the translation were labeled either "Retained form (marked)" or "Retained form (unmarked)", depending on whether the form can be considered an established word in the target language. "Retained form (unmarked)" refers to instances where the term exists in identical form in the target language, as is the case with loanwords. For example, instances where the ST word *gay* was retained in the TT as *gay* were labeled "Retained form (unmarked)", since the word is an established loanword in Swedish. "Retained form (marked)" refers to instances where a foreign word, that is not entrenched in the TL and thus unlikely to be familiar to the target audience, is retained unchanged in the translation. Since, as stated above, the categorisation is not meant to distinguish correct translations from incorrect ones, any retained form not entrenched in the target language was labeled "Retained form (marked)", regardless of whether or not a more appropriate term exists in the target language. For example, the word *okie* is not an established Swedish term. When retained as *okie* in the TT, it was thus labeled "Retained form (marked)", even if there is (to my knowledge) no established alternative translation for the term in Swedish (see § Loanwords and Degrees of Entrenchment for further discussion). The full taxonomy is shown in Figure 3.

Finally, the offensiveness of the translated term as compared to that of the source term was graded on a three-point scale with the alternatives *More offensive – Roughly equivalent – Less offensive*. The level of offensiveness for a given term was assessed based on a hypothetical situation where the hearer is a member

[17] For this part of the analysis, all translations were compared to the slur sense of the ST term, regardless of whether it appeared as a slur or as a non-slur homonym in the original context.

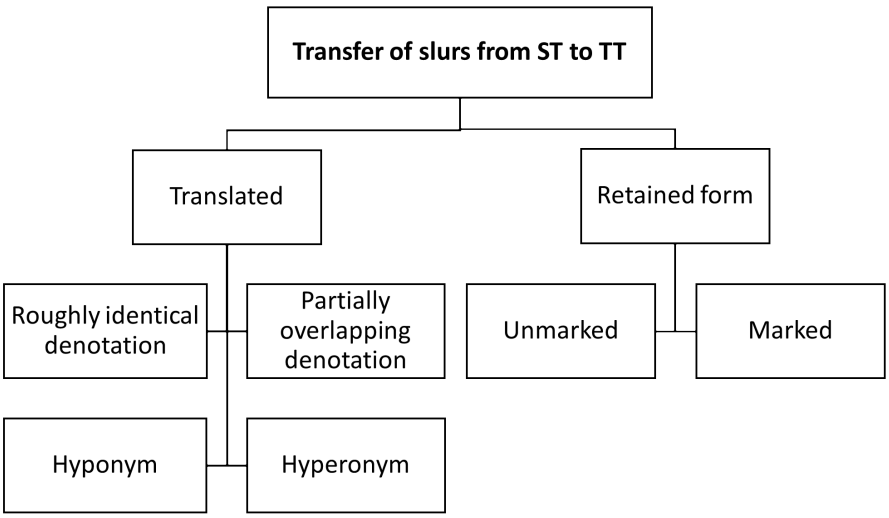

Figure 3: Taxonomy for classifying machine translated slurs.

of the target group and the term is uttered by a non-member. This further analysis was applied to all instances categorised as "Roughly identical", "Hyperonym", "Hyponym", "Partial overlap" or "Retained form (unmarked)", but not to those categorised as "Retained form (marked)". This was based on the reasoning that if an expression is not established in the target language, and thus not in the vocabulary of the target audience, the target audience will not have any preexisting intuitions about its offensiveness.

Results

Analysis of Matches

Of the 100K lines in the data, 164 (0.164%) contained tokens matching items from the cleaned list of discriminatory terms. With one line containing two matching tokens, the total number of matches amounted to 165. In the subsequent analysis, 73 (44%) of these were labeled false positives, leaving 92 actual slurs in the English material. In turn, out of the 755 items on the list of terms, only 39 (5%) yielded matches in the

Term	*n*	(% of total)	Term	*n*	(% of total)	Term	*n*	(% of total)
gypsy	29	(18%)	bung	2	(1%)	hunky	1	(<1%)
gay	28	(17%)	spig	2	(1%)	shyster	1	(<1%)
queer	17	(10%)	ubangi	2	(1%)	shysters	1	(<1%)
bitch	13	(8%)	saracen	2	(1%)	darkie	1	(<1%)
trash	10	(6%)	jiggers	2	(1%)	toff	1	(<1%)
gypsies	8	(5%)	bimbo	2	(1%)	batty	1	(<1%)
tinker	7	(4%)	slut	2	(1%)	ike	1	(<1%)
darkies	4	(2%)	lesbian	1	(<1%)	sooty	1	(<1%)
mock	4	(2%)	bitches	1	(<1%)	jigger	1	(<1%)
dyke	3	(2%)	gypped	1	(<1%)	pickaninnies	1	(<1%)
gyp	3	(2%)	papist	1	(<1%)	mongrels	1	(<1%)
blockhead	3	(2%)	spic	1	(<1%)	spick	1	(<1%)
okies	3	(2%)	nip	1	(<1%)	toffs	1	(<1%)

Table 2: Number of occurrences for each slur present in the material (total = 165).

material. The number of times each individual term appeared in the material is shown in table 2.

The fact that so few of the terms on the list occurred in the material suggests that tools adapted from use in hate speech detection may not in fact be optimal for identifying slurs in a corpus of subtitles. On the other hand, AVT applies to a wide range of genres, all of which are not necessarily represented in the OpenSubtitles corpus. It is possible that the list would be better suited for detecting sensitive language in a corpus mainly consisting of more contemporary material, such as reality shows.

Translation Error Analysis

The coarse analysis of the translated terms showed a rather high rate of severe errors for all three systems (25%, 35% and 28% respectively). Error counts and distribution by category for each system are reported in Table 3. Note that this is only a coarse categorisation of the more severe translation errors; it says little about the quality of the translations that were not categorised as errors.

	Google Translate	Amazon Translate	Microsoft Translator
WSD-error; slur to non-slur	11 (7%)	16 (10%)	11 (7%)
WSD-error; non-slur to slur	25 (15%)	38 (23%)	34 (21%)
Other mistranslations	6 (4%)	4 (2%)	-
Omission	-	-	1 (<1%)
Total errors:	42 (25%)	58 (35%)	46 (28%)

Table 3: Translation error count for each system. Percentages of the total number of matched tokens in the source text (n=165) are reported in parenthesis.

