

# JUST

JOURNAL OF LANGUAGE  
RIGHTS & MINORITIES

— — — —  
REVISTA DE DRETS  
LINGÜÍSTICS I MINORIES

---

## Should raw machine translation be used for public-health information? Suggestions for a multilingual communication policy in Catalonia

Anthony Pym 

Nune Ayvazyan 

Jonathan Prioleau 

### To cite this article:

Pym, Anthony, Nune Ayvazyan & Jonathan Prioleau. 2022. "Should raw machine translation be used for public-health information? Suggestions for a multilingual communication policy in Catalonia." In "Language Policies for Social Justice," edited by Christopher D. Mellinger & Esther Monzó-Nebot. Special issue, *Just. Journal of Language Rights & Minorities, Revista de Drets Lingüístics i Minories* 1 (1-2): 71-99.

<https://doi.org/10.7203/Just.1.24880>.



© 2022 JUST. JOURNAL OF LANGUAGE RIGHTS & MINORITIES, REVISTA DE DRETS LINGÜÍSTICS I MINORIES, VOL 1 (1-2), CÀTEDRA DE DRETS LINGÜÍSTICS - UNIVERSITAT DE VALÈNCIA.

## **Should raw machine translation be used for public-health information? Suggestions for a multilingual communication policy in Catalonia**

Anthony Pym, Nune Ayvazyan & Jonathan Prioleau

Universitat Rovira i Virgili, Spain

### **Abstract**

As Catalonia becomes increasingly multicultural as a result of immigration on many levels, official government communications are received by speakers of numerous different languages. For some languages, this is achieved through the employment of qualified translators. For the non-official languages, however, there is increasing reliance on the use of free online machine translation, explicitly without human correction (“post-editing”). Here we survey the use of Google Translate on the official website of the Catalan health service, focusing on COVID-19 vaccination information in 2021 and 2022. We identify the strategic advantages of machine translation and then survey the nature of the main errors made, some of which border on the incomprehensible and self-contradictory. It is proposed that a multilingual communication policy is needed for the mitigation of errors not just through the judicious use of translation memory software and efficient post-editing, but most especially by editing texts in such a way that the machine translation problems are solved before they appear (“pre-editing”). In the relative absence of policies for non-official languages, strategic planning is required in order to ensure that the benefits of machine translation can work towards a more inclusive society, rather than alienate users who merely see their language being abused.

**Keywords:** machine translation, post-editing, pre-editing, social inclusion

## 1. Introduction

The COVID-19 pandemic presented new challenges for many social practices, among them basic language policies. Traditional territorial policies would seek to ensure the rights of long-term social groups to official-language status over a temporal dimension measured in generations: in the case of Catalonia, this fundamentally concerns relations between Spanish, Catalan, and Aranese Occitan. In times of a pandemic, however, communication is needed with all the people actually living in a territory, no matter what their language, since collective well-being depends on actions such as limiting mobility, wearing masks, and accepting vaccination. Territorial policy thus has to address complex issues of mobility and inclusion (cf. Grin 2022). Only when *all* the people in the territory change their behavior can collective well-being be enhanced to the fullest extent. So what happens when traditionally territorial policy meets an urgent communication challenge that involves mobility and inclusion? Here we look at the communication practice of the Catalan government health service, particularly its public website, where the prime solution to this particular problem was to use raw machine translation output.

Within the field of healthcare communication, vaccination information is of particular interest in that it requires a high degree of trust on the part of the receiver of the communication (Pym 2020b; Pym & Hu 2022). The complexity of the raw medical information is such that the general public does not have direct access to it in terms of interpretative skills. The actual risks are thus very difficult for the individual to assess, and in this case the issue was further complicated by the circulation of conspiracy theories for all tastes (Gualda et al. 2021). There are few kinds of communication that are so dependent on the perceived trustworthiness of not only the message, but more particularly the sender of the message. We are thus particularly interested in the quality of texts that are visibly mediated by unedited machine translation. What kinds of errors are involved and how might those errors be mitigated so as to produce translations that are more trustworthy?

## 2. Previous research

Our own interest in these questions stems from data collected within the large research project *Mobility and Inclusion in Multilingual Europe* (2014–2018), where we were looking at the mediation strategies used by speakers of marginalized or minoritized languages. Asylum seekers were found to be using machine translation to access official information in detention centers in both Leipzig (Fiedler & Wohlfarth 2018, 279–280) and Ljubljana (Pokorn & Čibej 2018, 297–298). This typically involved short-term communication solutions such as looking up key terms prior to a visit to the doctor, thus enhancing communication in the host language, and as a tool for language learning. In our interview study of the Russian community in Tarragona–Salou (Ayvazyan & Pym 2018, 2022), we found that 78% of our 50 respondents reported using machine translation, especially the younger community members, even though they were generally aware of the possible errors. Such widespread use of machine translation is sprinkled with occasional comments that the resource could be used to check on human mediators who were not entirely trusted (Ayvazyan & Pym 2018, 350; Pym 2018, 261). Machine translation might make mistakes, but it offers relative independence (the user is in control of the input) and confidentiality (what you say might not be reported to an authority). Many of the asylum seekers come from countries where public officials are quite likely to be biased and might indeed operate as spies (Allaby 2018). User-initiated machine translation thus offers clear advantages over mediation via interpreters, for example, with respect to speed, cost, user independence and ostensible confidentiality (Pym 2018, 260–261). These virtues are to be counterbalanced by a lack of translation quality, which the user may or may not perceive adequately.

This pragmatic use of machine translation would appear to have become a general social phenomenon. It has been calculated that human translation accounts for less than 0.69% of the words translated in the world (Pym & Torres-Simón 2021, working from Wood 2018), although little is known about how well or badly all those users actually interact with the output. The need for some basic training in the area is underscored by growing attention to machine translation literacy, which would include knowing when and where *not* to use machine

translation and how to negotiate clear translation errors (cf. Bowker 2009, 2019; Bowker & Buitrago-Ciro 2019).

