REVISTA DE LLENGUA I DRET // JOURNAL OF LANGUAGE AND LAW #73

BUILDING MACHINE TRANSLATION SYSTEMS FOR MINOR LANGUAGES: CHALLENGES AND EFFECTS

Mikel L. Forcada*

Abstract

Building machine translation systems for disadvantaged languages, which I will call *minor* languages, poses a number of challenges whilst also opening the door to new opportunities. After defining a few basic concepts, such as *minor language* and *machine translation*, the paper provides a brief overview of the types of machine translation available today, their most common uses, the type of data they are based on, and the usage rights and licences of machine translation software and data. Then, it describes the challenges involved in building machine translation systems, as well as the effects these systems can have on the status of minor languages. Finally, this is illustrated by drawing on examples from minor languages in Europe.

Keywords: machine translation; minor languages; language resources; Aragonese; Breton; Saami; Norwegian *Bokmål*; Norwegian *Nynorsk*; Occitan; Catalan; Valencian.

CONSTRUIR SISTEMES DE TRADUCCIÓ AUTOMÀTICA PER A LLENGÜES MENORS: REPTES I EFECTES

Resum

La creació de sistemes de traducció automàtica per a llengües desfavorides, que anomenaré llengües menors*, presenta diversos reptes alhora que obri la porta a noves oportunitats. Després de definir conceptes preliminars com ara els de* llengua menor *i* traducció automàtica*, i d'explicar breument els tipus de traducció automàtica existents, els usos més comuns, el tipus de dades en què es basen, i els drets d'ús i les llicències del programari i de les dades de traducció automàtica, es discuteixen els reptes a què s'enfronta la construcció de sistemes de traducció automàtica i els possibles efectes sobre l'estatus de la llengua menor, usant com a exemples llengües menors d'Europa.*

Paraules clau: traducció automàtica; llengües menors; recursos lingüístics; aragonés; bretó; sami; noruec bokmål*; noruec* nynorsk*; occità; català; valencià.*

* Mikel L. Forcada, full professor of Software and Computing Systems at the Universitat d'Alacant, president of the European Association for Machine Translation (EAMT) since 2015, founder and chair of the Project Management Committee of Apertium, an open-source machine translation platform, and co-founder and research director of language technology company Prompsit Language Engineering. mlf@dlsi.ua.es 0000-0003-0843-6442

Article received: 14.01.2020. Blind reviews: 26.02.2020 and 22.04.2020. Final version accepted: 22.05.2020.

Recommended citation: Forcada, Mikel L. (2020). Building machine translation systems for minor languages: challenges and effects. *Revista de Llengua i Dret, Journal of Language and Law*, *73*, 1-20. https://doi.org/10.2436/rld.i73.2020.3404

Summary

1 Introduction

In today's digital society, a major part of our lives is spent online, which means that we depend on tools to efficiently process what we write. When these tools are not available for certain languages, their users are forced to switch to another language in order to conduct their online activity. This, in turn, can limit their online life experience. One important tool for processing linguistic content is *machine translation*. Knowing this, I aim to describe the challenges involved in building machine translation systems for disadvantaged languages, which I will call *minor* languages, as well as the effects of these systems on the languages involved.

The paper is structured as follows: in Section 2, I define basic concepts such as *minor language* and *machine translation*; I provide a brief overview of the types of machine translation available today, their most common uses and the type of data they are based on, and I discuss machine translation software and data usage rights and licences. In Section 3, I describe the challenges posed by building machine translation systems and, in Section 4, I address the potential effects these systems can have on the status of minor languages. Preceding my final remarks, Section 5 exemplifies the foregoing by looking at five minor languages in Europe, Breton, Occitan, Aragonese, Norwegian *Nynorsk* and Northern Saami, and one not-so-minor language, Catalan.

2 Basic definitions

2.1 Minor languages

For the purposes of this paper, I employ the term *minor language* to encompass an array of disadvantageous situations faced by languages, echoing a decision originally made in a paper published fourteen years ago.[1] Many expressions, which prove synonymous with our use of *minor language* to varying degrees, appear in the literature and on the Internet. Besides *minor languages*, the following related phrases obtain a relevant number of hits on Google (provided in parentheses): *minority languages*, or languages spoken by the minority of people in a nation or region (1,160,000); *lesser used languages*, or languages with comparatively smaller numbers of speakers (59,800); *small languages* (70,500); *smaller languages* (41,800), and *under-resourced languages* (47,300). In this paper, I choose to disregard the unique connotations of each term (for example, the fact that a minority language in one country could, in fact, be a rather large language on the world scale, such as Gujarati in the UK); rather, I will employ the term *minor languages* (89,000) to refer to all languages that have some, if not all, of the following traits (see Streiter et al., 2006).

A small number of speakers (or, as the paper deals with machine translation and, therefore, written texts, a small number of *literate* speakers).

- A use that is somehow far from normality (for instance, common at home or in family situations but not at school or in trade or government affairs; discriminated against by society, neglected politically, poorly funded, forbidden or repressed).
- Lack of a single writing system, consistent spellings or a widely accepted standard variant.
- Extremely limited presence on the Internet.[2]
- Few linguistic experts.
- An absence of computer-readable resources, such as dictionaries and corpora.
- A dependency on technologies that are not readily accessible by users of the language in question.

In contrast, the paper employs the terms *main languages*, *major languages* or *dominant languages* to refer to languages that do not have any of the above-described limitations.

To illustrate this, let us consider the languages of Europe, or more specifically, the languages of the European Union (EU). The EU is—and strives to be, at least in its official declarations—multilingual. Europeans'

[1] Forcada (2006); parts of this paper have, in fact, been taken and updated from the 2006 publication.

[2] Or, as Williams et al. (2001) put it, "non-visible in information system mediated natural interactivity of the information age".

language identity is at the core of how they perceive themselves and understand their relationships with others. Linguistic diversity is a highly valued asset in Europe, but also a challenge and a source of fragmentation. Functionally speaking, the majority of Europeans are (still) monolingual,[3] which may lead to injustices, as it prevents full social participation in European society from extending beyond their language barriers. In fact, multilingual individuals are considered to have a definite advantage in Europe.[4]

Not all European citizens have the same rights, however. Even within the same EU Member State, users of any of the 24 official languages (for example, Dutch in the Netherlands) have more rights than users of unofficial languages (such as the Basque language in France). When it comes to engaging with EU institutions, users of languages not listed in the *Treaty of Lisbon* face severe hindrances. Since communication is key to building social cohesion among citizens, it comes as no surprise that citizens experience varying levels of such cohesion depending on what their first language (L_1) is.

At the present time, a major part of our lives is spent online and, as time goes on, we engage more often with institutions and businesses via the Internet. Consequently, for a language to be useful, it has to be useful online. This means that texts in this language need to be available and efficiently processed on the Internet, making language technologies a cornerstone of cohesion in today's digital society.

This paper deals with one specific type of language technology: machine translation. Since machine translation acts as a bridge between languages, the effects that its availability has on a minor language (*M*) will depend on which languages the bridge leads to:

1. A major language (*P*) may have a translation-based link with a minor language (*M*) thanks to the availability of a machine translation system between *P* and *M*. For example, *M* = Breton and *P* = French.
2. A major language (*P*) may have a translation-based link with another language (*N*) that is very closely related to a minor language (*M*). If *M* and *N* are indeed so similar, it will be relatively easier to create a machine translation system between them. Thus, if a machine translation system already exists between *N* and *P*, the minor language, *M*, can also benefit from it, albeit indirectly by way of *N*. For example, we could indirectly translate from *P* = English into *M* = Asturian by way of *N* = Spanish.

