Facilitating the implementation of the European Charter for Regional or Minority Languages through artificial intelligence

Council of Europe Secretariat
of the European Charter
for Regional or Minority Languages

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

Publication by the Council of Europe
Author: Miriam Gerken
Conception and editing: Secretariat of the European Charter for Regional or Minority Languages
Cover page image: Shutterstock
MIN-LANG(2022)4
February 2022

Contents

Introduction **4**

1. General reasons for using artificial intelligence (AI) to facilitate the implementation of the Charter 5

2. Machine translation 7

2.1 Different methods for machine translation **7**

2.2 Use of already existing applications **8**

2.3 Developing new applications **9**

3. Other natural language processing (NLP) applications and their use for the implementation of the Charter 10

3.1 Use of regional or minority languages in private life (Article 7.1.d) **10**

3.2 Use of regional or minority languages in education (Articles 7.1.g, 8.1.f.ii, iii) **11**

3.3 Use of regional or minority languages by judicial authorities (Articles 9.1.a.i-iv, 9.1.b.i-iii, 9.1.c.i-iii, 9.1.d, 9.3) **13**

3.4 Use of regional or minority languages by administrative authorities and public services **13**

3.4.1 Chatbots (Articles 10.1.a.i-iv, 10.2.a, 10.2.b, 10.3.a-c) 13

3.4.2 Smart search (Article 10.2.g) 14

3.4.3 Speech synthesis for street name announcements (Article 10.2.g) 15

3.4.4 Machine translation (Articles 10.1.a.i-v, 10.1.b, 10.1.c, 10.2.a-f, 10.3.a-c, 10.4.a) 15

3.5 Use of regional or minority languages in the media **15**

3.5.1 Automatic generation of subtitles (Articles 11.1.a.i-iii, 11.1.c.i-ii) 16

3.5.2 Automatic information extraction (Article 11.1.e.i-ii) 16

3.6 Use of regional or minority languages in cultural activities and facilities **17**

3.6.1 Data structuring (Articles 12.1.g, 12.1.h) 17

3.6.2 Machine translation (Articles 12.1.a, 12.1.b, 12.1.c) 18

3.6.3 Automatic generation of subtitles (Articles 12.1.b, 12.1.c) 18

3.7 Use of regional or minority languages in economic and social life **18**

3.7.1 Sentiment analysis (Article 13.1.c, 13.1.d, 13.2.b) 18

3.7.2 Machine translation (Articles 13.1.a, 13.1.d, 13.2.a, 13.2.b, 13.2.d, 13.2.e) 19

3.8 Use of regional or minority languages in transfrontier exchanges (Articles 7.1.i, 14) **19**

Outlook **21**

Acknowledgements

This report was researched and written by Miriam Gerken, in co-operation with the Secretariat of the European Charter for Regional or Minority Languages, during her study visit at the Council of Europe in February 2020 and updated in March 2022. Miriam Gerken graduated in translation studies in 2018 and in computational linguistics in 2021 at the Universities of Hildesheim and Bielefeld (Germany), both with a focus on machine translation. She now works as an expert for language technologies at Deutsche Bahn AG. The Charter Secretariat is grateful to Miriam Gerken who, thanks to her in-depth knowledge of language use in the context of AI, has produced a comprehensive and practical guide for policymakers and practitioners working on the implementation of the Charter and the promotion of the daily use of regional or minority languages in public and private life.

Introduction

In most European states, regional or minority languages have been historically used in parts of the territory by a minority of the population. The demographic and legal situation of such languages varies greatly. However, what many have in common is a greater or lesser degree of precariousness.

The legal frame of reference worldwide for the promotion of these languages is the European Charter for Regional or Minority Languages of the Council of Europe. The Charter promotes the use of regional or minority languages in different fields of public life: education, judicial authorities, administrative authorities and public services, media, cultural activities and facilities, economic and social life, and transfrontier exchanges. The Charter has been ratified by 25 states;[1] a further nine states have signed, but not yet ratified it.[2]

The drafting of the Charter began in 1984. Since the adoption of this treaty by the Committee of Ministers of the Council of Europe in 1992, different new technologies have changed the conditions for its implementation by the states parties. For example, the internet has helped to increase the media offer and made requests for radio frequencies or more convenient broadcasting times for programmes in regional or minority languages to some extent obsolete.

Due to the rise of AI technologies, the states parties are facing new opportunities, but also new responsibilities for implementing the Charter. Artificial intelligence – a term that used to be science fiction has arrived in the everyday life of many people today: alerting them to leave their house sooner when there is a lot of traffic on their commute, suggesting what kind of movies they probably want to watch next or enabling them to read messages in languages they do not know. But what exactly is artificial intelligence? And how can it provide solutions for a specific real-life problem, the support of regional or minority languages?

AI generally describes intelligent machines that are able to analyse their environment and make decisions based on these analyses. AI systems are supposed to mimic human abilities, like learning or problem-solving through statistical data processing in neural networks. The use of AI for handling language problems is called natural language processing (NLP). NLP is one of the main sub-fields of AI and computer sciences. It is concerned with human-computer-interaction through natural languages. The goal of NLP is developing programmes that can read, process, analyse and ultimately understand natural languages in all their complexity. To achieve this, large amounts of natural language data are necessary.

The Council of Europe supports states parties in implementing the Charter. The present report contributes to this effort. Its goal is to show how different NLP applications, and thereby AI, can facilitate the everyday use and promotion of regional or minority languages and hence support states parties in implementing the Charter provisions which they have ratified. In so doing, the report follows the structure of the Charter, focusing on its Articles 7-14.