As can be seen in Table 3, WSD-errors where a non-slur was translated as a slur were by far the most common translation error as captured by this categorisation. Possible causes for and implications of this will be explored in § Discussion.

Semantic Analysis

On this level of the analysis, all tokens that had been transferred to the target segment as slurs were taken into account, regardless of whether the ST term was categorised as a slur or a non-slur homonym. This included terms whose form remained unchanged in the translation. The distribution of translations with regards to the semantic relationship between ST term and translation are shown in Figure 4.

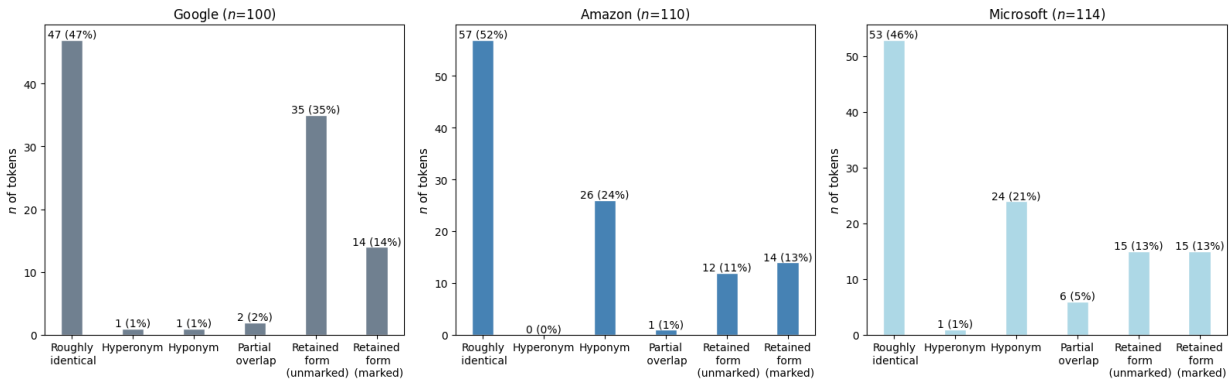

Figure 4: Denotation of the translated terms as compared to the ST terms (note that the total differs between the three systems, since they did not have the same number of WSD-errors).

As can be seen, the most common category for all three systems was “Roughly identical”. Apart from that, however, some differences between the systems can be observed. For example, the proportion of translated and retained tokens differ noticeably between Google and the other two systems. Google translated 51% of the tokens and retained 49%. The corresponding numbers for Amazon are 76% translated and 24% retained, and for Microsoft 74% translated and 26% retained. While the proportion of instances categorised as “Retained form (marked)” was similar for all systems, Google Translate had a noticeably larger proportion of instances in the “Retained form (unmarked)” category as compared to the other systems. On the other hand, Google Translate had a much lower proportion than either of the other systems in the “Hyponym” category. These differences will be expanded upon in § Discussion.

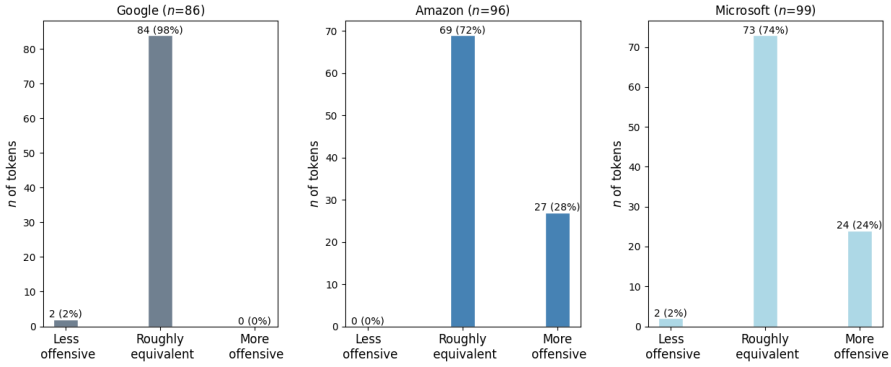

Figure 5: Offensiveness of translated terms as compared to the ST terms (once again the total differs between the three systems, since they did not have the same number of WSD-errors and retained words).

The final part of the analysis targeted the offensiveness of the translated term as compared to the source term. The results of this analysis can be seen in Figure 5. Once again, there is a noticeable difference between Google Translate and the other two systems, where both Amazon Translate and Microsoft translator have a stronger tendency to produce translations that are more offensive than the original. However, it should be kept in mind that offensiveness is a vague concept and that the perceived offensiveness of a term is bound to differ between individuals. Performing an analysis like this with only one human annotator unavoidably brings with it a certain degree of subjectivity. For this reason, while these numbers can give an inkling about how the offensiveness of a word may be affected in translation, they should be taken with a grain of salt. This, and other issues pertaining to this part of the study, will be discussed in § Offensiveness and Connotations.

Discussion

Using Tools From Hate Speech Detection to Identify Slurs in Subtitles

As shown in § Results, only 5% of the items on the manually cleaned list of slurs yielded matches in the data. This suggests that there is a mismatch between the type of language used in the OpenSubtitles corpus and the type of language for which the Weaponized Word lexicon was originally developed. Moreover, despite cleaning the list from words that were deemed too common in non-slur senses, 44% of the segments containing tokens from the list were labeled false positives. For example, the word *mock* appeared four times as a verb (in the sense *make fun of*) but never as a noun (as a derogatory term for a Jewish person). Another example is the word *queer*, which appeared 17 times in the sense *strange* but not once to denote a person of non-normative sexuality or gender identity (see § Word Sense Disambiguation and Semantic Change for further discussion of word senses).

It makes intuitive sense that movie subtitles, which are normally made up of scripted dialogue, do not contain

the same types of expressions used in online hate speech, or at least not with the same frequency. However, subtitling does not exclusively concern movies; unscripted content such as reality TV is also translated and subtitled, and likely to contain a larger variety of offensive expressions. Since the lexicon from which the list was extracted is made to be used on current online material, its accuracy in detecting actual instances of offensive language could be expected to increase the more recent the material. It is therefore possible that the list used in this study would be more efficient as a tool for identifying offensive speech in, for example, a corpus of recent reality TV subtitles. For the corpus used in this study, however, a smaller sample of unambiguous terms may have been better suited as a starting point for developing a consistent way of assessing how meaning is affected in translation.