2.2 Machine translation

2.2.1 Definition

Machine translation handles written, and specifically, *computerised* texts. These are text files stored in a computer-based medium, such as files that you generate or edit using a word processor. It is called *machine* translation because it is carried out by a *computer system*—computers with the proper software installed—without the need for human intervention. Accordingly, the term *machine translation* signifies any use of computer systems to transform a computerised text written in a *source language* into a different computerised text written in a *target language*, thereby generating what is known as a *raw translation*.

Machine translation is not without its limitations. Generally speaking, raw translations generated by machine translation systems differ from those produced by translation professionals and may or may not be suitable for certain communicative purposes. If they prove unsuitable, this could be due to a number of challenges, including the ambiguity of human texts (which are often riddled with polysemous words[5] and sentences with more than one possible syntactic structure[6]) and syntactic divergences between the source and target

[3] In the Eurobarometer Special Survey (2012), 46% of Europeans reported only being able to hold a conversation if it is in their mother tongue.

[4] See also Rivera et al. (2017), a study commissioned by the European Parliament's Panel for the Future of Science and Technology.

[5] Such as the Spanish verb *registrar*, which can mean either "enter or record on an official list" (in English, *register*) or "search someone or something to find something hidden" (in English, *search*, *examine, frisk*); or the Spanish homograph *libertad*, which can be a noun (in English, *freedom, liberty*) or a second person plural verb in the imperative mood (in English, *free, let free*).

[6] Such as the English sentence *George received the oranges he had bought from Amy*, in which the prepositional phrase *from Amy* can modify either the verb phrase of the subordinate clause, *he had bought*, or the verb phrase of the main clause containing it, *George*

languages.[7] These challenges are addressed using methods which, in general, simplify the translation process quite radically. The resulting simplifications, while allowing for the creation of rather simple translation mechanisms on which to build swift and compact machine translation systems in a reasonable amount of time, do prevent the solutions from being anywhere near optimal.

In light of these limitations, we can leave it up to a good machine translation system to take care of the more mechanical (or "mechanisable") aspects of translating. However, regardless of how good it is, we cannot expect it to understand the text, consistently translate ambiguous phrases correctly or produce genuine variants of the target language.

2.2.2 Uses

Machine translation applications can be categorised into two main groups.

The first group comprises applications for *assimilation*, which is the use of machine translation to understand the general meaning of texts (for example, those posted on the Internet) written in another language. Examples:

- A patent office might use machine translation to determine whether patents in other languages infringe upon their own or, conversely, whether their patents are at risk of infringing upon those written in other languages (Nurminen, 2019).
- An online chat application could integrate machine translation, thereby allowing everyone involved to write in their own language and read the other members' comments, despite the different languages used.
- Social media networks like Twitter and Facebook can offer a "see translation"-type feature that uses machine translation to allow users to read posts written in other languages.

For this type of machine translation application to work, it needs to be extremely swift—ideally instantaneous. The raw translations generated are used as is, although they may not be read entirely (for instance, if the translated document is too long) and they are not usually kept or saved after being read. For these reasons, machine translation applications like this are not intended for translation professionals, but rather for the general public.

The second group of applications are for *dissemination*, which refers to the use of machine translation by translation professionals. In this case, the word "dissemination" is used because such applications are employed as an intermediate step towards producing a document, in a given target language, for publication or dissemination. Thus, raw translations are typically saved for subsequent revision and proofing by a (ideally specialised) translation professional, a process which has come to be known as *post-editing*.[8] Entire documents can be machine translated and then post-edited in a word processor. However, it is more comfortable to integrate machine translation into a computer-assisted translation environment, where it can serve as a support tool alongside translation memories (which allow users to retrieve previous translations for similar sentences), termbases (which contain the validated translations of specialised terminology) and tools for calculating costs and productivity. In any case, and simplifying a little, machine translation followed by post-editing constitutes an alternative to professional translation only if its joint cost is lower than that of

received the oranges he had bought. Whilst in the former case George has bought oranges from Amy, in the latter he has received them from her.

[7] The Basque sentence *Donostiatik etorri den gizonari eman diot* (in English, *I have given it to the man that has come from Donostia*), is literally "Donostia-from came is-that man-the-to given to-him-it-I-have". Given the syntactic divergence between these two sentences, translation from one to the other entails a radical (if not complete) reordering of the sentence fragments.

[8] For-profit and not-for-profit crowdsourcing models are increasingly being used to generate publishable content with machine translation. In such models, participants may not in fact be translation professionals and, in some cases, they do not post-edit target language texts as raw translations but rather proofread them without concern for the source language.

traditional professional translation, or when one wants to speed up the translation process while maintaining the cost.[9]

2.2.3 Types

One can also distinguish between two main types of translation technologies.

From roughly 50 years ago, when the first attempts at machine translation were made, to the 1990s, the prevailing approach was called *rule-based machine translation* (RBMT), which we can find in systems such as Lucy,[10] ProMT[11] and Apertium.[12] Typically, RBMT begins with word-to-word translation, ideally building up to process entire sentences. To develop an RBMT system:

- Firstly, translation experts must compile electronic dictionaries, write rules that analyse the source language and transform source structures into equivalent structures in the target language, and carry out other similar actions. Note that translators' intuitive, non-formalised knowledge has to be converted into rules and coded in a computationally efficient way. This can lead to drastic simplifications which, if chosen wisely, can be useful in most cases.
- Secondly, computer experts write programes (called *translation engines*) which analyse and subsequently translate original texts by consulting the relevant dictionaries and applying the written rules in their set order.

More recently—since the beginning of the 1990s—we have witnessed a rise in *corpus-based machine translation* (CBMT). In this case, machine translation programes "learn to translate" by drawing on enormous corpora of bilingual texts in which hundreds of thousands or even millions of sentences in one language have been paired, or *aligned*, with their corresponding translation in another. Such alignment gives rise to vast *translation memories*. For CBMT, the role of translation experts could appear less important if we fail to consider the amount of effort put into translating (ideally, but not always, professionally) the texts making up the training corpora (see 2.3.2 for other possible sources of bilingual training corpora).

Corpus-based machine translation relies on two main strategies: *statistical* machine translation and *neural* machine translation.

- Statistical machine translation was first devised towards the end of the 1980s and has been commercially marketed since 2003. Programes of this nature learn and use probabilistic models that count the number of times certain occurrences appear in the bilingual training corpora. This could mean counting the number of times a certain word collocates with another in target language sentences or the number of times a given word is used in a target language when another specific word appears in the source language.
- First showing up on the market in 2016, neural machine translation is relatively new to the scene. It draws on artificial neural networks inspired (vaguely) by the way the brain learns and generalises information. In this case, learning and generalisation are based on observation of bilingual corpora (Forcada, 2017; Casacuberta & Peris, 2017). In fact, the main online systems—those made publicly available by Google[13] and Microsoft,[14] for instance—are based on neural machine translation, and newer systems, such as DeepL,[15] also rely on this type of technology.

[9] Sometimes, to save on post-editing costs, one can lightly pre-edit the original text to make it more machine translatable, i.e. void of some of the problems the specific machine translation system being used is known to have.

[10] http://lucysoftware.com/catala/traduccio-automatica/kwik-translator/

[11] https://www.online-translator.com/

[12] http://www.apertium.org

[13] http://translate.google.com

[14] https://www.bing.com/translator

[15] https://www.deepl.com/translator

There are, of course, *hybrid* systems which blend the two strategies. To provide an example, this could mean using morphological rules to analyse texts before translating them with a system that has been trained on a corpus of morphologically analysed texts.