The Framework Convention for the Protection of National Minorities of the Council of Europe contains rights of persons belonging to national minorities in different fields, including language and culture. It

[1] Armenia, Austria, Bosnia and Herzegovina, Croatia, Cyprus, the Czech Republic, Denmark, Finland, Germany, Hungary, Liechtenstein, Luxembourg, Montenegro, the Netherlands, Norway, Poland, Romania, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Ukraine and the United Kingdom.

[2] Azerbaijan, France, Iceland, Italy, Malta, Republic of Moldova, Portugal, Russian Federation and North Macedonia.

has been ratified by 39 states, several of which have not yet ratified the Charter. Considering that measures supporting the implementation of the Charter will also have a positive effect on the implementation of the linguistic rights in the Framework Convention, this report will refer to relevant provisions of the latter.

At the Council of Europe, the Ad hoc Committee on Artificial Intelligence (CAHAI) has examined possible elements of a legal framework on artificial intelligence, based on the Council of Europe's standards on human rights, democracy and the rule of law. During its work from 2019 to 2021, the committee has noted that AI applications may enable persons speaking minority languages to participate more actively in public life, debate, deliberation or decision making, but also raise awareness of such languages, notably it was pointed out that using minority languages in AI applications could be a way to avoid linguistic discrimination.[3]

The present report might thus also contribute to the Council of Europe's wider reflections about AI.

1. General reasons for using AI to facilitate the implementation of the Charter

Before presenting the specific ways in which AI can support the implementation of individual Charter provisions, this section shows how it can help to achieve the overall objectives of the treaty.

Researchers refer to **"Digital Language Extinction"** when a language fails to broaden its reach to new means of communication and technology. This concerns not only less-widely used ("small") languages, but also languages that, while being the majority language of a country, are in a threatened minority situation in another country. To face this challenge, strong commitment to introduce regional or minority languages to new technologies is needed. If used in the right way, new means of technology and communication entail huge possibilities to promote these languages.

The applications described in this report are all easy to use, hands-on applications for real life situations. Many of them help to introduce regional or minority languages to larger audiences. If regional or minority languages can be used in common AI applications like smart home assistants, the possibilities for using the language in daily life are significantly increased. In addition, new language learners are motivated and encouraged, which contributes to increasing the number of users of such languages. The presence of regional or minority languages on different internet platforms also increases their visibility. One example is Catalan, which belongs to the 20 most used languages on Wikipedia.

With the help of AI, authorities can, at a relatively low cost, quickly make available an offer for users of regional or minority languages. That way, AI supports authorities in taking "resolute action to promote regional or minority languages in order to safeguard them", which is one of the central objectives and principles of the Charter (Article 7.1.c).

All applications mentioned in this report are transnational and can be used in other countries where the same regional or minority language is spoken. Once an application is developed for one area and one language, it can easily be adapted to other areas and other regional or minority languages. As

[3] See Ad-hoc Committee on Artificial Intelligence (CAHAI): Feasibility Study, CAHAI(2020)23, p. 19; Analysis of the Multi-Stakeholder Consultation, CAHAI(2021)07, p. 22; Compilation of responses to the Multi-Stakeholder Consultation (F to M), CAHAI(2021)06, p. 80.

described in more detail below, AI thereby promotes exchanges and connections between groups from different states using the same regional or minority language, as well as between authorities.

The fields of study concerned with AI and NLP are relatively new and future-oriented. Including regional or minority languages into these research fields is therefore of significant help to their survival and development as living languages. The promotion of research on such languages is foreseen in Article 7.1.h, which applies to every regional or minority language used in a state party. On several occasions, this report will point out different starting points for future research. The most important basis for developing useful NLP applications is, however, the gathering of relevant natural language data of the regional or minority languages in text **corpora** (large data sets of different texts in one or more language(s), depending on the NLP task at hand). Without sufficient data, the development of working NLP applications is not possible. Therefore, the gathering of natural language data of regional or minority languages should be the first step to promote further study and research.

There are already several relevant research projects and groups established. One example is the working group "SALTMIL" (speech and language technology for minority languages), which follows the goal of promoting research and development in the field of speech and language technology for lesser-used languages, especially of Europe. Another research group is "Ixa" at the University of the Basque Country (Spain), which aims to develop and provide algorithms, tools and linguistic resources that allow computers to process and understand human languages, especially Basque. The group has already published several corpora and NLP tools for Basque, like a wordnet (lexical database of semantic relations between words), a spell checker and a parser (programme to analyse the syntax of sentences) as well as tools for information retrieval and extraction, machine translation and language learning. As a third example, the University of Helsinki/Helsingfors is conducting study projects regarding language technologies like treebanking (compiling a text corpus with information about the sentence's syntactic and semantic structures) for different minority languages of Finland.

A positive example of government support for language technologies is the Welsh language technology plan of 2018. It forms part of the Welsh government's general strategy to promote the everyday use of Welsh and to increase the number of active speakers. The plan aims to encourage technological developments and set the direction for work on Welsh language technology. To achieve this goal, the Welsh government is funding many different projects, like the coding of Welsh language learning applications for children, the expansion of OpenStreetMap (crowdsourced, free and opensource world map) with a Welsh version, establishing a Welsh speech database or the development of Macsen, a Welsh language digital assistant. The government also encourages the use of already existing parallel corpora and other tools like stopword lists (list of frequently used words of a language, like prepositions or articles) or parser for research projects. In addition, it organises annual conferences for researchers at Bangor University (Wales) and creates vacancies at universities reserved for this area of research. These measures are able to raise considerable awareness for present and future Welsh language technologies and this language in general. Given their positive impact, they could function as a model for other countries.