Word Sense Disambiguation and Semantic Change

It was far more common for all three systems to translate a neutral word as a slur than the other way around. While this may seem remarkable at a first glance, it is less surprising when considering that there is very likely a discrepancy between the tested material and the majority of the data the systems have been trained on. Books, movies and other forms of fiction do not always take place in the present time, and fictional works depicting events from the past often use words and expressions that would be considered odd and archaic if uttered in a contemporary context. Thus, as a collection of subtitles from movies and TV series, the OpenSubtitles corpus contains examples of word uses that are rare in current English. For example, due to a semantic shift the use of the word *queer* as a slur or, in the reclaimed use, as a self-identifying term for persons with a non-normative sexuality and/or gender identity (Kolker et al., 2020) has almost completely replaced the older sense of *strange* or *unusual*, so that it is now near impossible to use the word in the second sense without also invoking the first (MacCabe & Yanacek, 2018). In the tested material, however, all appearances of *queer* were in the sense *strange*. Unsurprisingly, this presented a challenge to all three systems, since they were more familiar with the current use. Example (2) shows how one such segment was handled by the three systems.

(2) Source: She smiled in a **queer** sort of way.
Reference: Hon log på ett **konstigt** sätt.

Google Translate:	**Hon**	**log**	**på**	**ett**	**konstigt**	**sätt.**
	She	smiled	in	a	strange	way.
Amazon Translate:	**Hon**	**log**	**på**	**ett**	**bögigt**	**sätt.**
	She	smiled	in	a	faggoty	way.
Microsoft Translator:	**Hon**	**log**	**på**	**ett**	**queert**	**sätt.**
	She	smiled	in	a	*queer*	way.

As can be seen, only Google Translate managed to correctly disambiguate the sense of the target word in this sentence. The other two systems translated *queer* into *bögigt* (“faggoty” or faggot-like) and *queert*[18] respectively. Out of the 17 instances where *queer* appeared in the material in the sense *strange*, it was correctly disambiguated 10 times by Google Translate, twice by Microsoft Translator, and never by Amazon Translate.

[18] Queer, in the currently more common sense of the word, is established as a loan word in Swedish. Here it is correctly inflected to agree with the gender of the noun.

The word *gay* has gone through a similar shift as the word queer, from being synonymous with *happy* and *lively* to being more common as a synonym of *homosexual* (Rosenfeld & Erk, 2018). While the word appeared in both senses in the material, the *happy* sense was somewhat more common (17 instances out of 28). This too caused difficulties for all three systems; in fact, neither of the systems managed to correctly disambiguate the *happy* sense of gay.

(3) Source: And the song was bright and **gay**.
Reference: Flickan var förtjusande och visan var **glad**.

Google Translate:	**Och**	**sången**	**var**	**ljus**	**och**	**gay.**
	And	the song	was	bright	and	*gay*.
Amazon Translate:	**Och**	**låten**	**var**	**ljus**	**och**	**homosexuell.**
	And	the tune	was	bright	and	homosexual.
Microsoft Translator:	**Och**	**låten**	**var**	**ljus**	**och**	**gay.**
	And	the tune	was	bright	and	*gay*.

In Example (3), two of the systems retained the form *gay* in the translation, while the third translated it to *homosexuell* (homosexual). Since *gay* in the sense *homosexual* is an established loan word in Swedish, the most obvious analysis is to regard this retained form as a disambiguation error rather than simply an untranslated word. This will be further discussed in § Loanwords and Degrees of Entrenchment below.

Translations of non slur homonyms into slurs mainly concerned the words *queer* and *gay*, words which have both undergone drastic semantic change. In contrast, the instances where slurs were translated as non-slurs, while fewer in number, applied to a greater variety of source text terms. Two such examples can be seen in (4) and (5) below.

(4) Source: You won't fight over that **bitch**!
Reference: Ska ni slåss för det där **kräket**? Kräket?

Google Translate:	**Du**	**kommer**	**inte**	**att slåss**	**om**	**den där**	**tiken!**
	You (sg)	will	not	fight	over	that	female dog!

(5) Source: **Darkies'** pay would break us.
Reference: **Svarta arbetare** är dyra.

Amazon Translate:	**Mörkrets lön**	**skulle**	**knäcka**	**oss.**
	The pay of darkness	would	break	us.

Since different senses of a word often differ radically, incorrectly translated ambiguous words may result in translations that are incomprehensible or misleading (Marvin & Koehn, 2018). However, this very incomprehensibility may in fact make WSD errors a less serious problem in translations that are to be post-edited

by humans, since humans presumably have access to enough contextual cues to notice that a translation error has occurred. As mentioned in § Machine Translation and Subtitling, a study by Farrell (2018) suggests that post-editors, when faced with an acceptable translation, tend not to edit it. In the section below, I will argue that more subtle changes in the translated meaning may in fact constitute a bigger problem for translations that are to be used in industry, since they run the risk of slipping through undetected by post-editors.

Transferring Meaning, and the Things That Get Dropped or Picked up Along the Way

Denotation

Two of the systems (Amazon and Microsoft) translated a large proportion of the ST terms into hyponyms (21% for Microsoft and 24% for Amazon). A closer look at the material shows that most of these cases, once again, concern the words *queer* and *gay*. Both systems show a strong preference for translating both of these terms as *bög*,[19] a word that can be roughly translated as *faggot*. As previously mentioned, the term *queer* applies to individuals with a non-normative sexuality and/or gender identity, while *gay* applies to homosexual individuals of any gender. *Bög*, on the other hand, is specifically used about gay men. While this group is indeed a subset of the groups denoted by both *gay* and *queer*, it is unlikely that they constitute more than approximately half of the former, and even less of the latter. Yet, they are strongly over-represented in the translations, even in cases where the source segment includes a female pronoun such as in Example (6).

(6) Source: She looked ill, **queer**.
Reference: Hon såg sjuk ut och plötsligt reste hon sig och kom emot mig.[20]

Google Translate:	**Hon**	**såg**	**sjuk ut,**	**konstig.**
	She	looked	sick,	strange.
Amazon Translate:	**Hon**	**såg**	**sjuk ut,**	**bög.**
	She	looked	sick,	faggot.
Microsoft Translator:	**Hon**	**såg**	**sjuk ut,**	**bög.**
	She	looked	sick,	faggot.