There is some evidence on the way machine translation is used in provider-initiated healthcare communication. A literature review by Al Shamsi et al. (2020) compares 14 studies on communication solutions in healthcare and finds that cost and time delays are major factors in medical consultations. They further conclude that online translation tools present viable solutions in some situations and can be combined with the provision of interpreting services. Flores (2005) reviews 36 studies on English-language health services and concludes that “multiple studies document the positive impact that both trained, professional interpreters and bilingual providers have on LEP [Limited English Proficiency] patients’ quality of care” (2005, 255). A large study by Lindholm et al. (2012) suggests that the savings of professional mediation are most pronounced when interpreters are present at admission only or at both admission and discharge; there may be no savings at all when there is a less targeted use of human mediation (cf. Wallbrecht et al. 2014). This suggests that in the more run-of-the-mill low-stakes medical encounters, there are situations when provider-initiated machine translation could provide workable solutions, alongside intercomprehension, medical staff who speak something of an L2, and volunteer mediators such as family members. As Al Shamsi et al. (2020) conclude, machine translation is indeed being used in low-stakes exchanges, although there is still little empirical evidence of its actual effects.

A similar literature review by O’mara and Carey (2019) looks at seven recent studies on government information for culturally and linguistically diverse communities in Australia. Although the survey generally finds that effective strategies mix translation and interpreting services with other communication strategies, the reviewed studies that included machine translation were almost exclusively restricted to the education field. There was very little empirical evidence on the actual effects of machine translation: “At present, it is not clear whether information technology is effective for translating government information” (2019, 19).

These studies are to be placed against a background of industry claims that parity has been reached between neural machine translation and human

translation (Hassan et al. 2018). That assessment is nevertheless based on assessment of the content (not form) of isolated sentence pairs, where non-professional users could not distinguish between the two with significant frequency. Real-life usage tends to concern texts rather than isolated sentences (Läubli, Sennrich & Volk 2018), and harm can come from actual translation errors rather than a receiver's incapacity to distinguish between human translation and machine translation.

Considerable academic attention has been paid to how translators correct raw machine translation output ("post-editing") (see, for example, volume 38 of the *Journal of Specialised Translation* in 2019), which would constitute a mode of machine translation literacy that requires specialized training to be carried out effectively (Nitzke, Hansen-Schirra & Canfora 2019). In a review of empirical studies prior to the advent of neural machine translation, Koponen (2016) concluded that post-editing *can* give results similar to professional human translation. More recent research, however, tends to find that the inclusion of machine translation in the work process has mixed but generally negative effects on the final translation quality, although it tends to offer some time gains (cf. the reports in Moorkens et al. 2018; Macken, Prou & Tezcan 2020).

In comparison, there are fewer empirical studies on the way texts can be edited especially for machine translation ("pre-editing"). This involves removing instances of "negative translatability indicators," in other words elements that are likely to be problematic for machine translation (O'Brien 2006, cf. Pym 2020a). There are general guidelines for pre-editing (see for example the "translation-friendly writing" outlined in Bowker & Buitrago 2019). The guidelines in many respects follow those of controlled language, which might be dated from Ogden's project for Basic English (1932), so on one level the principles are by no means new. Marzouk and Hansen-Schirra (2019) nevertheless report that their use of controlled language made no significant improvement to neural machine translation output between German and English. There are also other principles that seem more language-pair-specific (for example, Hiraoka & Yamada 2019 for Japanese-English). Ideally, specific indicators should be identified for specific domains and for particular language pairs or language families. Further, our informal classroom experiments over the years (Pym 2019, 333-334) suggest

(but do not prove) that pre-editing usually takes more time than post-editing, which implies that it is only worth doing when a given start text is to be translated into more than a few target languages.

Despite these various studies, there is widespread belief that machine translation should not be used for high-stakes texts. Even the most innocuous and well-intentioned errors can leave speakers of a language feeling relegated to an inferior status (cf. Angermeyer 2017; Bowker 2009, 147). Hale and Liddicoat (2015), writing just prior to the use of neural machine translation, claimed that machine translation was basically unsuited to situations where accuracy and cultural values were important, especially in healthcare. When press reports made it known that the Australian government used machine translation in the early stages of its COVID-19 messaging (Dalzell 2020), there was considerable outrage across the community of professional translators (cf. American Translators Association 2020). It became general Australian government practice to avoid machine translation in healthcare messaging.

Little did the Australians know that precisely the opposite communication solution was being adopted in Catalonia at the same time.

### **3. The GenCat website as an application of non-policy**

Language policy in Catalonia is squarely focused on the defense of Catalan in terms of historical territorial rights. The official status of Catalan is implied in Article 3 of the Spanish Constitution of 1978 and has been developed in Catalan legislation on language normalization in 1983 and on language policy in 1998. Public education adopts a basic policy of immersion in Catalan, with ongoing debates and legal decisions about the relative use of Spanish. There appears to be no policy specifically dealing with the provision of government services in non-official languages, although there is indirect mention in some provisions for bilingual education (Ali & Ready 2021). There is, however, a general policy trend towards the use of artificial intelligence in government services, and particularly towards online solutions in healthcare. The Digital Spain 2025 initiative (Ministerio de Asuntos Económicos y Transformación Digital 2022), embedded in the wider European agenda for digital transformation, includes goals such as “empowering

patients with telemedicine tools, self-diagnosis and *greater accessibility*" (emphasis ours). The use of automated language services is part of that vision.

Here we focus on the official website of the health services of the Catalan government, the Generalitat de Catalunya (GenCat). This is the site to which most municipal websites refer in Catalonia; it gives the official updates on all aspects of health services, from how to deal with mosquito bites through to how to survive COVID-19. The COVID-19 information is regularly updated in Catalan as the start text. A drop-down menu at the top of the page invites speakers of Spanish, English, and French to select their language and then see the page as translated by Google Translate. Speakers of other languages can do the same if they know how to go to the Google Translate site and insert the URL (133 languages are currently provided for by Google) but the GenCat page limits its healthcare menu to Spanish, English, and French.

Why should official healthcare communication be given in raw machine translation? Apart from the advantages of speed and cross-language coherence (since only one language version needs to be updated), a policymaker might argue that only a negligible percentage of the population actually need these translations. It is difficult to estimate the number of people concerned. Since the start texts (STs) for the translations are indicated as being always in Catalan, the users could be anyone in Catalonia who does not understand that language. According to official statistics for 2018, that number would be 512,068 people, some 6.6 percent of the population (Idescat 2018a). We note, however, that the official survey asked respondents whether they were "able to understand a conversation on an everyday topic" in Catalan (Idescat 2018a), which would be the operative definition of "understand" here. One doubts that official COVID information really counts as an everyday conversation, so the potential population could be considerably greater than 6.6 percent – enough to compromise attempts at universal vaccination.