2. Machine translation

Machine translation technologies can be of significant use for the implementation of the Charter. Many of its undertakings are concerned with the translation of different types of documents into regional or minority languages. These undertakings could be implemented with a lower time and financial effort by using machine translation. Consequently, the following sections will focus on machine translation, its different methods, which applications already exist, which regional or minority languages they support and how it is possible to develop new translation applications for specific domains.

2.1 Different methods for machine translation

The term “machine translation” describes a software that automatically translates a text from one language into another. It is one of the most researched aspects of **computational linguistics**. Machine translation is a complex challenge since the software must be able to decode the meaning from the input or source language and then encode the meaning in a new way in the output or target language. However, to decode the meaning of a phrase, semantic understanding of a language would be necessary, which is not (yet) achievable - only natural language processing is possible, not natural language understanding. This missing semantic understanding of the software gives rise to various problems regarding the translation process: ambiguity (lexical and syntactical), translation of proper names (for example of institutions) and structural difficulties, not to mention cultural differences from source to target language. All these problems can be traced back to the evident fact that translation is not just transmitting words into another language, but rather conveying the meaning of phrases, while taking the background of both languages into account.

However, there are already domains in which machine translation can achieve good results. This applies mostly to topics in which standard, formulaic language with similarly structured sentences and unambiguous wording is used. Examples for these topics are weather reports or application forms.

There are three main methodological approaches to machine translation.

The first is **rule-based machine translation**. For this method, a linguistic description with a pre-defined lexicon and explicit rules on how to form a grammatical sentence in the source and target language are needed. Due to the complexity of these descriptions and the unreliability of the translation results, rule-based machine translation is mostly outdated.

The second approach is **statistical machine translation**. Here, the software uses translated data for both input and output language, called parallel corpora, and derives probabilities from it, for example that the word “movie” is often followed by the word “was”. When given a new input, the algorithm uses these probabilities to find the best translation from a statistical perspective, which is the translation with the highest probability. This approach is easier to set up than the first one because it does not need manually crafted rules, only parallel corpora. Statistical machine translation cannot achieve convincing results for all language pairs, though, especially for those with largely varying syntax.

The third approach is the one that is used in most machine translation applications today. It is called **neural machine translation** and uses AI and neural networks to “learn” the pairing of sentences, similar to how biological neurons learn relevant connections. In the neural machine translation process, a network first encodes the input sentence into so called **word vectors**. These are vector representations

that allow the system to transfer semantic relationships between words to a mathematical space. A distance in meaning between two words is therefore captured through a difference in values between their vector representations. For example, by subtracting the values of the word vector for “man” from the word vector for “king” and adding the vector values for “woman”, one would get a vector that is identical to the word vector for “queen”. This means significant progress for machine translation and is one of the reasons why neural machine translation achieves good results today. After this encoding process, the network decodes the word vectors to the output language. This is made possible through the use of the already mentioned neurons that have “learned” the connections between input and output language through parallel corpora, similar to the statistical machine translation approach. The method of neural networks, consisting of connected neurons, learning from data is called **training**. Neural machine translation is able to achieve better results than ever before while also using less memory than statistical machine translation. For these reasons, neural machine translation is widely used by most machine translation applications today.

2.2 Use of already existing applications

Many machine translation applications used today already support various regional or minority languages that are covered by the Charter. The following table lists the most important ones and the supported languages in alphabetical order.

MT application	Supported regional or minority languages	Number of languages
DeepL	Bulgarian, Czech, Danish, Finnish, French, German, Greek, Hungarian, Italian, Lithuanian, Polish, Romanian, Russian, Slovakian, Slovenian, Swedish	16
Google Translate	Albanian, Armenian, Basque, Belarusian, Bosnian, Bulgarian, Catalan, Croatian, Czech, Danish, Finnish, French, Frisian, Galician, German, Greek, Hungarian, Irish, Italian, Kurdish, Lithuanian, Macedonian, Polish, Romanian, Russian, Scottish-Gaelic, Serbian, Slovakian, Slovenian, Swedish, Tatar, Turkish, Ukrainian, Welsh, Yiddish	35
Microsoft Translator	Bosnian, Bulgarian, Catalan, Croatian, Czech, Danish, Finnish, French, German, Greek, Hungarian, Irish, Italian, Kurdish, Lithuanian, Macedonian, Polish, Romanian, Russian, Serbian, Slovakian, Slovenian, Swedish, Tatar, Turkish, Ukrainian, Welsh	27
PROMT	Finnish, French, German, Greek, Italian, Russian, Tatar, Turkish, Ukrainian	9
Watson Language Translator (IBM)	Basque, Bosnian, Bulgarian, Catalan, Croatian, Czech, Danish, Finnish, French, German, Greek, Hungarian, Irish, Italian, Lithuanian, Polish, Romanian, Russian, Serbian, Slovakian, Slovenian, Swedish, Turkish, Ukrainian, Welsh	25
Yandex	Albanian, Armenian, Basque, Belarusian, Bosnian, Bulgarian, Catalan, Croatian, Czech, Danish, Finnish, French, Galician, German, Greek, Hungarian, Irish, Italian, Lithuanian, Macedonian,	33

	Polish, Romanian, Russian, Scottish-Gaelic, Serbian, Slovakian, Slovenian, Swedish, Tatar, Turkish, Ukrainian, Welsh, Yiddish	

This list is in no way exhaustive. There are several other companies offering machine translation services, and since the market is rapidly changing, new companies could quickly change its landscape. For example, DeepL was only introduced in 2017 and has quickly gained a good reputation.