Disregarding for the moment the fact that *queer* in this context means *strange* rather than *person with a non-normative gender identity and/or sexuality*, we see in (6) an example of a term denoting a person of unspecified gender being translated as a term denoting a man. This is an example of gender bias, which is a serious problem in NLP (Leavy, 2018). Among LGBTQ groups, just like in society at large, men have been more visible than people of other genders, both historically (Campbell, 1998) and in more recent media (The Guardian, 2015). With this in mind, it is not unlikely that men have been over-represented in the training data fed to MT systems as the referents of words such as gay and queer, and that this may have created a bias in the translations.

[19] The adjective uses of *queer* and *gay* were generally translated as the adjective *bögig*, roughly "*faggoty*" or *faggot-like*.
[20] The reference translation of this segment does not include a translation of the target word.

Offensiveness and Connotations

As mentioned in § Results, offensiveness is a vague concept, and the perceived offensiveness of a term is bound to differ between individuals. Furthermore, offensiveness is merely part of the connotations of a word. Two words that cause a similar amount of offense might do so for different reasons, and bring with them different sets of connotations that affect the tone of the utterance as a whole.

(7) Source: You'd think a man of the cloth and a psychiatrist could put a pretty compelling case to the Lord for some **gay** rights.
Reference: En präst och en psykiatriker borde kunna prata **gay**rättigheter med Gud.

Google Translate: **Man skulle kunna tro att en man av tyget och en**
One could think that a man of the fabric and a
psykiater kan lägga ett ganska övertygande fall till Herren
psychiatrist can place a pretty convincing case to the Lord
för några homosexuella rättigheter.
for some homosexual rights.

The translation of *gay* into *homosexuell* is one example where a comparison of offensiveness is insufficient to describe the resulting divergence in meaning. In the US, the more clinical term *homosexual* has been, and still is, used by some conservative and Christian groups to evoke associations to pathology, and this seems to have strongly influenced the connotations of the term in American English (Smith et al., 2018). In Swedish there is no comparably visible, widespread practice of using the term *homosexuell* to deliberately evoke negative associations. Further, before the word *gay* was established as a loanword, options were scarce for neutral alternatives to the pejorative *bög* when referring to male homosexuality. Thus, the word *homosexuell* is generally regarded as neutral in Swedish, and in this analysis translations of *gay* to *homosexuell* were labeled as having "Roughly equivalent offensiveness". However, it should be noted that the connections to medical diagnoses and pathology are present in Swedish as well, even if they are not as widely exploited in political discourse. Recall that Hom (2008) regards the derogatory content of a slur to be tied to discriminatory practices in the linguistic community to which the speaker belongs (see § Conceptualising Slurs – Some Different Approaches). Viewed through this lens, using the term *homosexuell* about an individual could still be perceived as invoking the threat of pathologisation and discriminatory medical practices. The label "Roughly equivalent offensiveness" does in no way convey this substantial difference in connotations between *gay* and *homosexuell*.

Another example of a case where compared offensiveness does not satisfactorily cover the differences in usage is the word *gypsy*, consistently translated as *zigenare* by all three systems. Both *gypsy* and *zigenare* are used about members of the Romani people, but they carry different sets of connotations. In Swedish the term *zigenare*, which is highly derogatory and carries connotations of "otherness" and possible criminality, stands in contrast to the more neutral *rom*, which is the more widely used term and recommended by Svenska Akademiens Ordlista (SAOL, 2015). The English term *gypsy* has similar connotations as *zigenare*, but it is also used in a romanticising way to describe "a vagabond, free-spirited lifestyle" (NOW, 2017). While efforts have been made to introduce the term *Roma* as a non-pejorative alternative (similar to *rom* in Swedish), the term *gypsy* is still the most widely known and used by non-members of the target group (Challa, 2013), and many use it without being aware of the level of offense it causes (NOW, 2017). Thus, to an English speaker who is not a member of the target group, the ST segments containing the word *gypsy* are likely to come across as less offensive than the corresponding translations containing the word *zigenare* would to a Swedish speaker. To a member of the target group, however, they are both likely to be perceived

as highly offensive, and for that reason these translations were labeled “Roughly equivalent offensiveness”.

Both of these examples show that it is very hard, if not impossible, to completely disentangle the meaning of a slur from the culture of which it is a part. The case of the word *gypsy* also shows that conclusions about the offensiveness of a term may have to be made relative to some imagined hearer. These issues will be further discussed in § Separating M-Meaning and C-Meaning below.

Loanwords and Degrees of Entrenchment

When an expression was kept identical to the source text token in the translation, it was labeled “Retained form (unmarked)” if the token is an established word in the target language and “Retained form (marked)” otherwise. A large proportion of Google’s translations were in the category “Retained form (unmarked)” (35%, as compared to Microsoft’s 13% and Amazon’s 11%). In the category “Retained form (marked)”, the three systems were more or less equally represented. In other words, while the systems seem equally likely to simply retain an unknown word untranslated in the target text, Google seems to favor TL loanwords borrowed from the SL over TL-specific expressions in cases where both are available.

(8) Source: Maybe he was gonna announce that he was **gay**.
Reference: Han tänkte kanske tillkännage att han var **gay**?

Google Translate:	**Han**	**kanske**	**skulle**	**meddela**	**att**	**han**	**var**	**gay.**
	Maybe	he	was going	to announce	that	he	was	*gay*.
Amazon Translate:	**Han**	**kanske**	**skulle**	**meddela**	**att**	**han**	**var**	**bög.**
	Maybe	he	was going	to announce	that	he	was	a faggot.
Microsoft Translator:	**Han**	**kanske**	**skulle**	**meddela**	**att**	**han**	**var**	**bög.**
	Maybe	he	was going	to announce	that	he	was	a faggot.

For example, (8) shows how Google used the loanword *gay* where Amazon and Microsoft translated into the more offensive *bög* (faggot). This suggests that Google’s system is more up to date with recent loanword uses than both Amazon and Microsoft. Considering the company’s resources and prominence in the data mining field (Google Research, n.d.), it is hardly a surprising finding.