Rule-based machine translation systems require a hefty amount of linguistic and translation work and take more time to build, since the linguistic data (rules and dictionaries) must be explicitly coded so that the systems can use them. By contrast, corpus-based systems generally take less time to build, although this only holds true if available corpora provide a large enough volume of sentence-aligned translated text. As such, the latter form of machine translation can be difficult to apply to minor languages with few digitised corpora. In these cases, RBMT may be the only strategy with any chance of success, which is why greater focus will inevitably be placed on it in this paper.[16]

It is important to bear in mind that statistical machine translation systems can produce deceivingly natural-sounding texts that are not, in fact, suitable translations of the original texts, given the importance they place on imitating the target language texts used to train them. The risk of this is even greater when it comes to neural systems.

Regardless of the type of technology used, machine translation systems comprise three key components:

- An *engine*, the programe that carries out the machine translation.
- *Data* in the form of language resources and corpora (more on this in the next sub-section).
- *Tools* for managing and maintaining the data and converting them into the format required by the engine. In corpus-based systems, this conversion includes the learning process that yields the corresponding probabilistic or neural models.

2.3 Data

Machine translation, irrespective of the type, requires data on the source-target language pair in computer-readable format. The nature of this data, however, will vary depending on the type of machine translation, as described below. For the purposes of this paper, we will distinguish between two types of data: *language resources* and *corpora*.

2.3.1 Language resources

For rule-based machine translation to work, we need to provide the engine with *language resources* (referring here to resources in computer-readable, rather than human-readable, formats), such as monolingual dictionaries that describe the morphology of a source or target language, rules for handling homographic and polysemous words, rules for transforming source language structures into equivalent structures in the target language, and bilingual dictionaries. These resources must be stored in the format that the *tools* and translation *engine* expect. As previously noted, resources like this are difficult to build and require the support of linguistic and translation experts who are familiarised with the formats used by the system in question. Indeed, the experts have to create resources from scratch or somehow transform the ones already available.

Language resources can also be used to automatically transform, *annotate* or prepare the linguistic corpora described in the following sub-section, in order to render them more useful in training corpus-based systems; for example, indicating to which lexical category (noun, adjective, etc.) each word in the texts belongs.

2.3.2 Corpora

For corpus-based machine translation to work, we need a huge number (hundreds of thousands or millions) of sentence pairs, each made up of a source language sentence and its translation.[17] Putting together this type

[16] It should be noted that CBMT for under-resourced languages is an active field of study. For example, the author is the scientific coordinator of the European project known as GoURMET (Global Under-Resourced Media Translation, http://gourmet-project.eu), which deals with neural machine translation between English and languages such as Amharic, Kyrgyz and Swahili.

[17] For statistical machine translation, it is advisable to have even larger numbers of sentences in the target language (a monolingual

of corpus also requires considerable effort. Namely, there must be enough professionally translated texts available when it comes time to train the system, and the translations must be aligned sentence by sentence (alignment, of course, can be done by a machine with a certain margin of error). It is not uncommon for corpora to contain *noise*; in other words, translations that cannot be considered suitable and must therefore be singled out and deleted.

Recent advancements enable systems to crawl multilingual websites in order to populate their corpora. This is one of the methods used by commercial systems such as Google, Microsoft and DeepL. To do this, documents in the languages of interest are downloaded, and original texts and their translations are matched up by examining their length and structure and consulting available bilingual resources. Next, the texts are split into sentences, which are paired up as best as possible. Lastly, simple techniques are used to discard paired sentences that are not, in fact, a real match (for example, because one sentence in the pair is much longer than the other).[18]

There are projects, such as OPUS,[19] which attempt to gather all the open-source parallel corpora for dozens of languages from around the world. Nonetheless, religious and computer-related texts tend to be the only ones available for minor languages.[20]

2.4 Usage rights and licences

2.4.1 Software

Two types of software are used in rule-based machine translation. On the one hand, as explained above, there is an *engine* that carries out the machine translation, which should be as independent from the languages as possible. On the other, it needs *tools* that are able to manage the language resources used by the system, in order to edit them (create them from scratch or update them) and convert them into the format used by the engine. In terms of access, machine translation software may be designed for installation on a local computer, such as a desktop, laptop, smartphone or a server housed on an institution's or business's premises. Otherwise, it can be stored on a remote computer (server), thereby allowing people to access it via the Internet, which is how Google Translate and DeepL operate, for instance. In the latter case, usage rights are determined by the conditions of use established for the remote service. Although it is true that Internet systems are increasingly supporting minor languages, machine translation still remains unavailable for many of them. Section 5 looks at cases of languages with a single machine translation system, such as Aragonese, Breton and Northern Saami.

When software needs to be used on a local computer, users need to pay particular attention to the *licence*. All software can be classified as either *free* or *non-free.* Free software[21] can be:

- Freely run for any purpose;
- Freely examined to see how it works and modified to make it suitable for new requirements or applications. For this to be possible, the *source code* (the programe's *editable* code written in a programming language) and the *executable* derived therefrom must both be available. Hence the alternative denomination *open-source software;*[22]
- Freely redistributed to anyone; and

text corpus) in order to ensure natural-sounding translations.

[18] For example, Paracrawl (http://www.paracrawl.eu) develops software for creating bilingual corpora from texts available on the Internet and, what's more, releases its corpora for use by the public.

[19] http://opus.nlpl.eu

[20] On the one hand, the Quran or texts by the Jehovah's Witnesses, and on the other, documents related to free/open-source software, such as LibreOffice.

[21] For further information, visit http://www.gnu.org/philosophy/free-sw.html.

[22] For the purposes of this paper, the definition put forth by the Open Source Initiative (http://www.opensource.org/docs/definition.php) is considered a near equivalent.

- Freely improved and released so that the entire community of users can benefit from it (the source code must be available for this as well).

Software that does not fulfil the above-listed conditions is non-free. This may be the case even though the software comes free of charge; Opera (a browser), Adobe Acrobat (software for viewing documents) and WhatsApp (a text messaging system) all provide good examples of this. In the case of non-free software, use could be limited to one person, commercial exploitation could be prohibited or the source code may not be available.

The engine and tools used in machine translation for minor languages should ideally be free. The machine translation platform called Apertium is a good example to follow: thanks to the free licence, improvements to the engine can be made public, which benefits all users, regardless of the languages they are working with.[23]

2.4.2 Data

As we saw above, machine translation software is special because it relies heavily on data. Rule-based machine translation depends on language resources such as morphological dictionaries, bilingual dictionaries, computational grammars and lists of rules for transforming the structures of one language into those of another. Meanwhile, corpus-based machine translation (whether statistical or neural) relies either directly or indirectly on the availability of parallel texts aligned sentence by sentence.

As is often the case when it comes to many minor languages, and despite the efforts of initiatives like OPUS (cited in 2.3.2 above), obtaining and preparing enough parallel texts with paired sentences (with word counts in the hundreds of thousands or millions) to yield reasonable results through statistical or neural CBMT may still be impractical. In such cases, it may be easier for expert users of the minor language to acquire the skills needed to encode their knowledge into dictionaries and rules needed by the RBMT system. The rights of access and conditions of use of these systems will depend on the licence under which the newly created language resources are made available. These resources (rules, dictionaries) can be free, like Apertium's, in the same way that software can be free, as laid out in the preceding sub-section. This will enhance the effect of the resulting matching translation system on the status of the minor language in question, since the linguistic community will not only be able to access and use it, but also improve upon and disseminate the language resources it is based on.