Here we are not focusing on Spanish because the quality of the machine translation between Catalan and Spanish tends to be high: the two languages are highly cognate and the paired databases are very extensive (daily newspapers are machine translated from Catalan to Spanish). This leaves the English and French versions of the COVID-19 information provided by GenCat, which are the



machine translations made available to people in Catalonia who say they do not speak Spanish or Catalan. How numerous would those people be?

According to the official 2018 language survey, only 0.2 percent of the total population say they do not understand Catalan or Spanish (Idescat 2018b). This percentage is perhaps small enough for a policymaker somewhere to have dismissed it as not worth including in a developed communication strategy, even though it is still a sizeable population of 15,507. There are several reasons to suspect that the actual number is considerably greater.

First, once again, the 0.2 percent only represents the people who say they understand an *everyday conversation* in Spanish or Catalan. But healthcare directives are not everyday conversations.

Second, in 2021 there were 30,270 interpreting services rendered in the Catalan courts, almost twice the number of people that the language survey records as not speaking Spanish or Catalan (Suport Judicial 2021, 2). One notes that some 43 percent of those services were actually for Arabic (Suport Judicial 2021, 2, 8), for which the lingua franca on the GenCat website would tend to be French.

Third, the speakers of non-official languages form relatively transitory social groups, based on wealthy mobility from the North and economic migration from the South. This means their presence is likely not to be fully reflected in the official language surveys. These are also groups that tend to use English as a lingua franca, significantly among retirement immigrants (Gustafson & Laksfoss Cardoso 2017). In fact, the language survey itself indicates that many speakers of languages other than Spanish and Catalan turn to English as a second or third language: some 52 percent of non-Spanish citizens in Catalonia say that they understand English, as opposed to 44 percent for Spanish citizens (Idescat 2018b). This is another reason to suspect that the number of potential users of the raw machine translations is greater than 0.2 percent of the population.

#### 4. Methodology

Here our mission is to use text comparison to identify errors in the raw machine translations and to test post-editing and pre-editing as ways of mitigating those errors.

The GenCat website for COVID-19 information has been studied by our research students Marc Anguiano Musons in 2021 and Jonathan Prioleau in 2022. Both surveyed the entire website, although here we focus only on vaccination information. The 12-month difference between the two studies allows us to track certain changes made to the site, especially with respect to the correction (and mostly non-correction) of errors. Prioleau (2022), who was employed at the time as an interpreter at a clinic in the United States where COVID-19 vaccinations were given, also undertook post-editing of the site from Catalan to English, which we have compared with the raw machine translation available in 2022.

The information provided by these comparisons affords a general overview of the use of raw machine translation. In broad terms, the translations are surprisingly readable, to the extent that almost all the basic information could be understood correctly. The advances made in neural machine translation since 2016 are palpable. At the same time, however, there are errors that stand to impair the comprehension and sometimes the actionability of the texts. Our focus is on those errors.

A methodological difficulty here is the fact that not all errors are of the same magnitude and many have different consequences for different receivers, who have variable bilingual skills, health literacy, machine translation literacy, and culture-specific propensities to trust. We have conducted a small reception study with 19 subjects, using eye-tracking and questions about comprehension and trust with respect to a few of our examples, but the significant subject variables extend well beyond the scope of the present report. For example, in follow-up interviews, our university exchange students were particularly skilled at navigating around errors (they have good machine translation literacy) but approached the problems in very different ways: a German-speaking student was not too flustered by contradictions in the machine translation (“I would speak with friends to clear that up”) whereas a Japanese-speaking student assumed the errors were with her language skills and there was nothing wrong with the machine translation (“I trust it because it is from the government”). We leave those subject variables for another day. We nevertheless draw on that input implicitly in privileging the errors that are most obvious to both us and to our subjects, proceeding as would

a grammarian who fundamentally relies on their own internalized language competence.

We thus basically fall back on textual comparison to ascertain: 1) how human post-editing can repair errors (by comparing the 2022 raw machine translation with Prioleau), and 2) how minimal pre-editing can avoid some machine translation errors before they occur.

## 5. The variable need for post-editing

Here we present an overview of the main types of translation errors. Prioleau (2022) offers a bottom-up categorization of errors, which we adopt here. Since our interest is in how the errors might be avoided by post-editing and pre-editing, we offer illustrative examples rather than a quantitative analysis. Our presentation goes from the most obvious errors that concern actionability as well as comprehension, to those around which there is more scope for interpretation and some receivers can construe correct readings.

### *Untranslated images blocking actionability*

One of the guidelines for effective healthcare information is that images be used to lead readers through the text (Shoemaker et al., 2014). Unfortunately, machine translation will usually not render text that is embedded in images. In some cases, this can lead to a loss of actionability. In 2022, the GenCat website included an image with a text announcing information on who was eligible to receive a booster shot of the COVID-19 vaccine. Here is the image with the accompanying machine translated text in English:



Clearly, the translated text below the image did not make mention of the key linguistic information: that this announcement concerns the booster shot. For the English-language user, this instruction is entirely readable (a group has opened, somewhere) but wholly unactionable (what is it open for?). The recommendation for website managers is clear and simple enough: do not embed text in images. At best, the user will wonder what they are missing.

### *Omissions blocking actionability*

Neural machine translation can make minor omissions in order to smooth a text and enhance its readability. The omissions are often inconsequential, but not always. In the following case, an important negation goes missing:

ST: Fes gestions de forma no presencial amb el sistema sanitari.

MT: Make arrangements in person with the healthcare system.

Human PE: Access your healthcare system online.

In other words, the machine translation omission of “no” to qualify “presencial” effectively produced the opposite meaning, in this case with consequences for actionability.