All services listed above are accessible online with a standard web browser. DeepL and PROMT also offer a desktop version of their product, which makes offline translation possible. However, the desktop version of PROMT is not free of charge.

To use the services of Microsoft Translator and Watson Language Translator it is necessary to create an account. The use remains free of charge, although in the case of Watson Language Translator only until reaching a specific translation word count.

These different applications have varying levels of accuracy depending on the chosen source and target language. For some languages, especially those with fewer data available, these translation services may only be useful to get a basic understanding of what a text or document is about to decide if further translation by a professional translator is needed. This process is called **"gisting"**. As could be seen from the information above, all applications have different focus, advantages, and disadvantages. For this reason, it is not possible to recommend one solution for all purposes.

2.3 Developing new applications

Although the existing machine translation services already cover a wide range of regional or minority languages, there are several languages that lack such a service. However, machine translation is also possible for these languages.

As mentioned in the first section, machine translation works best if used for a restricted topic, with limited vocabulary and formulaic sentences. Since this is the case for many translations referred to in the Charter (legal documents, administrative forms, financial documents etc.), it is possible to develop a machine translation service for these domains that will probably provide convincing results.

To develop such a system, the main prerequisite is parallel data of the same texts in the source and target language, preferably translated by a professional translator. Next, one could either train an already working programme with the data (there are several services where one can put in own training data into an already existing programme) or develop one's own translation service by following one of the many tutorials that can be found online. This could also be an opportunity to promote research and study on regional or minority languages, since, given the available data, developing such a system is a feasible task (e.g. for student groups from different programming or computational studies).

3. Other natural language processing (NLP) applications and their use for the implementation of the Charter

3.1 Use of regional or minority languages in private life (Article 7.1.d)

Today, a large part of social life, especially of the younger generation, takes place online on **social media platforms**, which incorporate many AI technologies into their services. Especially during the Covid-19 pandemic, social media platforms have been an important tool for communication and maintaining social connections. In order to protect and promote regional or minority languages, it is only natural to also look at social media platforms and encourage the use of such languages there. In fact, many social networks are already available in regional or minority languages. The supported languages of the three most important social media platforms can be seen in the table below.

Network name	Supported regional or minority languages	Number of languages
Facebook	Albanian, Armenian, Basque, Belarusian, Bosnian, Bulgarian, Catalan, Croatian, Czech, Danish, Finnish, French, Frisian, Galician, German, Greek, Hungarian, Irish, Italian, Kurdish, Lithuanian, Macedonian, Polish, Romanian, Russian, Serbian, Slovakian, Slovenian, Swedish, Tatar, Turkish, Ukrainian, Welsh	33
Instagram	Bulgarian, Croatian, Czech, Danish, Finnish, French, German, Greek, Italian, Polish, Romanian, Russian, Serbian, Slovakian, Swedish, Turkish, Ukrainian	17
Twitter	Czech, Danish, Finnish, French, German, Greek, Hungarian, Italian, Polish, Romanian, Russian, Swedish, Turkish, Ukrainian	14

There are four main possible actions authorities and individuals can take to make use of social media in regional or minority languages. The first one is to use social media platforms in such languages and to encourage young people, for example in regional or minority language learning classes, to do the same. This way, the language becomes an active part of their everyday (online) social life and the language's timeliness, reach and visibility are increased. Secondly, speakers of regional or minority languages that already have a translation on social media platforms can contribute to improve these translations, for example through online translation fora. Especially on Facebook, there is a big **translation community** that helps to incorporate more regional or minority languages into the platform's structure. Users can submit their own translations or improve the translations that are already given. Thirdly, speakers of regional or minority languages which do not yet have a translation can request it and encourage others to do the same, in order to show the social networks that there is a need for translations in this specific language. This is an important opportunity for advocates of regional or minority languages to support and encourage the language's use on different social media platforms. Lastly, social media platforms are also a new way to connect different groups of regional or minority language speakers or learners, especially from different states. Groups can organise meetings or initiatives, communicate in the regional or minority language and simply network. This way, transfrontier exchanges can be facilitated, encouraged, and promoted.

3.2 Use of regional or minority languages in education (Articles 7.1.g, 8.1.f.ii, iii)

Ensuring the teaching and learning of a language is one of the most important measures, if not the most important one, to provide for its understanding, use in various social situations and continued existence. With different **platforms of online learning**, AI can facilitate the **learning of regional or minority languages**. Especially during the Covid-19 pandemic, it became clear that online learning platforms are a vital tool in which such languages should be included. They may also contribute to the implementation of Article 14.2 of the Framework Convention.

Online learning platforms are either websites or mobile applications directed at people who want to learn or improve their skills in a language. The different platforms use various types of media (texts, sound recordings etc.) and often also methods of **gamification** (use of game elements, like rewards or competition, in non-game activities) to make the learning process entertaining and efficient for the users. Moreover, most platforms use the learner's data to improve their products. Many of the existing platforms already support various regional or minority languages, as can be seen in the following table.