When it came to more unusual source text terms, the differences in performance between the three systems were smaller. To a large extent, the instances labeled “Retained form (marked)” concerned the same ST terms for all systems.

(9) Source: Them **Okies** got no sense and no feeling.
Reference: De har varken förnuft eller känslor.[21]

Google Translate:	**De**	***Okies***	**hade**	**ingen**	**mening**	**och**	**ingen**	**känsla.**
	They	*Okies*	had	no	meaning	and	no	feeling.

[21] The reference translation for this segment does not include a translation of the target term.

Amazon Translate:	***Okies***	**har**	**ingen**	**mening**	**och**	**ingen**	**känsla.**
	Okies	have	no	meaning	and	no	feeling.
Microsoft Translator:	***Okies***	**har**	**inget**	**vett**	**och**	**ingen**	**känsla.**
	Okies	have	no	sense	and	no	feeling.

Example (9) shows how the systems handled the term *okie*, which is a derogatory term for "a migrant agricultural worker; esp: such a worker from Oklahoma" (Oklahoma Historical Society). As can be seen, the term seems to have been unfamiliar to all three systems, and they have all retained the source text form of the word in the translation. Since the term *okie* is not an established loan word in Swedish, these translations were marked "Retained form (marked)". However, there is no corresponding Swedish word with the same denotation that would have constituted the "correct" translation. While a human translator might choose to translate the term as *Oklahoma-bo* (person who lives in Oklahoma), this would in no way convey the connotations or offensiveness of the original term.

"Retained form (marked)" is thus not an error category, but rather a way to differentiate between instances where a term exists in both the SL and the TL (as is the case for loanwords) and instances where the translated text introduces a new expression in the target language (as is the case for untranslated words). The reason is that to a speaker of the target language, these will be perceived differently; a loanword will likely already be part of the speaker's vocabulary, while an untranslated word will be unfamiliar and possibly require extra processing. However, this distinction between marked and unmarked uses of SL words is a simplification, since the establishment of loanwords in a language is a gradual process. Rather than fall into one of two distinct categories, terms that are in the process of being established as loanwords are more likely to fall somewhere on a scale from "completely foreign" to "completely established".

A similar process can be observed for metaphors, which develop from completely new and original metaphors to lexicalised or "dead" metaphors.[22] It is generally agreed that the force of a metaphor increases inversely with the degree of entrenchment, so that unfamiliar (and therefore more marked) metaphors convey a stronger imagery than conventional, commonly used metaphors (Pedersen, 2017b). It is possible that the retention of an unfamiliar term in a translation could serve a similar purpose by communicating more strongly that the group denoted by the word is specific to the source culture and therefore cannot be captured by any existing word in the target language. For example, there is no term in Swedish that corresponds to the term *okie*, since the concept denoted by *okie* is tied specifically to events in North American history. By using the original source text term instead of an existing target language term, the concept and the term are introduced simultaneously to the audience, who can then add it to their vocabulary. in this way, new terms start the process of becoming established as loan words.

Separating M-Meaning and C-Meaning

Since this study has been limited to what Bender & Koller (2020) call *conventional meaning*, context-specific factors such as communicative intent and speaker identity were not included in the analysis. However, when analysing real language, such as when working with corpus-data, it is hard to ever fully escape the influence of contextual factors. The results of the study should be viewed with this in mind. For example, in the case of reclaimed slurs, the attitudes communicated with the use of of a certain term can change drastically depending on the group membership of the speaker (Burnett, 2020; Hom, 2008; Bolinger, 2017). In this

[22] "Dead" metaphors are those that are so entrenched in the language that the average speaker is hardly conscious of the metaphorical origin of the expression. One example of a dead metaphor is the use of the spatial preposition *past* to talk about time, as in *half past nine*.

study, I did not take the identity of the speaker into account when determining the offensiveness of a specific utterance; instead, for each occurrence of a slur the speaker was assumed to be a non-member and the hearer a member of the target group. This means that reclaimed uses of slurs are not addressed by my analysis. It is highly likely that different assumptions about the context of the target words would have yielded different results for some of the points of analysis. Any conclusions drawn about how the tested systems handle the target phenomenon must, then, be drawn in relation to this imagined context.

The fact that contextual factors such as speaker identity affect the interpretation of a slur (Burnett, 2020) of course has implications for the translation process. A human translator may need to make different decisions about how to translate a specific term depending on who is uttering it and the perceived intent behind the utterance. For example, a word that has both a slur use and a reclaimed use in the source language may translate into different words in the target language depending on whether it is used in the reclaimed sense by a target group member, or as an insult by a non-member. The intended audience of the translated content may also play a part in these decisions, since different groups sometimes have different intuitions about the offensiveness of a term (the term *gypsy* is an example, see § Offensiveness and Connotations).

As mentioned in § Machine Translation and Subtitling, some of the contextual information needed to interpret an utterance in audiovisual media may only be available in the visual image. An MT system that only takes text as input is unlikely to have access to that information. Thus, unless there are other, purely linguistic clues to speaker identity (and the system manages to pick up on them), the system's translation of slurs should be independent of who the speaker is. Reclaimed uses of slurs would then be translated the same way as derogatory uses by non-members of the target group. To test this hypothesis, and to determine which speaker the system is implicitly assuming in these cases, would be an interesting topic for future studies.

DTS and MT Evaluation

Unlike conventional evaluation of MT output, which focuses on fluency and adequacy of the output text, the approach used here is inspired by those used in DTS for analysis of human translation. As such, it aims to capture descriptively the ways a translation can relate to its source text on a lexical-semantic level, rather than to provide quality judgements. This has proven to be a useful way to look at human translation. The results of this study suggest that it could also be of use in fine-grained analyses of MT output, in cases where the focus is a specific phenomenon of interest.

As a branch of study focused on human translation, the DTS framework of course assumes a human intent behind each choice of wording in the target text. The decisions made in response to difficult translation problems, and the norms of behavior that govern these decisions, are the very object of study in DTS. An MT system, on the other hand, has no communicative intent and does not make conscious decisions (see § Machine Translation and Meaning). Nevertheless, a human reading a translated text is likely to construe a communicative intent behind the content whether the decisions therein are made by a human translator or an MT system (Bender et al., 2021). Furthermore, since MT systems are trained on already existing translations, they too could be said to be affected by translation norms. If one, like Toury (1995/2012), considers translations as facts of the target culture and analyses them as such, the methods used need not necessarily differ between analyses of human translations and MT output.