A similar reversal occurs with the Catalan preposition “a,” which can mean “to” but also “at”:

ST: Descarrega el teu certificat COVID a La Meva Salut

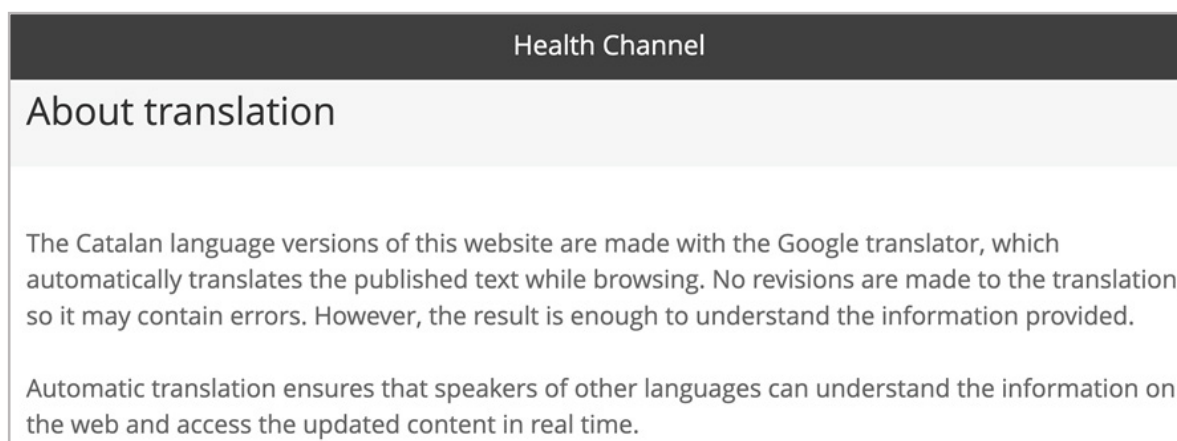
MT: Download your COVID certificate to My Health.

Human PE: Download your COVID certificate at My Health (La Meva Salut).

These are examples of cases where some degree of post-editing is essential, if only to have a pair of human eyes check the text and authorize it, as might a notary in the legal field.

### *Hallucinations with potential legal consequences*

Perhaps the most high-stakes translation error comes in the webpage where the use of machine translation is explained. Here is a screen shot from June 2022:



This disclaimer appropriately informs the user that the translation “may contain errors,” which might offer some degree of legal protection in the case of harm ensuing from misinterpretations. The text also identifies the legitimate advantages of the strategy: a basic understanding of the information and real-time updating of the contents. The translated text, however, refers to “the Catalan language versions of this website,” whereas the Catalan start text makes it clear that it should refer to all the versions *except the one in Catalan*: “Les versions en idiomes diferents del català d’aquest web es fan amb el traductor de Google” (italics ours). This could be one of those mysterious errors that are called “hallucinations” in research on neural machine translation (for example, Raunak et al. 2021), since the pathology is clear but the causes are not. One doubts the disclaimer will offer much legal protection.

### *Domain misidentification challenging comprehension*

Since Google Translate is designed for very general use, it can fail to identify the specific terms used in a given field. The most egregious example of this is the rendering of “dosi de record” as “record dose” and occasionally “memory dose,”

rather than the accepted English term “booster shot.” The machine translated names are simply not recognized by users who do not speak Catalan or Spanish.

Similar examples include the Catalan term “quarantena” (quarantine) rendered as “forties,” which it can also mean sometimes but not in this context: “If you are in your forties because you have become in close contact with someone with COVID-19 . . .” Another instance concerns the word “convocatòria,” which can elsewhere be translated as “call” but not here:

ST: Si ja tens administrada una primera dosi de la vacuna contra la COVID-19, descobreix com es duu a terme la convocatòria de la segona dosi.

MT: If you have already received a first dose of the COVID-19 vaccine, find out how the second dose is called.

Human PE: If you already have your first COVID-19 shot, find out how to get your second COVID-19 shot.

A rather more innocent example of the same problem is found in the subheading “Salut A-Z,” where the healthcare information is presented in alphabetical sections, from A to Z. The word “salut” has multiple meanings in Catalan, including not just “health” but also “cheers” or even “hello.” The machine translation unfortunately went for the last-mentioned option: “Hi AZ.” The website is apparently greeting an unknown interlocutor by the name of AZ.

In these examples, the use of machine translation leads to merely useless pieces of language. Although comprehension is made difficult, there is little risk of a false action being taken. A moderately socialized reader is unlikely to, assume that contact with COVID changes their age, fret over a secret name for a second vaccination, or wonder who AZ is. We would hope that basic machine translation literacy enables users to filter out these infelicities.

In all these cases, the problems could be avoided by preparing a field-specific glossary and/or translation memory, feeding it into any standard translation memory suite, then selecting settings that allow it to override the general machine

translation feed or otherwise instructing post-editors to apply the glossary, not the machine translation proposals. The same GenCat website lists several such glossaries under “recursos” (resources).

### *ST polysemy challenging comprehension*

There are also cases where the potential polysemy of the start text creates problems:

ST: Dispensadors de gel hidroalcohòlic

MT: Hydroalcoholic ice dispensers

This is a simple mistranslation that appeared in 2021 (Anguiano-Musons 2021, 16). The Catalan term “gel” can mean both “gel” and “ice,” and the machine translation preferred the latter. In 2022, the term was replaced on the Catalan website by “solucions hidroalcohòliques” (hydroalcoholic solutions) or “preparats hidroalcohòlics” (hydroalcoholic preparations), both of which successfully avoid the suggestion that people should be washing their hands with ice. But the cat was out of the bag. When we searched for “hydroalcoholic ice dispensers” in July 2022, we found some 67 hotels and public institutions that had repeated the same error, perhaps because they copied the official terminology. We nevertheless have no reports of people washing their hands with ice.

### *Complex grammar challenging comprehension*

A general rule of thumb in machine translation assessment is that the longer the sentence, the more likely the errors of syntactic reference. An example:

ST: Totes les persones, incloent-hi els infants, si són un cas positiu de COVID-19 o si estan realitzant una quarantena per contacte estret no es poden vacunar.

MT: Not all people, including children, can be vaccinated if they have a positive case of COVID-19 or if they are undergoing close contact quarantine.

Part of the problem here lies in the tortuous logic of the Catalan sentence, but the technical problem for the machine translation is that the negative particle “no” comes near the end of the sentence and the algorithms have a tough time knowing what it applies to. In principle, “Not all people can be vaccinated if they have a positive case” implies that some people who have been exposed to COVID-19 can, in fact, be vaccinated. This is not the case, as can be made clear in a human translation that applies considerable syntactic simplification:

Human PE: If you are COVID-19 positive or in quarantine due to close contact, you cannot be vaccinated. This includes children.

The remedy here is clear and well-known: avoid long sentences and keep the logic simple.

## 6. The discrete presence of human translations

Despite the overriding reliance on machine translations in this website, there are a few fully human translations to be found. Our search conducted in June 2022 required quite a few clicks to locate a well-translated brochure *The COVID-19 vaccination guide*. Here the medical terms are generally correct (“vaccination shot” instead of “dose,” for example). But when we looked for the term “booster shot,” it was nowhere to be found. The brochure was published in October 2021, so when we were looking for information in June 2022, it was woefully out of date. This serves to illustrate the major theoretical advantage of machine translation: updates are made to the Catalan text, and all other versions are updated automatically.