Application name	Supported regional or minority languages	Number of languages
Beelingu	French, German, Italian, Russian, Swedish, Turkish	6
Busuu	French, German, Italian, Polish, Russian, Turkish	6
Clozemaster	Albanian, Armenian, Basque, Belarusian, Bulgarian, Catalan, Cornish, Croatian, Czech, Danish, Finnish, French, Galician, German, Greek, Hungarian, Irish, Italian, Lithuanian, Macedonian, Polish, Romanian, Russian, Scottish-Gaelic, Serbian, Slovakian, Slovenian, Swedish, Turkish, Ukrainian, Welsh, Yiddish	32
Drops	Bosnian, Croatian, Danish, Finnish, French, German, Greek, Hungarian, Italian, Polish, Russian, Serbian, Swedish, Turkish	14
Duolingo	Czech, Danish, Finnish, French, German, Greek, Hungarian (testing status), Irish, Italian, Polish, Romanian, Russian, Scottish-Gaelic, Swedish, Turkish, Ukrainian, Welsh, Yiddish (testing status)	18
Memrise	Danish, French, German, Italian, Polish, Russian, Slovenian, Swedish, Turkish	9
Mondly	Bulgarian, Catalan, Croatian, Czech, Danish, Finnish, French, German, Greek, Hungarian, Italian, Lithuanian, Polish, Romanian, Russian, Slovakian, Swedish, Turkish, Ukrainian	19

These learning platforms are all accessible as mobile applications. Except for Beelingu and Drops, all platforms also offer their services as a website. To use the applications, one has to create an account so that the learning process can be saved, and the user can resume at the last saving point. All platforms mentioned above are free of charge, although some offer a premium membership at the user's expense that e.g. stops the show of advertisements between learning sessions. There are also

several other fee-based language learning offers, like Babbel or Rosetta Stone, which are not included in the list above since a required payment may exclude some language learners.

There are additional aspects to note about the learning platforms mentioned above:

- **Beelingu** is a special language learning application that does not offer normal courses but instead side-by-side reading and listening of the learner's language and the language the learner wants to learn.
- **Clozemaster** offers a wide variety of languages. However, it can mainly be recommended to advanced learners because it does not offer normal language courses, but the learning of new vocabulary by inserting it in different contexts.
- **Drops** has a special language learning app for children called "Droplets". It teaches the languages with special drawing games and without ads. Additionally, parents can configure that in-app purchases require a second password.
- **Duolingo** is probably the most relevant language learning application in the present context because it offers high quality courses in many regional or minority languages with several new ways of language teaching, like storytelling or podcasts for language learners. Moreover, there is a special Duolingo version for the use of language teaching at schools. Teachers can create classrooms with the accounts of their pupils, monitor their learning process and assign special tasks.
- **Memrise** offers so called "community courses" that are created by users. There are already several regional or minority languages covered by these community courses , like Albanian, Armenian, Basque, Catalan, Finnish, Greek, Hungarian, Irish, Lithuanian, Romanian, Scottish-Gaelic, North Sami, or Welsh. To incorporate additional languages would be simple since every registered member can create his/her own language course that can then be used by every other user of the platform.
- **Mondly** also has a special language learning app for children called "Mondly Kids". It teaches the languages with simple, child appropriate gamified lessons.

In addition to these more classic language learning applications, there are also several applications (like HelloTalk) that connect language learners with native speakers of the language so that they can talk or chat together. These **online language tandem applications** may also be of interest for regional or minority language learners.

One major disadvantage of these learning platforms is that not all combinations of learning language and learned language exist. Most platforms only offer a few learning languages, mostly English. Therefore, online language learning platforms can so far only support regular education, not replace it. The different learning applications are, however, of interest for older language learners from other countries who already have a working knowledge of e.g. English and want to learn a regional or minority language.

To sum up, there are already many language learning services offering and supporting regional or minority languages. The use of their services has many advantages, like enabling people who are not living in the region or even the country to learn and master regional or minority languages and enhancing the visibility of such languages. The states parties can make use of these offers by either promoting the availability of already existing online language courses through these applications or by creating new language courses or vocabulary cards for regional or minority languages that are not yet covered by a language learning application on the learning platforms. This would make the language learning materials available to a large number of new potential learners and therefore increase the language's visibility.

3.3 Use of regional or minority languages by judicial authorities (Articles 9.1.a.i-iv, 9.1.b.i-iii, 9.1.c.i-iii, 9.1.d, 9.3)

Article 9 is concerned with the use of regional or minority languages in one of the core areas of the state, namely judicial authorities. In this context, NLP applications, notably machine translation, can be of use for the processing of documents in regional or minority languages. Machine translation is of particular interest regarding Article 9.1.d since it can reduce the high cost often linked to professional translation.

How machine translation works and what kind of applications exist has already been discussed in the previous section. As described there, machine translation of judicial documents could be highly promising, due to their formulaic structure and language. An investment in this area could therefore lead to valuable results.

Machine translation can also be useful if the already existing machine translation applications are used for **"gisting".** This can also be done for regional or minority languages with a low level of machine translation accuracy.

Machine translation may also contribute to the implementation of Article 10.3 of the Framework Convention.