On the other hand, translations made by an MT system do differ somewhat from those made by human translators, mainly by showing less variety and inventiveness in lexical choice (Farrell, 2018). When it comes to difficult cases like metaphors (Pedersen, 2017b), extralinguistic cultural references (Pedersen, 2007b) and slurs, some solutions used by human translators are not available to an MT system. For example, if they deem it necessary for the target audience to understand the intended meaning, human translators may intervene and add information that was not present in the source text (Pedersen, 2007b). This strategy

requires world knowledge that an MT system does not have. For that reason, a DTS taxonomy that is to be used on MT output may need to be adapted to better capture the translation alternatives available to MT systems specifically. Such a taxonomy may then be used to find patterns in the translation choices made by MT systems, and identify and measure differences between systems. The taxonomy I used in this study was an initial attempt at such an adaptation. I believe that further revision could achieve an even better fit between the taxonomy categories and the translation options available to MT systems.

Sensitive Language and Translation Quality

As mentioned in § Evaluating Meaning Transfer, the norms governing MT evaluation stem to some extent from the part evaluation plays in the development of MT systems. The overall goal, of course, is to create better machine translation systems. However, what constitutes "better" machine translation is rarely discussed or defined beyond the notion of similarity to human translation. Consequently, most evaluation metrics measure global quality in terms of similarity to human translations, without ever discussing what constitutes quality in human translations. This may have been less of a problem in times when the quality of state-of-the-art machine translation was still far from that of human translation, and may still not be a problem in metrics used for quick evaluations during the training process. However, the more the quality of MT improves, the more important it becomes to define what is meant by quality, to determine which problems need to be solved in order to improve it.

It is reasonable to assume that some aspects of a translation affect the overall quality more than others. For example, Marvin & Koehn (2018) argue that WSD is crucial for MT quality, since "[i]f the NMT systems do not correctly translate ambiguous words, the resulting translations could be incomprehensible or misleading." While some research on MT has targeted specific phenomena that are perceived as particularly problematic, such as WSD (Carpuat, 2013; Marvin & Koehn, 2018; Liu et al., 2018, among others), there is, to my knowledge, no previously existing research on how MT handles slurs and other sensitive language. Yet, it could be argued that the translation of sensitive language has a significant effect on the overall translation quality, since small nuances in meaning can have a large impact on the perceived hostility of an utterance, and may even result in emotional damage where none was intended.

According to Pym (2002), translation is first and foremost a mode of linguistic mediation. Seen in that light, he argues, the main task of Translation Studies becomes that of solving social problems (§ DTS and Ways of Classifying Translation strategies). By looking into how available NMT systems handle slurs, I have gained some insights into how their performance may affect marginalised communities by perpetuating biases and discriminatory language present in the training data. As discussed in § Transferring Meaning, and the Things That Get Dropped or Picked up Along the Way, this can result in misleading and offensive translations of texts that were not originally meant to offend. Some of the issues discussed, like gender bias (see § Denotation), are already well-known problems to MT researchers (e.g. Vamvas & Sennrich, 2021). Others, like differing connotations of expressions with the same denotation, and the effect of speaker identity on the interpretation of an expression (§ Offensiveness and Connotations), are largely unexplored in MT research, even as they are increasingly recognised in other areas of NLP such as hate speech detection (e.g. Palmer et al., 2020).

In the EU guidelines on ethics in artificial intelligence, it is stated that "the social impacts of these systems (i.e. on people's physical and mental wellbeing) must be monitored and considered" (Madiega, 2019). Similarly, (Bender et al., 2021) emphasise that " research and development of language technology, at once concerned with deeply human data (language) and creating systems which humans interact with in immediate and vivid ways, should be done with forethought and care", and advocate for "research that centers the people who stand to be adversely affected by the resulting technology, with a broad view on the possible ways that technology can affect people" (Bender et al., 2021, p.619). With this in mind, it seems appropriate to start taking ethical considerations into account when estimating translation quality. To that

purpose, a fine-grained analysis like the one used here could function as a kind of diagnostic for specific issues.

Limitations

The aim of this study was exploratory, and for that reason the net was thrown very wide. These findings should therefore be seen as pointers to potential areas for more focused investigations in the future – on their own, they are not sufficient basis for any definite conclusions about general patterns in machine translation of slurs. Some limitations to this study are discussed below, along with suggestions for future research:

Data. This study only used a small part of the OpenSubtitles corpus, and the lines were not randomised. This means that the content of the sample is likely to depend on the specific movies encountered in the beginning of the corpus. This could, for example, explain the prominence of the word *gypsy* in the data. Future work should use a larger and preferably randomised sample. It would also be interesting to perform a similar study on data containing more contemporary language, such as reality TV subtitles.

Language pairs. This study only considered one language pair, and only translations in one direction (from English to Swedish). Future studies should investigate whether the trends observed in translations from English to Swedish can also be observed translations from Swedish to English, and whether the same holds for other language pairs.

Target-words. This study targeted a large amount and variety of slurs in order to explore whether any general patterns could be identified. However, because of the variety of the terms, and the potential mismatch between the type of language used in the OpenSubtitles corpus and the type of language for which the Weaponized Word lexicon was developed, it is hard to make any definite claims. In retrospect, it may have have been more suitable to limit the focus to a smaller set of target terms in a larger sample of data. Future studies could then continue to explore whether the identified patterns extend to other terms as well.

Subjectivity in annotation. The categorisation of the translated target tokens were all done by me, and are therefore dependent to some degree on my own intuitions about the denotation and offensiveness of each expression. Employing multiple annotators in future studies may mitigate this problem to some extent. One possibility might be to recruit the help of members of the target communities themselves to access their knowledge about the terms used about them.

Context. This study was limited to what Bender & Koller (2020) call conventional meaning. However, it is hard to ever fully escape the influence of contextual factors when dealing with naturally occurring language (§ Separating M-Meaning and C-Meaning). Future research on the translation of slurs should therefore, if possible, take contextual factors into account as well. For example, in some cases a use of a slur meant to insult would translate into a different word than a reclaimed use of the same slur by a member of the target group. One interesting subject for future research would be to investigate if NMT systems can differentiate these uses based on linguistic cues alone, or if additional information is necessary.