A second human translation was found in quite a different part of the website. Under information for “people from Ukraine with temporary protection,” we find a PDF with the basic Catalan information, an English translation that might



be human (it correctly refers to the “booster shot”), and a human translation into Ukrainian. All foreigners are equal, but some are more equal than others, it seems.

<b>Atenció sanitària a les persones d'Ucraïna amb protecció temporal</b>	<b>Охорона здоров'я громадян України в ситуації тимчасового захисту</b>	<b>Healthcare for people from Ukraine with temporary protection</b>
Si necessites atenció sanitària, tindràs la mateixa que tenim els ciutadans de Catalunya. L'atenció sanitària es gratuïta.	Ви маєте однакові з громадянами Каталонії права на охорону здоров'я та медичну допомогу. Медичне обслуговування є безкоштовним.	If you need healthcare, you will have the same coverage as Catalan people. Healthcare is free of charge.
També tindràs accés a les vacunes que hi ha al calendari de vacunes sistemàtiques i també a la vacuna contra la COVID-19, o la dosi de record, si ja la tens administrada.	Ви також маєте право на вакцинацію згідно з календарем систематичних щеплень, а також на вакцини проти COVID-19 або на бустерну дозу, якщо ви вже вакциновані.	You will additionally be able to get the vaccines on the routine vaccination schedule and also the Covid-19 vaccine, or alternatively the booster shot if you have already received it.

The intriguing thing about the texts here is that the last sentence is potentially nonsensical in English if read as follows: “you will be able to get [...] the booster shot if you have already received it.” The same apparent contradiction is in the Catalan – this is a case of challenging writing, not bad machine translation. But the Ukrainian here uses explication to make the meaning clear: it back-translates as “or the booster shot, if you are already vaccinated.” Here, as in many of the examples above, the basic cause of the problem is the way the Catalan start text is written. And that is something that could be changed.

## 7. The virtues of pre-editing

The above examples include the most common pitfalls of machine translation: domain-inappropriate terminology, lack of contextualization, grammatical confusion, pronoun misattribution and unwarranted smoothing, all of which is well-known. One could also argue, however, that the problems lie not with machine translation as such but with the way it has been used as a once-and-for-all solution. This concerns more than the kind of poor workflow management that results in untranslated images. It also has to do with not testing the translated website before publishing it. And it further ensues from not considering any of the many ways in which relatively simple technologies and workflows could have removed most of the errors we have just seen.

Most of the solutions are within easy reach. The most obvious and perhaps expensive kind of improvement involves translators regularly checking the machine translation output, doing post-editing before mistranslations are released to the population. A more cost-effective solution is to use translation memory software into which one has not only whatever machine translation feeds one wants but also authenticated propositions from all previous translations, within which the new updates stand out and can be focused on immediately. The software also includes domain-specific glossaries that can be set to override the proposals coming from raw machine translations, thus solving the problem of indiscriminate terminology. The translation automation association TAUS provided free translation memory databases at the beginning of COVID-19, including for Spanish–English, but there is no sign of them here. The political decision to use Catalan as a start language, understandable enough in political terms, meant that the extensive resources available for work between Spanish and English could not be used.

An alternative solution is to edit the start text in such a way that the MT problems are avoided before they appear. This basically involves simplifying text. It is technically called “pre-editing,” as opposed to “post-editing,” which is an intervention *after* the passage through machine translation. As noted, our classroom experiments find that pre-editing typically takes more time than post-editing and might thus appear less cost-effective. Its benefits, however, can automatically appear in *all* the languages into which machine translation is carried out. If one is going into three languages, as is the case here, pre-editing is likely to be more cost-effective than post-editing. And if it is done well, then there is no reason why machine translation should not be provided in many more languages as well, especially those that are relatively cognate.

To illustrate the virtues of pre-editing, here we use a little reverse engineering. In the examples below, we take the problematic cases we have seen in the start texts above (ST in Catalan), we give the raw machine translation with the problems indicated in italics (MT in English), we revise the Catalan input so as to avoid the machine translation problems (Revised ST in Catalan), and finally we present the raw Google Translate version of that revised text as it was rendered in July 2022 (New MT), without applying any special glossary or translation memory. Here we

also add the raw French MT of the original ST and the revised ST, to indicate that the problem can potentially be solved not just for English but for other languages as well (and bearing in mind that many speakers of Arabic, as mentioned, are likely to use French as a lingua franca). It is important to note here that this activity has been carried out bottom-up, working from the errors in order to remove their causes, rather than as a top-down application of an abstractly defined controlled language – one recalls that Marzouk & Hansen-Schirra (2019) found that controlled language had no significant effect in a similar situation. This is because we want to capture the kinds of solutions that can work in this particular domain, for this kind of text, and for more than one target language. After each case below, we nevertheless give a guideline that might apply more generally when healthcare messaging is being written.

And so to the examples:

ST: Descobreix com es duu a terme la convocatòria de la segona dosi.

MT-EN: Find out how the second dose is *called*.

MT-FR: Découvrez comment s'effectue l'appel pour la deuxième dose.

Revised ST: Descobreix quan i com pots rebre la segona dosi.

New MT-EN: Find out when and how you can receive the second dose.

New MT-FR: Découvrez quand et comment vous pouvez recevoir la deuxième dose.

Proposed guideline: Since the term “convocatòria” is problematically ambiguous, spell out what it means in this context. (Note that if we want “dosi” to be rendered as “shot,” we would have to add a specialized glossary or translation memory.)

ST: Salut A-Z

MT-EN: *Hi* A-Z

MT-FR: Santé de A à Z

Revised ST: Temes de salut de A a Z

New MT-EN: Health issues from A to Z

New MT-FR: Les thématiques santé de A à Z

Same principle: Explicitate the potentially ambiguous terms. Note that the first French MT was fine here but was not excessively harmed by the pre-editing.

ST: Fes gestions de forma no presencial amb el sistema sanitari.

MT-EN: Make arrangements *in person* with the healthcare system.

MR-FR: Prendre des dispositions *en personne* avec le système de santé.

Revised ST: Fes gestions en línia amb el sistema sanitari.

New MT-EN: Make arrangements online with the healthcare system.

New MT-FR: Prendre des dispositions en ligne avec le système de santé.