3 Challenges

3.1 Technophobic and Luddite-esque attitudes

Even if a minor language is backed by a group of motivated, well-trained linguistic activists, they must be able to combine their language-related expertise with information technology skills. This is not always a simple endeavour: what could be deemed *technophobic* or *Luddite*[24] attitudes have been detected in many linguistic communities. This refers to individuals who, despite being literate in a minor language and generally well-trained, distrust technology either because they are clinging to an idealised view of language and human communication or because they undervalue non-formal or non-literary expressions of the language.[25] Any group of people making an effort to build open-source machine translation systems for a minor language must be ready to face this sort of, let's call it *socio-academic*, adversity.

[23] Rather than being completely free, another solution would be to make the engine and the tools publicly available, with the data in well-documented formats. If this were the case, systems could be built by creating relevant linguistic data, although the person actually running the software would need to have the right to use the engine and the tools.

[24] Based on the pseudonym *Ned Ludd*, the term Luddite was coined to refer to the activists who were determined to destroy the machines being built during the United Kingdom's Industrial Revolution out of fear of losing their jobs. In mid-nineteenth-century Catalonia, these same tensions took shape in the so-called *conflicte de les selfactines* (conflict of the self-acting machines).

[25] Perhaps because many of these language professionals are in the habit of focusing on generally infrequent phenomena that are unique to a given language (the *gems*), which are not typically handled well by machine translation systems, instead of recognising the ability of these systems to process the commonly used words and structures making up 95% of everyday texts (the language's *basic building blocks*).

3.2 Standardisation

The lack of a commonly accepted writing system, standard spellings or a recognised prestige dialect can signify a serious hurdle for anyone attempting to build a machine translation system for a minor language (what we could call the *pioneer syndrome*). If a system is programmed to follow a standard that is not generally accepted, its overall usefulness is compromised. Section 5 discusses Occitan and Aragonese, two languages whose standards are still unstable.

It is worth highlighting here that, for corpus-based machine translation systems, a more informal approach to this problem is possible, although this can clearly lead to consequences: if a system is trained using corpora that are inconsistent in the standard they follow,[26] the resulting machine translations will also be inconsistent.[27]

3.3 Knowledge elicitation and linguistic complexity

Creating language resources presents two closely related challenges.

Firstly, minor languages may face an undersupply of the explicitly encoded linguistic knowledge needed to create language resources. In order to generate useful linguistic data, the intuitive knowledge of language users must be made explicit; in other words, it needs to be *elicited* from them. Of course, certain types of language resource are simpler to build than others. For example, using a well-designed form interface, the knowledge of volunteer language users can be harnessed to create and maintain dictionaries. They could be asked to add monolingual and bilingual dictionary entries via this form interface, which would allow them to select inflectional paradigms, choose translation equivalents in any direction, etc. However, certain necessary linguistic data, such as the rules for transforming the syntactic structures of one language into those of another, are not so easily elicited from non-experts.

Beyond this, given the ongoing need to edit and update the language resources, it is important to keep the required linguistic knowledge to a minimum. The aim is to encode the knowledge of a community of language users via levels of representation that can be easily learned based on the basic grammar and language skills and concepts that can be acquired at primary or secondary school. However, that does not prevent volunteers from learning the formats into which this knowledge must be encoded. Indeed, novel developers should be able to benefit from a combination of written documentation and support from other developers.

3.4 Organising the creation of resources

Open-source machine translation technology has the potential to benefit minor languages in a number of ways, namely thanks to the creation of machine translation systems that bridge them to other languages. In this regard, the work of communities of people that develop language resources or collect and process texts to populate corpora is especially important. Frequently, these people have to volunteer their time due to a lack of funding, a common issue faced by minor languages.[28] Many minor languages on the periphery of normality or officiality rely on groups of activists, generally from the field of education, with the linguistic and translation skills necessary to collaboratively create linguistic data (dictionaries and rules) and build and manage corpora. Despite being absolutely vital for minor languages, linguistic and translation skills and volunteered time are simply not enough; the work of experts needs to be coordinated by a smaller group of people with full proficiency in the translation engine and tools being used.[29]

[26] Some inconsistencies can certainly be removed automatically before training takes place, but this comes with its own set of consequences. For example, the BBC publishes texts in Igbo (a tonal language spoken in Nigeria by roughly 45 million people), some with diacritics marking tone and others without. Since unequivocally restoring these diacritics is impossible (precisely because of their diacritical value), they can be removed from the entire corpus, at the cost of adding ambiguities.

[27] This can actually be seen at times with translators such as Google Translate.

[28] The free machine translation platform known as Apertium initially received public funding from the governments of Spain and Catalonia, which enabled the hiring of expert personnel (for the Spanish-Catalan, Spanish-Galician, Catalan-Aranese, Catalan-English, etc. language pairs). Since then, some language pairs have been developed almost entirely thanks to the work of volunteers, whilst others have received partial funding from companies, NGOs and programmes such as Google Code-in and Google Summer of Code.

[29] Apertium has a Project Management Committee on which seven developers sit. Committee members are elected every two years by the active developers.

3.5 Resource management: licences and commons

A good way of improving the lot of under-resourced minor languages is by creating a shared body of language resources and software that is easily accessible, can be freely used by anyone and for any purpose, can be easily modified and improved for new applications, and encourages developers to contribute to it by making modifications and improvements. Free licences (see 2.4.1) ensure that these practices can be carried out freely, which is why the choice of licence is such a crucial factor in reaching these objectives.

Commercial machine translation, in contrast, provides far fewer opportunities. Major machine translation companies have their sights set on the world's well-resourced languages, as they are used in more developed markets with higher business potential, leaving minor languages with limited advantages. Moreover, commercial machine translation engines and the language resources they use are normally non-free, making it difficult to modify them to make them suitable for under-resourced minor languages. Usage rights are highly restricted (for example, combining and redistributing them tends to be frowned upon or prohibited) and the licences are expensive, which makes adopting them difficult. This means that hopeful users need to request permission from the seller to make the system work for their intended application and, consequently the linguistic community is forced to rely on a specific seller.

After having chosen a free licence under which to share data and software, there is a chance that, since they can be freely used, modified and distributed by anyone and for any purpose, some developers will move in and take advantage of the free data and software to produce and distribute non-free products without giving anything in return. A good licence would ideally dissuade such private appropriation and benefit without any sort of trade-off. Instead, it should favour collaborative development and the aggregation of complementary skills, without keeping people from building businesses around the languages involved (necessarily based on service provision rather than the sale of licences).

One possibility is to choose licences that facilitate the creation of a *commons*. Before the age of computers, a *commons* was an undivided piece of land for community use, intended, for example, for grazing, or an area open to the public in a municipality. In the digital age, we now have *software commons*, or code that is subject to common use, as well as *language resource commons*. Both the software and the language resources available for minor languages should be managed as commons.

When *copyleft* (a play on *copyright* as well as a type thereof) is added to a free licence, it means that modifications must be distributed under the same licence. Thus, there are free licences *without copyleft*, such as the three-clause BSD licence,[30] the MIT licence,[31] the Apache licence[32] and the Creative Commons Attribution licence, CC-BY,[33] and those *with copyleft*, such as the GNU general public licence[34] and the Creative Commons Attribution-ShareAlike licence, CC-BY-SA[35]).

Copyleft can thus provide sturdy scaffolding for the creation and maintenance of software and data commons, as it discourages private appropriation by obliging derivatives to be distributed under equal terms and sets up an even playing field for all. As a result, it fosters collaborative development and allows communities of developers to build shared sets of free resources, as all the resulting work must be shared under the same free licence. Likewise, copyleft encourages the creation of service-based companies (adaptation, installation, redistribution), releasing minor languages from their commercial blockage in the process.