Conclusion

This work has explored how DTS approaches can be adapted for analysis of MT output. Using a descriptive taxonomy of ways a translation can relate to its source text on a lexical-semantic level, I investigated how slurs are translated by commercial MT systems. The results provide some preliminary insights into the kinds of semantic divergence that can be introduced in the translation process when using MT to translate slurs. These can be seen as pointers to areas for future investigations. They also illuminate some potential issues that may be of interest to end-users of MT systems, such as translators and post-editors who use MT as a tool in their work. For example, some instances of gender bias were identified, where systems tended to translate terms denoting individuals of any gender into terms specifically denoting men. Additionally, two of the tested systems showed a tendency to translate into more offensive terms.

The way sensitive language is translated can have a significant effect on the overall translation quality, since small nuances in meaning can have a large impact on the perceived hostility of an utterance, and may result in emotional damage where none was intended. The taxonomy used here makes it possible to capture some (though not all) of these nuances in a descriptive way. While further revision may be needed for the taxonomy to better fit the material to be analysed, this initial investigation suggests that DTS approaches could be well suited as the basis for more fine-grained analyses of MT output.

References

Alexandra, R. (2017). The other f-word: How homophobic language has ruined '80s teen movies. https://www.kqed.org/pop/97337/the-other-f-word-how-homophobic-language-has-ruined-80s-teen-movies. KQED.

Ashwell, L. (2016). Gendered slurs. *Social Theory and Practice*, 42(2), 228–239.

Bach, K. (2018). Loaded words: On the semantics and pragmatics of slurs. In D. Sosa (Ed.), *Bad Words: Philosophical Perspectives on Slurs* (pp. 60–76). Oxford: Oxford University Press.

Banerjee, S. & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* (pp. 65–72). Ann Arbor, Michigan: Association for Computational Linguistics.

BBC (2020). BBC receives 18,600 complaints over use of racial slur in news report. https://www.bbc.com/news/entertainment-arts-53676557. BBC.

Belinkov, Y., Màrquez, L., Sajjad, H., Durrani, N., Dalvi, F., & Glass, J. R. (2017). Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1–10). Taipei, Taiwan: Asian Federation of Natural Language Processing.

Bellés-Calvera, L. & Caro Quintana, R. (2021). Audiovisual translation through NMT and subtitling in the Netflix series 'Cable Girls'. In *Proceedings of the Translation and Interpreting Technology Online Conference* (pp. 142–148). Held Online: INCOMA Ltd.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21 (pp. 610–623). New York, NY, USA: Association for Computing Machinery.

Bender, E. M. & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Online: Association for Computational Linguistics.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Blanc, S. (2022). Translating swear words: F***ing challenge ! https://correspondent.afp.com/translating-swear-words-fing-challenge. AFP Correspondent.

Bolinger, R. J. (2017). The pragmatics of slurs. *Noûs*, 51(3), 439–462.

Burnett, H. (2020). A persona-based semantics for slurs. *Grazer Philosophische Studien*, 97(1), 31–62.

Bywood, L., Georgakopoulou, Y., & Etchegoyhen, T. (2017). Embracing the threat: machine translation as a solution for subtitling. *Perspectives*, 25, 1–17.

Campbell, K. (1998). Deviance, inversion and unnatural love: Lesbians in Canadian media, 1950-1970. *Atlantis Journal*, 23(1).

Carpuat, M. (2013). A semantic evaluation of machine translation lexical choice. In *Proceedings of the Seventh Workshop on Syntax, Semantics and Structure in Statistical Translation* (pp. 1–10). Atlanta, Georgia: Association for Computational Linguistics.

Carpuat, M., Vyas, Y., & Niu, X. (2017). Detecting cross-lingual semantic divergence for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation* (pp. 69–79). Vancouver: Association for Computational Linguistics.

Challa, J. (2013). Why being 'gypped' hurts the Roma more than it hurts you. https://www.npr.org/sections/codeswitch/2013/12/30/242429836/why-being-gypped-hurts-the-roma-more-than-it-hurts-you?t=1649749369010. NPR Code Switch, December 30, 2013.

Denkowski, M. & Lavie, A. (2010). Choosing the right evaluation for machine translation: An examination of annotator and automatic metric performance on human judgment tasks. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers* Denver, Colorado, USA: Association for Machine Translation in the Americas.

Diaz-Cintas, J. (2012). Clearing the smoke to see the screen: Ideological manipulation in audiovisual translation. *Meta: Journal des traducteurs*, 57, 279–293.

Emerson, G. (2020). What are the goals of distributional semantics? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7436–7453). Online: Association for Computational Linguistics.

Farrell, M. (2018). Machine translation markers in post-edited machine translation output. In *Translating and the Computer 40: proceedings* (pp. 50–59).: Asling: International Society for Advancement in Language Technology.

Filmer, D. (2012). The 'gook' goes 'gay': Cultural interference in translating offensive language. *inTRAlinea*, 14.

Google Research (n.d.). Research areas – data mining and modeling. https://research.google/research-areas/data-mining-and-modeling/. Accessed: May 26, 2022.

Grice, H. P. (1968). Utterer's meaning, sentence-meaning, and word-meaning. *Foundations of language*, 4(3), 225–242.

Hämäläinen, M. & Alnajjar, K. (2021). The great misalignment problem in human evaluation of NLP methods. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)* (pp. 69–74). Online: Association for Computational Linguistics.

Harris, Z. S. (1954). Distributional structure. *WORD*, 10(2-3), 146–162.

Hom, C. (2008). The semantics of racial epithets. *Journal of Philosophy*, 105(8), 416–440.

Kolker, Z. M., Taylor, P. C., & Galupo, M. P. (2020). "As a sort of blanket term": Qualitative analysis of queer sexual identity marking. *Sexuality & Culture*, 24(5), 1337–1357.

Leavy, S. (2018). Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In *2018 IEEE/ACM 1st International Workshop on Gender Equality in Software Engineering (GE)* (pp. 14–16).

Lison, P. & Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 923–929). Portorož, Slovenia: European Language Resources Association (ELRA).