3.4 Use of regional or minority languages by administrative authorities and public services

NLP applications can facilitate the use of regional or minority languages by, and in communication with administrative authorities and public services. Relevant applications are chatbots, different methods of smart search, the use of speech synthesis to generate street name announcements and, again, machine translation. The applications mentioned in this section may also contribute to the implementation of Article 10.2 (use of minority languages with authorities) and 11.3 (topographic names) of the Framework Convention.

3.4.1 Chatbots (Articles 10.1.a.i-iv, 10.2.a, 10.2.b, 10.3.a-c)

Chatbots are software programmes that conduct conversation in order to cater for customer requests and queries directly. They are able to answer standard questions or refer the customer to administrative personnel. In theory, these **conversational agents** should mimic written or spoken human speech for the purposes of simulating a conversation and interaction with a real person. They achieve this by first processing the text input from the customer and afterwards responding based on an algorithm that interprets what the customer said and then determines an appropriate answer. There are chatbots that can recognise key words and phrases and then give an output of pre-prepared or pre-programmed responses. This method is especially useful for straight-forward information that can be classified into predictable categories. There are also chatbot programmes that are able to learn new responses based on customer interaction through the use of AI. Different ways to run a chatbot software exist: on popular messaging applications, via SMS, on stand-alone applications or websites. Many companies offer their services to programme chatbots. When one already has pre-written dialogues, there are also websites that make it possible to create simple chatbots using an intuitive graphic interface with drag-and-drop gestures.

Chatbots are a popular and widely used method. They are able to automate simple request tasks through a conversational interface, which makes the automation appear less mechanical and more human. Chatbots are used in many areas of **customer interaction**. They can be a support for administrative authorities since the user can make appointments, download forms, get answers to standard questions, set a reminder for appointments, or get important information about e.g. opening hours through them. There are already many administrative authorities that use chatbots to extend and simplify their work: the cities of Berlin, Bonn, and Würzburg in Germany for example, as well as different administrative authorities in Finland (Immigration Service, Tax Administration, Patent and Registration Office) or the Transport Office for London in the United Kingdom. The European Commission has published a report on chatbots for this domain, called "Architecture for public service chatbots", to help administrative authorities that want to incorporate chatbot programmes into their services.[4]

Since the communication of chatbots is entirely available in written form, a **combined use with machine translation** is possible. Therefore, it would be simple to adapt existing chatbots to regional or minority languages. With the use of chatbot technologies, administrative authorities can ensure that speakers of regional or minority languages can easily apply to them, submit written applications and receive a written reply in their language.

3.4.2 Smart search (Article 10.2.g)

The internet is used for many activities that are related to **searching for place names**, like booking trips, looking for directions or the weather forecast. On some websites, it is also possible to conduct these activities using place names in regional or minority languages. This is due to different smart search techniques. For example, when typing a place name in a regional or minority language into the search panel of booking.com, the website automatically conducts a cross search on other search engines, like Google, to find more results. This way, different versions of place names can be connected through search results from other websites, like Wikipedia.

Smart search therefore describes methods of connection between different search engines and of linking search results to real world entities such as objects, persons, or places. With AI, these search methods have advanced significantly and are able to handle varieties, e.g. in spelling, like typos. The search methods include:

- **Keyword extraction**: determination of key phrases in a text by analysing the frequency of words and their co-occurrence with other words;
- **Entity extraction**: classification of named entities in a given text into pre-defined categories, such as person, place name, time expression etc.;
- **Entity linking**: connection of named entity to knowledge base of real-world entity;
- **Web crawling**: automatic and systematic search of other websites, typically for the purpose of web indexing to learn the content of webpages in order to retrieve the information when needed; used to update content or indices of another site's content;
- **Semantic similarity/word embeddings**: translation of definitional relationships between words into mathematical space as described in the section on machine translation.

[4] European Commission, Directorate-General for Informatics & ISA2 Programme. (2019). Architecture for public service chatbots. Available online at https://joinup.ec.europa.eu/sites/default/files/news/2019-09/ISA2_Architecture%20for%20public%20service%20chatbots.pdf

These methods prove that websites do not work in a vacuum, but rather depend on each other to get more relevant results for their users. It also explains why place names in regional or minority languages are accepted on some search engines. For persons outside the respective company, it is quite difficult to influence this search process. Further use of smart search methods in search engines that do not employ these methods yet should be encouraged.

3.4.3 Speech synthesis for street name announcements (Article 10.2.g)

Speech synthesis describes the **artificial production of human speech** created by concatenating different pieces of recorded speech from a database. The more specific the domain, the longer these pre-recorded phrases can be. They could therefore either be just combinations of vowels and consonants or complete words, like time announcements. Through machine learning, speech synthesis has advanced and is now producing good results, especially for restricted domains, like street name announcements.

Speech synthesis can be used for facilitating the implementation of the Charter in different ways. One way would be the use of **bilingual announcements in public transport**. The announcements of the next stop in public transportation are already often generated using speech synthesis and could easily be expanded to the corresponding street names in regional or minority languages. Another possibility would be to incorporate regional or minority languages into accessible pedestrian signals. These signals communicate information about “walk” and “stop” intervals at various intersections in non-visual formats, like auditive signals, to those who are blind or have low vision. Some accessible pedestrian signals also convey additional information, for example about the street name or the direction of the crosswalk. When this information is communicated through speech messages, it can also be presented in a regional or minority language.