This section could not come to an end without mentioning a problem that affects corpus-based machine translation. As explained in 2.2.3 above, these systems are trained using available translations, many of which are collected from documents posted on the Internet. On the one hand, in accordance with the Berne Convention, when a work does not explicitly express the terms under which it can be reused, it is understood

[30] https://opensource.org/licenses/BSD-3-Clause

[31] https://opensource.org/licenses/MIT

[32] https://www.apache.org/licenses/LICENSE-2.0.html

[33] https://creativecommons.org/licenses/by/4.0/deed.en

[34] https://www.gnu.org/licenses/gpl-3.0.ca.html

[35] https://creativecommons.org/licenses/by-sa/4.0/deed.en

that the author reserves all reproduction rights.[36] However, whether the translations generated by a system that has been trained using multilingual works should be considered the public reproduction of substantial parts thereof is still up in the air. Moreover, we would be wise to ask ourselves to what extent the copyright of original texts and their translations are respected when they are reused to create commercial machine translation systems. If individuals paying for a translation clearly state that it will be published openly, are they not satisfactorily completing the transaction with the translator? For a discussion on these matters, see, for example, Moorkens & Lewis (2019).

4 Effects

The existence of machine translation systems for minor languages can affect these languages positively. The easier it is to access these systems and the more readily available the data used to create them, the more intense the effects will be.

4.1 Normalisation and visibility

The availability of machine translation between a minor language and a surrounding dominant language can help to *normalise* the minor language, extending its use beyond the family and the home to more formal social contexts, such as schools, the media, government affairs and trade. Here are a few examples to illustrate this point:

- Educational material in a dominant language can be translated into a minor language so that children can be educated in the latter.
- News items published in a major language can be translated into a minor language in order to create media content for the relevant linguistic community.
- Laws, regulations, government announcements, advertisements, calls, etc. can be translated into a minor language.
- Companies can more easily market new products in a minor language via localisation. This is especially useful for products with a significant textual component, such as consumer electronics and mobile telephones.

Of course, in each of these scenarios, it is assumed that it is feasible to post-edit raw machine translations into suitable texts prior to publication. Therefore, the better the machine translation system is, the greater the aforementioned positive effects will be; for instance, when there are few linguistic divergences between the languages in question.

Additionally, the availability of machine translation from a minor language into one or more surrounding dominant languages can aid in the dissemination of materials originally written in the minor language. For example, web content could be written and managed directly in a minor language and machine-translated for users of other main languages. This could be done on the fly (in assimilation applications like those described in 2.2.2 above) or after a review process carried out by professionals (see, for example, the case of Northern Saami, described in 5.5 later on).

4.2 Literacy

The growing availability of texts in a minor language, obtained thanks to machine translating, post-editing and subsequently preparing materials that were originally written in a dominant language, can fuel efforts to improve literacy amongst speakers within the relevant linguistic community.

[36] The Berne Convention for the Protection of Literary and Artistic Works (https://www.wipo.int/treaties/en/ip/berne/index.html) states that "protection must not be conditional upon compliance with any formality (principle of 'automatic' protection)".

4.3 Standardisation

The use of machine translation systems can help to standardise a language by establishing a single writing system, encouraging set spellings, promoting a particular dialect and so on (see the cases of Aragonese and Occitan in Section 5 below).

4.4 Increased expertise and more available resources

Creating a machine translation system for a minor language implies, to a certain extent, reflecting on the language and subsequently specifying and codifying monolingual and bilingual knowledge. As long as it is in an open-source environment, the resulting linguistic expertise will be at the disposal of the entire linguistic community through the publication of new language resources.

5 Case studies

This section contains a brief overview of six cases in which a useful and publicly available machine translation system has been successfully created to bridge the gap between a minor language and a main language. Likewise, it addresses the challenges faced during development, the effects these systems have had on the minor languages involved and the resources that have been generated as a result. In light of the scarcity of bilingual corpora and the limited effect that commercial systems have when these do exist, emphasis will be placed on free rule-based machine translation systems housed on the Apertium platform.

5.1 Breton

Breton, called *brezhoneg* by users of the language, is a Brittonic language of the Celtic language family spoken in western Brittany (known as *Breizh Izel* or "Lower Brittany"), in France. The main language it has contact with is French, the country's sole official language. In fact, despite being spoken by roughly 200,000 people, Breton has virtually no legal recognition in France. To provide some indicators of its situation as a minor language, it should be noted that only 2% of education is carried out in Breton, only some road signs are bilingual and the language has a rather reduced presence in the media. The language's main promoting organisation is the *Ofis Publik ar Brezhoneg*,[37] and it has a well-established and generally accepted standard variant.

Browsers like Firefox, the Google server and some Microsoft tools such as Office and Skype have been *localised*[38] for Breton users, and Wikipedia in Breton has accumulated approximately 70,000 entries to date. There is not much software dedicated to the Breton language, although most of it, such as Apertium's machine translator (which we will look at shortly) and the LanguageTool grammar and spelling checker, are free (both in terms of cost and usage rights). This software, as well as other services (such as the online dictionary by Freelang[39]) are based on language resources such as morphological analysers and monolingual and bilingual dictionaries. In terms of bilingual text corpora, OPUS currently contains approximately 400,000 sentence pairs, the majority of which are highly specialised and from the field of computer science.

The Apertium project has a free Breton-to-French machine translator for assimilation purposes; in other words, to enable francophone readers to access content in Breton.[40] This machine translation system is the outcome of an initiative spearheaded by Francis Tyers, one of Apertium's lead developers. The system was initially presented in May 2009 and is the only one of its kind worldwide. As described in Tyers (2010), it was the result of the joint efforts of Gwenvael Jéquel and Fulup Jakez from the Ofis ar Brezhoneg (predecessor of

[37] The Public Office for the Breton Language (http://www.brezhoneg.bzh/) is a public cultural cooperation establishment set up by the French government, the Regional Council of Brittany, the Regional Council of Pays de la Loire and the departmental councils of Finistère, Morbihan, Côtes-d'Armor, Ille-et-Vilaine and Loire-Atlantique. Its mission is to promote the Breton language and foster its presence in all areas of language use.

[38] *Localisation* refers to the process of adapting a product to a certain *local* market, and language is often included amongst the characteristics that define local markets.

[39] https://www.freelang.com/enligne/breton.php

[40] The developers made the deliberate decision not to work on translation from French into Breton, as they considered it too risky given the sociolinguistic situation. Namely, they feared that people might mistakenly accept and use machine translations that are not actually correct (Jakez, 2009).

the Ofis Publik ar Brezhoneg), Valencia-based company Prompsit Language Engineering and the Universitat d'Alacant. As stated above, it is part of the Apertium project and is housed on its platform. The dictionaries were not put together from scratch, as Breton dictionaries were freely available on Lexilogos.[41] A primitive version of the system was used (Tyers, 2009; Sánchez-Cartagena et al., 2011) to augment the scant data available at the time, which Tyers (2009) himself had gathered using materials from the Ofis ar Brezhoneg and released under a free licence (some 31,000 translated sentences)[42] to train statistical machine translation systems.