Proposed guideline: Opt for a more clearly distinguished term – admittedly at the risk of an Anglicism in this case, but that might be considered a valid trade-off.

ST: Totes les persones, incloent-hi els infants, si són un cas positiu de COVID-19 o si estan realitzant una quarantena per contacte estret no es poden vacunar.

MT-EN: *All people*, including children, if they are a positive case of COVID-19 or if they are undergoing close contact quarantine *cannot be vaccinated*.

MT-FR: *Toutes les personnes*, y compris les enfants, si elles sont un cas positif de COVID-19 ou si elles sont en quarantaine en raison d'un contact étroit *ne peuvent pas être vaccinées*.

Revised ST: Ningú que tingui COVID-19 o que faci una quarantena per contacte estret pot vacunar-se. Això també s'aplica als nens.

New MT-EN: No one who has COVID-19 or who is in close contact quarantine

can be vaccinated. This also applies to children.

New MT-FR: Personne qui a le COVID-19 ou qui est en quarantaine en raison d'un contact étroit ne peut se faire vacciner. Cela s'applique également aux enfants.

Proposed guideline: Write short sentences in order to avoid syntactic complexity. Here the obscurity is reduced by moving the embedded clause away from the verb.

ST: Descarrega el teu certificat COVID a La Meva Salut.

MT-EN: Download your COVID certificate *to* My Health.

MT-FR: Téléchargez votre attestation COVID à La Meva Salut

Revised ST: Descarrega el teu certificat COVID des de la web de La Meva Salut

New MT-EN: Download your COVID certificate from the My Health website.

New MT-FR: Téléchargez votre certificat COVID sur le site Ma Santé.

Same guideline: The ambiguous preposition has been replaced by one that is not so ambiguous.

ST: Si el teu fill o filla és un cas positiu de COVID-19 no es pot vacunar. Tampoc si està realitzant una quarantena per contacte estret.

MT-EN: If your son or daughter is a positive case of COVID-19 you cannot be vaccinated. *Nor if you are performing a quarantine by close contact.*

MT-FR: Si votre fils ou votre fille est un cas positif de COVID-19, ils ne peuvent pas être vaccinés. Ni si vous êtes mis en quarantaine pour contact étroit.

Revised ST: Un nen no es pot vacunar si té COVID-19 o si fa una quarantena per contacte estret.

New MT-EN: A child cannot be vaccinated if they have COVID-19 or if they are in quarantine due to close contact.

New MT-FR: Un enfant ne peut pas être vacciné s'il a le COVID-19 ou s'il est en quarantaine en raison d'un contact étroit.

Guidelines: Position the subject first and as close as possible to the verb; avoid pronouns.

ST: Tindràs accés a la vacuna contra la COVID-19, o la dosi record, si ja la tens administrada.

MT-EN: You will have access to the vaccine against COVID-19, *or the record dose, if you already have it administered.*

MT-FR: Vous aurez accès au vaccin contre le COVID-19, *ou à la dose de rappel, si vous l'avez déjà reçu.*

Revised ST: Tindràs accés a la vacuna contra la COVID-19. Si ja ets vacunat, tindràs accés a la dosi de record.

New MT-EN: You will have access to the vaccine against COVID-19. If you are already vaccinated, you will have access to the booster dose.

New MT-FR: Vous aurez accès au vaccin contre le COVID-19. Si vous êtes déjà vacciné, vous aurez accès à la dose de rappel.

Guidelines: Split long sentences; avoid pronouns. There is also a domain-specific trick here: "dosi record" gives "record dose," but "dosi de record" translates as "booster dose," which is more likely to be understood.

From such bottom-up reverse engineering, one can extract some quite logical desiderata: short sentences, no grammatical complexity, no pronouns, and explicitation in cases of potential ambiguity. These principles all appear in the general guidelines for pre-editing (for example, Bowker & Buitrago-Cirio 2019, 55–78), so we claim no novelty on that level. At the same time, though, there are language-specific and domain-specific changes that go beyond the general guidelines: small changes in some start-text wordings can successfully avoid items like "alcoholic ice" for washing hands and a download that sounds like an

upload. We thus propose that at least one round of pre-editing should be done with the MT target text in live view (as can indeed be done in Google Translate and DeepL).

Of course, it might be argued that some of these results tend to sound like they come from English. At the end of the day, though, the aim of this messaging is to save lives, perhaps at the expense of a few items of deceptively autochthonous syntactic complexity. When you write the original Catalan messaging in such a way that it is easy to understand and act upon, that is good for Catalan users as well, in addition to the benefits it brings for machine translations of use to end users in other languages.

## 8. Conclusions

The use of machine translation for time-sensitive healthcare information seems not to correspond to any major policy decision in this case: territorial policies simply look the other way. The practice is nevertheless not without some justification. Its advantages include immediate updating of information, reduced costs, general understandability of basic information, and perhaps a reasonable level of acceptance among users who have basic machine translation literacy, particularly in situations where the translation connects with further checking processes and spoken exchanges.

If one wanted to isolate only the translation errors, especially the more comical ones, it would be easy to condemn machine translation out of hand. One could be quite reasonably outraged and then declare that machine translation should never be used for official communication, especially in the case of supposedly actionable healthcare messaging. One might further claim that, even if machine translation does not actually infringe on language rights, in this case it certainly risks alienating a considerable number of residents in the territory, thereby losing trust, reducing behavior change, and effectively compromising public health. It nevertheless seems more judicious to view these linguistic problems as hazards that, first, stand in a trade-off relationship to the several clear advantages of machine translation, and second, appear in machine translation as just one element in a more complex and ongoing communication practice.

Our proposal here is not to exclude translation technologies, but to work with them in order to find solutions to their current shortcomings. Basic translation memory software can override terminological and phraseological errors, highlighting updates, and allowing reasonably quick post-editing. In this particular case, however, with translation into three or more languages, the most cost-effective solution is certainly pre-editing, or just a start text written in clear, simple language, produced by professionals who are properly trained to write it.

If there is one policy recommendation to be made with respect to the use of machine translation for official messaging, it would be to have a policy. Territorial language debates, not just in Catalonia but generally across Europe, tend to sideline non-official languages as a problem for the speakers of those languages, who need to learn the official languages. In the case of urgent healthcare information, however, those priorities no longer apply: there is no time for learning languages; behavior changes are needed quickly. It is there, under time pressure, that policy is needed to regulate situations where technology can come to the fore. Such a multilingual communication policy must be designed to protect lives, not just languages.