Since a big database of pre-recorded speech is needed for speech synthesis, its use is of more interest for regional or minority languages which already have a database available, for example because they are the majority language of other countries. In this context, transfrontier study and research exchange is encouraged since already existing databases of the majority language countries could be transmitted to countries in which the language is a regional or minority language. If databases already exist for one language, the implementation of street name announcements in other countries is easily feasible and can complement and, in some cases, replace the production of street name signs in regional or minority languages.

3.4.4 Machine translation (Articles 10.1.a.i-v, 10.1.b, 10.1.c, 10.2.a-f, 10.3.a-c, 10.4.a)

As described in previous sections, machine translation applications can be used for **“gisting”** or for completely automatically translating submitted documents, administrative texts, forms, or other official documents. Due to the restricted domain and formulaic structure and language of the latter, it is also possible to develop new machine translation applications for regional or minority languages that are not yet covered by big machine translation services.

3.5 Use of regional or minority languages in the media

The presence of regional or minority languages in different types of mass media is crucial to their protection and promotion. In this field, AI can assist through automatic subtitle generation (regarding

television programmes) and automatic information extraction (regarding newspapers). These applications may also contribute to the implementation of Article 9.4 of the Framework Convention.

3.5.1 Automatic generation of subtitles (Articles 11.1.a.i-iii, 11.1.c.i-ii)

The automatic generation of subtitles is a subfield of **automated speech recognition** or **speech-to-text** methods. It describes the process of automatically transcribing audio files, e.g. from a video, to subtitles. Subtitles are only concerned with the spoken language of a sound file, while closed captions also include other audio cues, like music or background noises. Automated speech recognition has a long research history and has recently benefitted from new technological advances.

To automatically generate subtitles, both an acoustic and a written language model are needed. These are mostly trained together to achieve better accuracy. Speech recognition is a complex matter, since it has to take into account variations regarding accent, pronunciation, articulation, pitch, volume or speed, as well as background noises, echoes etc. It works better for languages with a lot of training data of different speakers and communication situations in order to master natural communication like overlapping or spontaneous speech.

There are different applications that are concerned with the automatic generation of subtitles. One that is free of charge, widely spread and easy to use is **Google's automated speech recognition algorithm** used in YouTube. Currently, it supports nine European languages, namely Dutch, English, French, German, Italian, Portuguese, Russian, Spanish and Turkish. To use it, one has to upload a video to the YouTube editor and then activate automatic captioning. There is an instruction on YouTube's Help-page on how to do this, depending on the type of video. After the automatic subtitle generation, it is possible to proofread what has been recognised and download a text file with the subtitles. Since YouTube's subtitle generation algorithm is based on machine learning, it is constantly improving and being updated for new languages. Another helpful YouTube tool is **automatic caption timing**. To use it, one has to provide a text file with all the words from a video. The algorithm then finds the exact time points in the video where the text is spoken. There is also an in-depth instruction for this available on the YouTube Help-page. This tool works for the same languages as the automatic speech recognition and can be useful since creating a transcript of a video is far less time-consuming than timing the subtitles.

Using automatic generation of subtitles in regional or minority languages can help to make television programmes in such languages accessible to the hearing-impaired speakers. In addition, it promotes the visibility of regional or minority languages. Lastly, automatic subtitle generation makes machine translation of television programmes possible. One could, for example, imagine translated subtitles in regional or minority languages for important television programmes, or subtitles in the majority language for programmes produced in regional or minority languages. There are already several television broadcasters working on using AI for automatic subtitle generation, like the Franco-German television network Arte or the Catalan television channel Betevé. Therefore, this is also an opportunity for transfrontier research exchanges.

3.5.2 Automatic information extraction (Article 11.1.e.i-ii)

Automatic information extraction describes the task of automatically excerpting structured information from unstructured machine-readable documents, in most cases natural language texts. It is especially useful for groups of documents that follow a specific template, where every document in the group describes entities or events in a similar manner but with different details. To extract the

desired information, the document is first edited through different pre-processes, like attaching a **part-of-speech tag** (e.g. verb, adjective, preposition) to every word or finding the stem of every verb of the document. Then, various subtasks (like named entity recognition or coreference resolution) are performed to extract the requested information.

Automatic summarisation is one of the main applications of automatic information extraction. It describes the process of computationally shortening a set of data to create a subset of said data, a summary that contains the most relevant information from the original content. This can either be done extraction-based or abstraction-based. The aim of the first approach is to find key phrases in the document and extract the most relevant sentences. It can therefore be described as a sentence ranking issue. In the second approach, a (machine learning) algorithm first generates an internal semantic representation of the original content and then creates a short, accurate and fluent paraphrased summary. Obviously, this is computationally far more challenging than the first approach, which is already well researched and capable of delivering good results.

There are several online resources available for automatic extraction-based summarisation. With its help, one could publish shorter versions of news articles in the regional or minority language on other pages or automatically shorten news articles in a majority language so that they can be translated faster and easier. Both methods can help to simplify the **publication of newspaper articles** in regional or minority languages on a regular basis.

3.6 Use of regional or minority languages in cultural activities and facilities

Regional or minority languages represent a rich cultural heritage, history and identity. Therefore, maintaining and promoting their cultural profile forms a significant part of supporting such a language. Furthermore, modern cultural initiatives can improve a language's image, especially among the younger generation, and play an important role in developing a living language. The importance of AI in this area may not be as obvious as in other areas, but NLP applications like data structuring, machine translation and automatic subtitle generation can be of help here as well. These applications may also contribute to the implementation of Article 5.1 of the Framework Convention.