Development on this machine translation system has recently slowed, although it is still the only one publicly available and under a free licence. Regarding the development issues, Fulup Jakez explained that "one of the main difficulties lies in the time it requires to improve the dictionaries and establish transfer rules. Besides that, we have always counted on the help of Fran[cis Tyers] and have never gained enough autonomy [from Apertium] because of our lack of computer skills. So, when he is unable to dedicate […] time like he did at the beginning […], the project's activity dwindles". Nevertheless, improvements continue; one way is by detecting those Breton words that the system does not recognise that are more common in the texts translated using the version installed on the Ofis's website,[43] and periodically adding them to the system. At the present time, the quality of the French in these translations makes them unsuitable for post-editing, but it is high enough for a French reader to obtain a rough idea of what the text in Breton is about.[44,45]

Regarding the reuse of resources generated whilst building the machine translator, a large portion of the data has also been used to set up the Breton grammar checker known as LanguageTool[46].

5.2 Occitan

The Occitan language is known as *lenga d'òc* by its native speakers, as well as sometimes being referred to using the name of one of its dialects, such as Limousin, Provençal or Gascon. This Romance language, which enjoyed great prestige during the Middle Ages, has since become a minor language. In the absence of an official census, it is estimated to have between 100,000 and 800,000 speakers, mainly in southern France, but also in the Val d'Aran, a region of the Spanish-governed Catalan territory, and areas of western Italy. The standardisation of Occitan faces a number of issues and is not free of controversy. It can be argued that this is largely due to the divergence amongst its dialects,[47] of which the Languedocien dialect has been given certain superiority in Standard Occitan, known as *occitan referenciau* or *occitan larg*. Standard Occitan is codified using what is known as the *nòrma clàssica*, but there also exists an alternative spelling system, called the *nòrma mistralenca*.[48]

Occitan checks many of the boxes for being considered a minor language. Like Breton, it has received virtually no legal recognition in France. It is, however, recognised as a *protected language* in Italy. Except in the case of the Val d'Aran, in which Catalonia's Statute of Autonomy guarantees schooling in this language as one of the official languages of the country, education carried out in Occitan in the so-called *calandretes*[49] is rare (limited to around sixty schools) and essentially *tolerated* by France. Another example: only in a few places in Occitan-speaking France are there any bilingual road signs, which contrasts with the Val

[41] https://www.lexilogos.com/

[42] http://opus.nlpl.eu/OfisPublik-v1.php

[43] The machine translator (http://www.fr.brezhoneg.bzh/42-traducteur-automatique.htm) is the most visited area on the Ofis's website (Jakez, 2009), accumulating over 60,000 clicks per month.

[44] In 2011, the system was able to translate 90% of the words found in texts on Wikipedia in Breton using a Breton dictionary with 17,000 entries and a Breton-French bilingual dictionary with 26,000 entries, 250 disambiguation rules and 250 grammar transformation rules.

[45] Whilst this paper was under review, another paper describing how to use Apertium's Breton-French machine translation system resources to improve the output of neural machine translators (see 2.2.3) in cases where corpora are scarce was accepted for presentation at the 2020 conference of the European Association for Machine Translation, EAMT (Sánchez-Martínez et al., 2020).

[46] http://languagetool.org

[47] At least as perceived by many users of the language.

[48] So called for having been used by Frederic Mistral, a writer and Nobel Prize winner.

[49] *Calandretes* are private (associative) schools that offer language immersion programmes. The public-school system in France offers bilingual education in very few early childhood and primary schools and only a certain level of continuity, with Occitan being offered as an optional subject during secondary education.

d'Aran, where they exist more systematically. Moreover, there is no leading organisation responsible for this language.

In terms of general-purpose software, a good portion of the free software (the Firebox browser, Google products, the LibreOffice system, etc.) has been localised, and Wikipedia in Occitan contains approximately 90,000 entries. As far as software for the Occitan language is concerned, there are two machine translation systems: one is based on the Apertium platform, is free and handles Occitan-Catalan/Spanish/French translation; the other is non-free and marketed by a company known as Sail Labs.[50]

Thanks to the development of these machine translation systems, there has been an upsurge in the number of monolingual and bilingual language resources available (dictionaries, rules); the free ones can be accessed via the Apertium project website. Likewise, there are resources which can be consulted on the Internet, such as Freelang and the dictionaries made available by Lexilogos.[51] With respect to bilingual corpora, OPUS has amassed approximately 400,000 sentence pairs, the majority of which hail from free software localisations.

Development on the first machine translation system for Occitan (Aranese Occitan-Catalan) kicked off in 2006 (Armentano-Oller & Forcada, 2006), receiving support from the Government of Catalonia as part of a joint project by the Universitat d'Alacant and Pompeu Fabra University, entitled "Open-source machine translation for Catalan", the following year. Later on, the Government of Catalonia commissioned a temporary business union (Taller Digital, from the Universitat d'Alacant, and Prompsit Language Engineering) with the creation of official translators between Occitan (Aranese and Standard) and Spanish/Catalan, since the Statute of Autonomy reform had established Occitan as Catalonia's third official language. In fact, a translators' version is still available on the Government of Catalonia website.[52]

In order to decide which model of Occitan the official system would produce, in 2007 the Government of Catalonia set up a language committee chaired by Aitor Carrera and mostly made up of leading linguists from different Occitan-speaking regions of France, the Occitan Valleys of Piedmont (which fall under the administrative remit of Italy) and the Val d'Aran. After a total of ten meetings, the committee agreed on a model in 2008 (Carrera, 2008). Although codifying the Occitan language was no untrodden path, numerous bumps still needed to be smoothed out. This resulted in coding based fundamentally on the Languedocien dialect. In fact, linguists who advocated for the alternative *mistralenc* model did not take part.

Only recently—as of 2018—has Apertium begun to develop a machine translation system that connects Occitan with the language it comes into contact with the most: French. Although many of the resources developed and made available during the construction of previous systems for Occitan were taken advantage of, the system has yet to feature a stable version.[53]

The availability of machine translation systems that work into Occitan has made it noticeably easier to produce Occitan-language content from texts originally in Spanish or Catalan. The entries on Wikipedia are a good example of this.[54]

5.3 Aragonese

Aragonese is an at-risk Romance language spoken by around 10,000 people in the Aragonese Pyrenees, especially in the Echo, Ansó and Chistàu valleys and the regions of Panticosa and Ribagorça Occidental. Differing proposals exist for its spelling system, which complicates normal use of the language.

Juan Pablo Martínez Cortés, professor of Signal Theory and Communications at the University of Saragossa, led the initiative to build the very first Aragonese-Spanish machine translator. In 2009, he contacted the Apertium community, immediately following which Jim O'Regan, a developer from the Apertium project, created an initial translator using available data on Spanish and the Aragonese inflection paradigms on

[50] Or was marketed until recently.

[51] https://www.lexilogos.com/occitan_dictionnaire.htm

[52] http://traductor.gencat.cat/text.do

[53] You can follow its development via the repository at https://github.com/apertium/apertium-oci-fra.

[54] The system is built into the Wikipedia entry translation service, Mediawiki Content Translation: https://www.mediawiki.org/wiki/Content_translation.

Wiktionary. Juan Pablo Martínez has since expanded the system with the support of other developers, such as Francis Tyers. An Aragonese-Catalan system was recently added, thereby completing the circle of languages in Aragon.[55] In 2019, Martínez spoke of his motivation for leading the initiative, saying that "making a machine translator, especially one meant to aid *non-literary* translation […], is fundamental for such an under-resourced language, as it reduces the amount of time required to translate from Spanish, […], for use by the administration, thereby awarding [Aragonese] a *quasi-official* status, […] or to help and build confidence amongst learners of Aragonese."