## References

- Ali, Farah & Carol Ready.** 2021. "Integration or assimilation? A comparative intertextual analysis of language policy in Madrid and Catalonia." *International Journal of Language and Law* 10: 24–47. <https://doi.org/10.14762/jll.2021.024>.
- Al Shamsi, Hilal, Abdullah G. Almutairi, Sulaiman Al Mashrafi & Talib Al Kalbani.** 2020. "Implications of language barriers for healthcare: A systematic review." *Oman Medical Journal* 35 (2): article e122. <https://doi.org/10.5001/omj.2020.40>.
- Allaby, Elaine.** 2018. "Are Eritrea government spies posing as refugee interpreters?" *Al Jazeera*, 26 February, 2018. Accessed 14 July 2022. <https://www.aljazeera.com/features/2018/2/26/are-eritrea-government-spies-posing-as-refugee-interpreters>.
- American Translators Association.** 2020. "Australian Government used Google Translate for COVID-19 messaging." Accessed 14 July 2022. <https://www.https://doi.org/10.7203/Just.1.24880>.



[atanet.org/industry-news/australian-government-used-google-translate-for-COVID-19-messaging/](https://atanet.org/industry-news/australian-government-used-google-translate-for-COVID-19-messaging/).

**Angermeyer, Philipp S.** 2017. "Controlling Roma refugees with 'Google-Hungarian': Indexing deviance, contempt, and belonging in Toronto's linguistic landscape." *Language in Society* 46 (2): 159–183.

<https://doi.org/10.1017/S0047404516001020>.

**Anguiano-Musons, Marc.** 2021. Translations of COVID-19 health information in Catalonia. BA thesis. Tarragona: Universitat Rovira i Virgili.

**Ayvazyan, Nune & Anthony Pym.** 2018. "Mediation choice in immigrant groups. A study of Russian speakers in southern Catalonia." *Language Problems and Language Planning* 42 (3): 344–364. <https://doi.org/10.1075/lplp.00027.ayv>.

**Ayvazyan, Nune & Anthony Pym.** 2022. "Portraying linguistic exclusion. Cases of Russian speakers in the province of Tarragona, Spain." In *Advances in interdisciplinary language policy*, edited by François Grin, László Marác & Nike Pokorn, 237–256. Amsterdam: John Benjamins.

<https://doi.org/10.1075/wlp.9.12ayv>.

**Bowker, Lynne.** 2009. "Can machine translation meet the needs of official language minority communities in Canada? A recipient evaluation." *Linguistica Antverpiensia* 8: 123–155. <https://doi.org/10.52034/lanstts.v8i.248>.

**Bowker, Lynne.** 2019. "Machine translation literacy as a social responsibility." Proceedings of the UNESCO International Conference on Language Technologies for All, Paris, 5 December 2019.

<https://lt4all.elra.info/media/papers/O7/145.pdf>.

**Bowker, Lynne & Jairo Buitrago Ciro.** 2019. *Machine translation and global research: Towards improved machine translation literacy in the scholarly community*. Bingley: Emerald. <https://doi.org/10.1108/9781787567214>.

**Dalzell, Stephanie.** 2020. "Federal Government used Google Translate for COVID-19 messaging aimed at multicultural communities." *ABC News*. Accessed 14 July 2022. <https://www.abc.net.au/news/2020-11-19/government-used-google-translate-for-nonsensical-COVID-19-tweet/12897200>.

**Fiedler, Sabine & Agnes Wohlfarth.** 2018. "Language choices and practices of migrants in Germany. An interview study." *Language Problems and Language Planning* 42 (3): 267–287. <https://doi.org/10.1075/lplp.00023.fie>.

- Flores, Glenn.** 2005. "The impact of medical interpreter services on the quality of health care: A systematic review." *Medical Care Research and Review* 62 (3): 255–299. <https://doi.org/10.1177/1077558705275416>.
- Grin, François.** 2022. "Principles of integrated language policy." In *Advances in interdisciplinary language policy*, edited by François Grin, László Marác & Nike Pokorn, 24–52. Amsterdam: John Benjamins. <https://doi.org/10.1075/wlp.9.02gri>.
- Gualda, Estrella, Andre Krouwel, Marisol Palacios-Galvez, Elena Morales-Marente, Ivan Rodriguez-Pascual & E. Begona Garcia-Navarro.** 2021. "Social distancing and COVID-19: Factors associated with compliance with social distancing norms in Spain." *Frontiers in Psychology* 12. <https://doi.org/10.3389/fpsyg.2021.727225>.
- Gustafson, Per & Ann Elisabeth Laksfoss Cardozo.** 2017. "Language use and social inclusion in international retirement migration." *Social Inclusion* 5 (4): 69–77. <https://doi.org/10.17645/si.v5i4.1133>.
- Hale, Sandra & Anthony Liddicoat.** 2015. "The meaning of accuracy and culture, and the rise of the machine in interpreting and translation." *Cultus: The Journal of Intercultural Mediation and Communication* 8: 14–26.
- Hassan, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang & Ming Zhou.** 2018. "Achieving human parity on automatic Chinese to English news translation." *arXiv*: 1–25. Accessed 14 July 2022. <https://doi.org/10.48550/arXiv.1803.05567>.
- Hiraoka, Yusuke & Masaru Yamada.** 2019. "Preediting plus neural machine translation for subtitling: Effective pre-editing rules for subtitling of TED talks." In *MT Summit 17*, edited by Mikel Forcada, Andy Way, Barry Haddow & Rico Sennrich, 64–72. Dublin: European Association for Machine Translation. Accessed 14 July 2022. <https://aclanthology.org/W19-6710.pdf>.
- Idescat.** 2018a. "Coneixement de llengües." Enquesta d'usos lingüístics de la població. Accessed July 2022. <https://www.idescat.cat/indicadors/?id=anuals&n=10367>.