Liu, F., Lu, H., & Neubig, G. (2018). Handling homographs in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 1336–1345). New Orleans, Louisiana: Association for Computational Linguistics.

Logsdon, G. (n.d.). Okie (term). https://www.okhistory.org/publications/enc/entry.php?entry=OK007#. in *The Encyclopedia of Oklahoma History and Culture*. Accessed: May 19, 2022.

MacCabe, C. & Yanacek, H. (2018). From "odd," "strange," and "bad," to reclaiming the word "queer". On *OUPblog – Oxford University Press's Academic Insights for the Thinking World*. https://blog.oup.com/2018/12/lgbtq-community-reclaiming-the-word-queer/.

Madiega, T. A. (2019). Eu guidelines on ethics in artificial intelligence: Context and implementation. EPRS: European Parliamentary Research Service.

Martínez Pleguezuelos, A. (2020). Translating the gay identity in audiovisual media: The case of Will & Grace. *Revista Espanola de Linguistica Aplicada*, 34.

Marvin, R. & Koehn, P. (2018). Exploring word sense disambiguation abilities of neural machine translation systems. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)* (pp. 125–131). Boston, MA: Association for Machine Translation in the Americas.

Mathur, N., Baldwin, T., & Cohn, T. (2020). Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4984–4997). Online: Association for Computational Linguistics.

Matusov, E., Wilken, P., & Georgakopoulou, Y. (2019). Customising neural machine translation for subtitling. In *Fourth Conference on Machine Translation (WMT19)*.

Moylan, B. (2015). Most LGBT characters on US TV are white and male, study finds. https://www.theguardian.com/tv-and-radio/2015/oct/27/most-lgbt-characters-on-us-tv-are-white-and-male-study-finds/. The Guardian, 2015-10-27.

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., & Liang, X. (2018). doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Naomi P. (2017). The "G" word isn't for you: How "Gypsy" erases Romani women. https://now.org/blog/the-g-word-isnt-for-you-how-gypsy-erases-romani-women/. NOW Blog, National Organization for Women, October 2, 2017.

Netflix (2022a). Timed text style guide: General requirements. https://partnerhelp.netflixstudios.com/hc/en-us/articles/215758617-Timed-Text-Style-Guide-General-Requirements. Accessed: February 25, 2022.

Netflix (2022b). Translation - offensive translation error (text). https://partnerhelp.netflixstudios.com/hc/en-us/articles/360050602953-Translation-Offensive-Translation-Error-Text-. Accessed: February 28, 2022.

Palmer, A., Carr, C., Robinson, M., & Sanders, J. (2020). COLD: Annotation scheme and evaluation data set for complex offensive language in english. *Journal for Language Technology and Computational Linguistics: Special Issue on Offensive Language*, 34(1), 1–28.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Pedersen, J. (2007a). Cultural interchangeability: The effects of substituting cultural references in subtitling. *Perspectives, studies in translatology*, 15(1), 30–48.

Pedersen, J. (2007b). *Scandinavian subtitles : a comparative study of subtitling norms in Sweden and Denmark with a focus on extralinguistic cultural references*. Stockholms universitet. Engelska institutionen. Doctoral thesis.

Pedersen, J. (2016). In Sweden, we do it like this: On cultural references and subtitling norms. *inTRAlinea*. Special Issue: A Text of Many Colours – translating The West Wing. Edited by: Christopher Taylor.

Pedersen, J. (2017a). The FAR model: Assessing quality in interlingual subtitling. *The Journal of specialised translation*, (28), 210–229.

Pedersen, J. (2017b). How metaphors are rendered in subtitles. *Target : international journal of translation studies*, 29(3), 416–439.

Pedersen, J. (2018). From old tricks to Netflix: How local are interlingual subtitling norms for streamed television? *Journal of Audiovisual Translation*, 1(1), 81–100.

Poliak, A., Belinkov, Y., Glass, J., & Van Durme, B. (2018). On the evaluation of semantic phenomena in neural machine translation using natural language inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (pp. 513–523). New Orleans, Louisiana: Association for Computational Linguistics.

Pym, A. (2002). *Translation studies as social problem-solving*. Pre-print version 2.1. `https://usuaris.tinet.cat/apym/on-line/research_methods/thessaloniki`.

Rosenfeld, A. & Erk, K. (2018). Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 474–484). New Orleans, Louisiana: Association for Computational Linguistics.

Sahlgren, M. & Carlsson, F. (2021). The singleton fallacy: Why current critiques of language models miss the point. *Frontiers in Artificial Intelligence*, 4.

SAOL (2015). *Svenska Akademiens Ordlista över svenska språket*. Svenska Akademien, 14 edition. Accessed: May 16, 2022.

Schmidt, A. & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (pp. 1–10). Valencia, Spain: Association for Computational Linguistics.

Smith, B. A., Murib, Z., Motta, M., Callaghan, T. H., & Theys, M. (2018). "Gay" or "homosexual"? The implications of social category labels for the structure of mass attitudes. *American Politics Research*, 46(2), 336–372.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers* (pp. 223–231). Cambridge, Massachusetts, USA: Association for Machine Translation in the Americas.

Szarkowska, A., Díaz Cintas, J., & Gerber-Morón, O. (2020). Quality is in the eye of the stakeholders: what do professional subtitlers and viewers think about subtitling? *Universal access in the information society*, 20(4), 661–675.

The Weaponized Word (2022). How are the lexicons generated? FAQs – The Weaponized Word. https://weaponizedword.org/faqs/screen%3Dfaq%7Cfaq_id%3DumRtrwxt7/how-are-the-lexicons-generated. Accessed: May 26, 2022.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 2214–2218). Istanbul, Turkey: European Language Resources Association (ELRA).

Toury, G. (2012). *Descriptive translation studies – and beyond*. Benjamins Translation Library ; v.100. Philadelphia: John Benjamins Publishing Company, rev. ed. edition. (Original work published 1995).

Vamvas, J. & Sennrich, R. (2021). Contrastive conditioning for assessing disambiguation in MT: A case study of distilled bias. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 10246–10265). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Ávila Cabrera, J. (2016). The treatment of offensive and taboo language in the subtitling of reservoir dogs into spanish. trans 20: 25-40. *TRANS. Revista de Traductologia*, 20, 25–40.