The spelling system chosen for Aragonese was proposed by Estudio de Filología Aragonesa (2010), although it was also used by many active initiatives, such as Wikipedia in Aragonese and Softaragonés, the latter of which develops software for the Aragonese language and distributes free software translated into Aragonese (Juan Pablo Martínez is also one of its founders). Nevertheless, when the translation system translates from Aragonese, it accepts lexical, orthographical and morphological variants that it does not generate when it translates to Aragonese. At the present time, its bilingual dictionary contains over 20,000 entries and hundreds of rules, covers 90% of the text on Wikipedia in Aragonese, and is widely used for localising software into Aragonese.

According to Martínez himself, "the translator has raised the profile of Aragonese in a range of fields and has helped to launch other projects (like Softaragonés), thereby multiplying its outward visibility". In fact, "in an Internet survey administered in 2014, of the 228 Aragonese speakers and learners in the sample, 72% reported knowing that the machine translator existed and 41% said that they had used it".

Nonetheless, in 2017 the Government of Aragon decided to follow and promote a different provisional spelling system (bearing a greater resemblance with the one proposed by the Consello d'a Fabla Aragonesa). It also gave up on the idea of building a corpus-based machine translation system, given the scarcity of bilingual corpora (some 100,000 parallel sentences in OPUS), and ended up hiring an external company to create an Apertium version using the chosen spellings.

The case of free machine translation for Aragonese illustrates the benefits of connecting a minor language with active commons, i.e. content, software and language resources like Apertium, Softaragonés and Wikipedia, to create tools that enable and foster use of the language. However, it also exemplifies the risks of regulatory fragmentation brought on by the use of free tools by stakeholders who promote alternative coding systems for minor languages that do not have a sufficiently stable system yet.

5.4 Translation between Norwegian *Bokmål* and Norwegian *Nynorsk*

The Scandinavian Germanic language that we informally refer to as *Norwegian* is a dialectal continuum with two written standards. Norwegian *Bokmål* (literally, "book language") is highly similar to Danish, as it stems from the Danish-Norwegian koine created and used by the elite when the country was politically linked to Denmark. Meanwhile, Norwegian *Nynorsk* (literally, "new Norwegian") is the fruit of an attempt to cover variants of the language that are based on original Norwegian and not so influenced by Danish. Standardisation reforms (one of which sought to create one sole Norwegian, or *Samnorsk*) have brought the two standards closer together. Without getting into too much detail, it could be said that *Bokmål* and *Nynorsk* share a level of divergence, interference and diglossia similar to that of Spanish and Catalan. An official division separates Norway into areas under the linguistic domain of *Bokmål* and areas under the linguistic domain of *Nynorsk*. This affects education, for example, as students in one area must also study the standard of the other.[56] However, *Bokmål* predominates in the media and other content, making *Nynorsk* a minor language.

Thanks to the similarities shared by *Bokmål* and *Nynorsk*, machine translation is a feasible endeavour. In fact, two systems currently exist. The first is a commercial (and, therefore, non-free) system called Nyno,[57]

[55] https://softaragones.org/traductor/index.arg.html

[56] This means that students in the *Bokmål* areas make up one of the most important groups of users of machine translation systems into *Nynorsk*, since, as the Norwegian press has expressed on multiple occasions, they use them to get *help* with their homework.

[57] The website (in Norwegian) states that, in light of divergences in the written standard, the system can translate into three different levels of *Nynorsk*: "radikal, moderat eller konservativ nynorsk" (https://www.nynodata.no/nyno).

which translates from *Bokmål* to *Nynorsk*, but not the other way round, perhaps due to a lack of demand. The second is Apertium's free, bidirectional system.[58] Beyond the fact that the latter can be used to translate from *Nynorsk* to *Bokmål*, the quality and usefulness of the two systems is otherwise comparable, the main difference being the community-based, open-source development of free linguistic data for the Apertium system. Furthermore, whereas both systems can facilitate the creation of content in a minor language[59] (*Nynorsk*, in this case), only the latter of the two has reused freely available resources and contributed to the creation of other free language resources (bilingual and monolingual dictionaries, disambiguation rules, grammatical transformation rules, etc.). For this reason, focus is placed here on the Apertium system.

Development on Apertium's *Nynorsk-Bokmål* bidirectional machine translation system began in March 2008, thanks to the availability of two free language resources (both under the GNU General Public Licence, which includes copyleft; see 3.5 above), specifically the Norsk Ordbank (literally, "Norwegian word bank") and the Oslo–Bergen Tagger disambiguation rules, both of which were converted into Apertium formats (Unhammer & Trosterud, 2009). Development continues to this day, despite matters lacking consensus, such as whether *Nynorsk* verb infinitives should end in *-e* or in *-a*. When translating news, currently only 5% of the words in raw translations produced by the system require post-editing.

Apertium's *Nynorsk-Bokmål* translator is used on a massive scale by students, by news agencies NTB and NPK, and for the creation of Wikipedia entries (via the content translation tool mentioned above in reference to Aragonese). Some institutions, such as the Språkrådet (The Language Council of Norway), have voiced their concern regarding whether machine translation may actually be curtailing *Nynorsk* variety.

5.5 The Northern Saami to Norwegian *Bokmål* translator

North or Northern Saami is the most widely used language amongst the Saami people[60]—with roughly 20,000 speakers—in the northern parts of Norway (where it has official status), Sweden and Finland (where it is recognised as a minor language). The Saami languages are grouped together in the Finno-Ugric language family. The language has a regulatory body called the Giellagáldu (literally, "Source of the Language").

The only machine translation system for Northern Saami is based on Apertium. It translates unidirectionally from Northern Saami to Norwegian *Bokmål*, in order to encourage the creation of content directly in Saami (on social media, for example), which can later be machine translated for users of *Bokmål*.

Development on the system began in 2010, led mainly by Trond Trosterud and Lene Antonsen, two members of Giellatekno, the UiT – Arctic University of Norway's research group for Saami language technology,[61] and Apertium developer Kevin Unhammer. It was based on free language resources from Giellatekno, namely morphological analysers and generators. In this case, the two languages are very different, so a large part of the completed development (and of what remains ahead) deals with writing syntactic transformation rules. The language resources generated can serve, for example, to create new machine translation systems between Saami and *Nynorsk* or Swedish, both of which are similar to *Bokmål*.

Few data exist regarding use of Apertium's Northern Saami-*Bokmål* system, but all suggest a rather intense use if the number of speakers is taken into consideration. The Arctic University of Norway's Northern Saami-*Bokmål* machine translator[62] receives approximately a hundred visits every day, with roughly five *actions* (translations) carried out per visit. The activity on this translator combined with that taking place on the Apertium's main website, which also offers this language pair, gives the total sum of machine translation activity involving Northern Saami.

[58] In fact, it is one of the most widely used systems on Apertium, especially by students.

[59] For example, Nyno helped to localise Microsoft software for users of *Nynorsk* (Unhammer, 2009).

[60] Saami has replaced the more traditional term *Lapp*, the use of which is now deprecated.

[61] http://giellatekno.uit.no

[62] http://jorgal.uit.no/index.sme.html?dir=sme-nob

5.6 The Government of Valencia's machine translator

The Catalan language is not currently as *minor* as the other five languages analysed in this section, as it has more resources, including several machine translation options. The aim of this sub-section is thus to illustrate how choosing a certain development and data access model can jeopardise a useful system in which substantial amounts of public money have been invested.

Despite being virtually finished in 1997, the Government of Valencia waited until 2000 to publish SALT, the first relatively widely accessible Spanish-Catalan (Valencian) machine translation system.[63,64] SALT, which was developed by a team headed up by Josep Lacreu, was initially distributed freely on floppy disks, although only as an executable, and was meant to be a system that could aid people in writing texts in Valencian, requiring only a certain level of interactivity from the user in cases of ambiguity. Neither the linguistic data that the system drew from, which included excellent dictionaries, nor the software's source code were made public. Subsequent versions (up to version 4.0) were distributed on CD-ROMs or as downloadable packages from the Internet, as well as being made available for use directly online, initially with some restrictions.