- Idescat.** 2018b. "Població segons coneixement de l'anglès i nacionalitat." Enquesta d'usos lingüístics de la població." Accessed 14 July 2022. <https://www.idescat.cat/pub/?id=eulp&n=4658>.
- Koponen, Maarit.** 2016. "Is machine translation post-editing worth the effort? A survey of research into post-editing and effort." *JoSTrans: The Journal of Specialised Translation* 25: 131–147. [http://www.jostrans.org/issue25/art\\_koponen.pdf](http://www.jostrans.org/issue25/art_koponen.pdf).
- Läubli, Samuel, Rico Sennrich & Martin Volk.** 2018. "Has machine translation achieved human parity? A case for document-level evaluation." Accessed 14 July 2022. <https://arxiv.org/pdf/1808.07048.pdf>. <https://doi.org/10.18653/v1/D18-1512>.
- Lindholm, Mary, J. Lee Hargraves, Warren J. Ferguson & George Reed.** 2012. "Professional language interpretation and inpatient length of stay and readmission rates." *Journal of General Internal Medicine* 27 (10): 1294–1299. <https://doi.org/10.1007/s11606-012-2041-5>.
- Macken, Lieve, Daniel Prou & Arda Tezcan.** 2020. "Quantifying the effect of machine translation in a high-quality human translation production process." *Informatics* 7 (2): article 12 (1–19). <https://doi.org/10.3390/informatics7020012>.
- Marzouk, Shaimaa & Silvia Hansen-Schirra.** 2019. "Evaluation of the impact of controlled language on neural machine translation compared to other MT architectures." *Machine Translation* 33: 179–203. <https://doi.org/10.1007/s10590-019-09233-w>.
- Ministerio de Asuntos Económicos y Transformación Digital.** 2022. "Digital Spain 2025." Accessed 14 July 2022. [https://portal.mineco.gob.es/RecursosArticulo/mineco/prensa/ficheros/noticias/2018/Agenda\\_Digital\\_2025.pdf](https://portal.mineco.gob.es/RecursosArticulo/mineco/prensa/ficheros/noticias/2018/Agenda_Digital_2025.pdf).
- Moorkens, Joss, Sheila Castilho, Federico Gaspari & Stephen Doherty,** eds. 2018. *Translation quality assessment*. Cham: Springer. <https://doi.org/10.1007/978-3-319-91241-7>.
- Nitzke, Jean, Silvia Hansen-Schirra & Carmen Canfora.** 2019. "Risk management and post-editing competence." *JoSTrans. Journal of Specialised Translation* 31: 239–259.

- Ogden, Charles Kay.** 1932. *Basic English: A general introduction with rules and grammar*. London: K. Paul, Trench, Trubner.
- O'Brien, Sharon.** 2006. Machine-translatability and post-editing effort: An empirical study using Translog and choice network analysis. Doctoral thesis, Dublin City University.
- O'mara, Ben & Gemma Carey.** 2019. "Do multilingual androids dream of a better life in Australia? Effectiveness of information technology for government translation to support refugees and migrants in Australia." *Australian Journal of Public Administration* 78 (3): 449–471. <https://doi.org/10.1111/1467-8500.12388>.
- Pokorn, Nike K. & Jaka Čibej.** 2018. "It's so vital to learn Slovene': Mediation choices by asylum seekers in Slovenia." *Language Problems and Language Planning* 42 (3): 286–304. <https://doi.org/10.1075/lplp.00024.pok>.
- Prioleau, Jonathan.** 2022. Machine translation in healthcare. Unpublished Master's dissertation. Universitat Rovira i Virgili.
- Pym, Anthony.** 2018. "Introduction: Why mediation strategies are important." *Language Problems and Language Planning* 42 (3): 255–266. <https://doi.org/10.1075/lplp.00022.pym>.
- Pym, Anthony.** 2019. "Teaching translation in a multilingual practice class." In *Translator education: The evolving curriculum*, edited by David Sawyer, Frank Austerlühl & Vanessa Enríquez-Raído, 75–79. Amsterdam: John Benjamins. <https://doi.org/10.1075/ata.xix.15pym>.
- Pym, Anthony.** 2020a. "Quality." In *The Routledge handbook of translation and technology*, edited by Minako O'Hagan, 437–452. Abingdon: Routledge. <https://doi.org/10.4324/9781315311258-26>.
- Pym, Anthony.** 2020b. "When trust matters more than translation." *Pursuit* (29 July), University of Melbourne, <https://pursuit.unimelb.edu.au/articles/when-trust-matters-more-than-translation>.
- Pym, Anthony & Bei Hu.** 2022. "Trust and cooperation through social media. COVID-19 translations for Chinese communities in Melbourne." In *Translation and social media communication in the age of the pandemic*, edited by Tong King Lee & Dingkun Wang, 44–61. London: Routledge. <https://doi.org/10.4324/9781003183907-4>.

- Pym, Anthony** & **Ester Torres-Simón**. 2021. "Is automation changing the translation profession?" *International Journal of the Sociology of Language* 270: 39–57. <https://doi.org/10.1515/ijsl-2020-0015>.
- Raunak, Vikas, Arul Menezes** & **Marcin Junczys-Dowmunt**. 2021. "The curious case of hallucinations in neural machine translation." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, edited by Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty & Yichao Zhou, 1172–1183. Association for Computational Linguistics. Accessed 14 July 2022. <https://aclanthology.org/2021.naacl-main.92.pdf>.  
<https://doi.org/10.18653/v1/2021.naacl-main.92>.
- Shoemaker, Sarah J., Michael S. Wolf** & **Cindy Brach**. 2014. *The patient education materials assessment tool (PEMAT) and user's guide*. Updated edition. U.S. Department of Health and Human Services. Accessed 14 July 2022. <https://www.ahrq.gov/health-literacy/patient-education/pemat.html>.
- Suport Judicial**. 2021. *Serveis d'interpretació i traducció judicial*. Barcelona: Subdirecció General de Suport Judicial i Coordinació Tècnica. Accessed 14 July 2022. [https://justicia.gencat.cat/web/.content/home/ambits/administraciojusticia/dades-estadistiques/servei\\_de\\_traduuccions\\_i\\_interpretacions/memoria-traduuccions-2021.pdf](https://justicia.gencat.cat/web/.content/home/ambits/administraciojusticia/dades-estadistiques/servei_de_traduuccions_i_interpretacions/memoria-traduuccions-2021.pdf).
- Wallbrecht, Joshua, Linda Hodes-Villamar, Steven J. Weiss** & **Amy A. Ernst**. 2012. "No difference in emergency department length of stay for limited English proficiency patients with language disparities." *Annals of Emergency Medicine* 60 (4): S72–S72. <https://doi.org/10.1016/j.annemergmed.2012.06.176>.
- Wood, Emily**. 2018. "Twenty years of building for everyone." *Google Company Announcements*. Accessed 14 July 2022. <https://blog.google/intl/en-ca/company-news/inside-google/twenty-years-of-building-for-everyone/>.