The 2015 Valencian elections ushered in a new Council of the Government of Valencia, leading to changes in the personnel responsible for SALT. During the process, the outgoing staff *forgot* the access passwords for the databases in which the linguistic data was stored, and the new staff members in charge were unable to continue development. Fortunately, there were very complete free data for Spanish-Valencian translation on Apertium, and it was therefore not necessary to start all over again.[65] With support from Prompsit Language Engineering and Apertium's community of developers, the Government of Valencia staff members were able to continue development themselves. SALT, as it stands today, is a brand-new software development: it is accessible via the Internet,[66] available as a mobile app for Android and iPhone, and based on the Apertium platform. The linguistic data are completely free and form part of the general language resources (apertium-spa [Spanish],[67] apertium-cat [Catalan][68] and apertium-spa-cat [Spanish-Catalan][69]) made available by Apertium,[70] where they have been added to and adapted to Government of Valencia uses.[71] Preservation of the resources used to create the current SALT system has been stepped up thanks to a public repository hosted on GitHub, a leading software development platform, away from the vulnerable situation that forced the technological shift.

6 Conclusions

Machine translation is one of the computer tools with the greatest potential to enhance a language's online presence and, therefore, to bolster its vitality in today's world, where communication is increasingly more digital. Machine translation, which can be harnessed to understand and create content written in another language, may *use language resources*, such as dictionaries and rules drafted by experts, or *learn from corpora* containing a huge number of already translated sentences. If gathering a sufficiently large collection of translated sentences proves impossible, which is all too often the case for *minor* languages, linguistic communities have to find a way to build, maintain and publish the necessary language resources under licences that will maximise their positive effect on the disadvantaged languages they are hoping to bolster. In light of all this, following a brief introduction on the uses and technologies of machine translation, the paper describes the general challenges faced by those working to build machine translation systems for minor

[63] The system could be used in automatic or interactive mode; in the latter, any ambiguities that would appear during the translation of the text could be cleared up manually.

[64] One of the reasons behind the delay was a lack of consensus regarding which Valencian model the system should produce. The acronym stands for Servei d'Assessorament Lingüístic i Traducció (Language Advice and Translation Service), put forth by the Government of Valencia.

[65] Even though some linguistic data could have been recovered by reverse engineering SALT 4.0.

[66] http://www.salt.gva.es/

[67] https://github.com/apertium/apertium-spa

[68] http://github.com/apertium/apertium-cat

[69] http://github.com/apertium/apertium-spa-cat

[70] Dictionary entries that diverge from the general model are marked with the tag "val_gva".

[71] http://www.salt.gva.es/va/criteris

languages, as well as the effects that these systems can have on the minor languages involved. Finally, case studies on five minor and one not-so-minor European languages are presented in more detail. In all cases, the machine translation systems are based on free language resources and built upon the same free machine translation platform: Apertium. Although some systems have faced specific challenges, such as those related with standardisation, most have had a positive effect on their minor languages. Most importantly, they have increased online presence and the creation of language resources that can be used to aid other languages in need.

Acknowledgements: I would like to thank Juan Pablo Martínez, from the University of Saragossa; Trond Trosterud, from UiT – The Arctic University of Norway; Kevin Brubeck Unhammer, from Trigram AS; Robert Escolano, from the Universitat d'Alacant; Fulup Jakez, from the Public Office for the Breton Language; Gema Ramírez-Sánchez, from Prompsit Language Engineering, and Carme Armentano i Oller, a freelance translator and computational linguist based in Lausanne, Switzerland. I would also like to thank the reviewers of the paper for their insightful suggestions.

Reference list

Armentano-Oller, Carme, & Forcada, Mikel L. (2006). Open-source machine translation between small languages: Catalan and Aranese Occitan. In *Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages)* (pp. 51–54) [organised in conjunction with the LREC 2006 (22-28.05.2006)].

Carrera, Aitor. (2008). Acòrds dera Comission Lingüistica deth traductor automatic catalan-occitan occitan-catalan". (164 p.).

Casacuberta Nolla, Francisco, & Peris Abril, Álvaro. (2017). Traducción automática neuronal. *Tradumática: Tecnologies de la Traducció*, *15*, 66–74.

Forcada, Mikel. (2006). Open source machine translation: an opportunity for minor languages. In *Proceedings of the Workshop "Strategies for developing machine translation for minority languages", LREC* (vol. 6) (pp. 1-6).

Estudio de Filología Aragonesa.. (2010). Propuesta ortografica provisional de l'Academia de l'Aragonés. Zaragoza: Edicions Dichitals de l'Academia de l'Aragonés

Eurobarometer Special. (2012). Europeans and their Languages. European Commission.

Forcada, Mikel. L. (2017). Making sense of neural machine translation. *Translation Spaces*, *6*(2), 291-309.

Jakez, Fulup. (2019). [personal communication].

Martínez, Juan Pablo. (2019). [personal communication].

Moorkens, Joss, & Lewis, Dave. (2019). Research questions and a proposal for the future governance of translation data. *Journal of Specialised Translation*, *32*, 2-25.

Nurminen, Mary. (2019). Decision-making, risk, and gist machine translation in the work of patent professionals. In *Proceedings of the 8th Workshop on Patent and Scientific Literature Translation* (pp. 32-42).

Rivera Pastor, Rafael, Tarín Quirós, Carlota, Villar García, Juan Pablo, Badia Cardús, Toni, & Melero Nogués, Maite. (2017). *Language equality in the digital age: Towards a human language project* [Report IP/G/STOA/FWC/2013-001/Lot4/C2]. European Parliament.

Sánchez-Cartagena, Víctor. M., Sánchez-Martínez, Felipe, & Pérez-Ortiz, Juan Antonio. (2011). Enriching a statistical machine translation system trained on small parallel corpora with rule-based bilingual phrases. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011* (pp. 90-96).

Sánchez-Cartagena, Víctor M., Forcada, Mikel L., & Sánchez-Martínez, Felipe. (2020). A multi-source approach for Breton-French hybrid machine translation [Accepted for presentation at the 22nd Annual Conference of the European Association for Machine Translation (EAMT 2020), 3-5 November 2020].

Streiter, Oliver, Scannell, Kevin. P., i Stuflesser, Mathias. (2006). Implementing NLP projects for non-central languages: instructions for funding bodies, strategies for developers. *Machine Translation*, *20*(4), 267-289.

Tyers, Francis M. (2009). Rule-based augmentation of training data in Breton-French statistical machine translation. In *Proceedings of the 13th Annual Conference of the European Association of Machine Translation EAMT09* (pp. 213-218).

Tyers, Francis M. (2010). Rule-based Breton to French machine translation. In *Proceedings of the 14th Annual Conference of the European Association of Machine Translation* (pp. 174-181).

Unhammer, Kevin. (2019) [personal communication].

Unhammer, Kevin, & Trosterud, Trond. (2009). Reuse of free resources in machine translation between Nynorsk and Bokmål. In Pérez-Ortiz, Juan Antonio, Sánchez-Martínez, Felipe, & Tyers, Francis M. (Coords.) *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation* (pp. 35–42) [Congress held at the Universitat d'Alacant in November 2009].

Williams, Briony, Nadeu, Climent, Sarasola, Kepa, Ó'Cróinin, Donncha, & Petek, Bojan. (2001). Speech and language technology for minority languages. In *Proceedings of Eurospeech 2001*.