3.6.1 Data structuring (Articles 12.1.g, 12.1.h)

Data structuring describes the process of automatically structuring or clustering large quantities of data with the help of AI. The task of **classifying documents** into different areas is one of the main applications of NLP.

There are various algorithms and techniques (Naive Bayes, tf-idf[5], Support Vector Machines, etc.) used to recognise patterns in databases. For each new document added to the database, the algorithm decides, based on previously categorised documents, to which category the new document belongs.

The training of algorithms for this kind of tasks can either be done through **supervised or unsupervised learning**. In the first case, a neural network is trained on a labelled dataset, in the latter case the network measures the content-related differences between documents by mapping them into a mathematical space and calculating the "document distance". This method was described in more detail in the section about machine translation and is in this case called **data clustering**.

[5] Term frequency-inverse document frequency

Data structuring can be a helpful tool especially for large unstructured data, but also for organising already structured data.

For the facilitation of the implementation of Article 12, it can be used for the **classification of publishing works** in regional or minority languages and for **structuring terminological databases** of such languages. Using intelligent data structuring methods for the development and maintenance of these kind of databases can help cultural institutions to reduce time and cost immensely. There are several providers and tutorials available for this, depending on the size, language and topic of data as well as the country concerned. Moreover, these data structuring projects could be study and research opportunities for students from different data study programmes.

3.6.2 Machine translation (Articles 12.1.a, 12.1.b, 12.1.c)

It is challenging to apply machine translation to cultural works. Even for professional translators, literary work is complicated to translate. For some highly language-dependent works, like poetry, translation may even be impossible. Obviously, translating literary works automatically is even more difficult. Machine translation can still be of use here, for example for **translating summaries or subtitles**. Through translated summaries, a higher interest in regional or minority language works could be generated because speakers of other languages could be more motivated to read them if the basic plot is already translated. Needless to say, this method also works the other way around to foster access in regional or minority languages to literary works produced in other languages. Furthermore, literary works from which the translated summaries are often requested could then be manually translated, since a general interest in the work could be established.

3.6.3 Automatic generation of subtitles (Articles 12.1.b, 12.1.c)

In Article 12, different “subtitling activities” are mentioned. As described earlier, these can be carried out automatically through the help of speech recognition. This reduces the time and financial efforts since the subtitles only have to be checked manually and not be done completely by hand. For already working examples of automatic subtitle generation and the different options in this area, see the more detailed description in the previous section on Article 11.

3.7 Use of regional or minority languages in economic and social life

For regional or minority languages to be fully functional, they must be used in every aspect of economic and social life. AI can help achieving this goal by using such languages in sentiment analysis and, again, machine translation.

3.7.1 Sentiment analysis (Article 13.1.c, 13.1.d, 13.2.b)

Sentiment analysis describes the use of NLP to **classify a document of subject information** (e.g. a review or survey response) into an opinion category, like positive/satisfied, neutral, negative/angry etc. It is useful for quickly gaining insights into large text data.

Sentiment analysis is a classification process that works similarly to the document classification described in the section on data structuring above. The neural network is first trained, either supervised or unsupervised, to learn to associate an input (e.g. a movie review) to a tag (e.g. “positive”). This can either be done completely through the use of machine learning, or with manually labelled rules, like “great” = “positive”. In the testing process, the network gets unseen input texts and

transforms them into word vectors. This process is called **feature extraction** and is described in more detail in the section about machine translation above. Afterwards, these vectors are processed by a classification algorithm, which then creates a tag for the document, stating for example that the tweet "The movie was really great!" is "positive" feedback.

Sentiment analysis is mostly used for **customer feedback analysis**. Developing programmes for sentiment analysis of regional or minority languages would mean to no longer exclude but include the speakers into these dialogues about customer satisfaction and ways to improve a company's service.

There are several companies providing services in sentiment analysis, depending on the country of origin, as well as the language, size and type of sentiment analysis desired. Again, this could also be an opportunity to promote research and study on regional or minority languages. Sentiment analysis is a highly relevant and current research topic with many online tutorials that is covered in all NLP and AI beginner classes. Given the available data, developing such a system is a feasible task e.g. for student groups from different programming or computational studies.

3.7.2 Machine translation (Articles 13.1.a, 13.1.d, 13.2.a, 13.2.b, 13.2.d, 13.2.e)

Several provisions in Article 13 require either directly or indirectly the translation of different documents into regional or minority languages, notably safety instructions and information on consumer rights as well as contracts, technical documents, payment orders or other financial documents. Through machine translation, these documents can be made available in the required language faster and cheaper. Since the documents all belong to specific domains and are comprised of formulaic language, an automatic translation is possible, either for "gisting" or for completely automatically translating the documents. It is also possible to develop new machine translation applications for regional or minority languages that are not yet covered by such applications. As already mentioned, this is also an interesting study and research opportunity. However, especially for the domain of safety instructions, **proofreading** is even more important here than it is for other areas. Small translation mistakes, as they still happen with machine translation, could have severe consequences.

3.8 Use of regional or minority languages in transfrontier exchanges (Articles 7.1.i, 14)

Crossborder co-operation significantly facilitates the promotion of regional or minority languages that are also official languages or minority languages in other states because the existing infrastructure in fields such as education (e.g., teaching materials, teacher training) or media can be taken over or adapted.

As could be seen in this report, using AI to facilitate the implementation of the Charter encourages **transfrontier research and study exchanges** regarding applications and datasets. These kinds of materials are international by nature and regional or minority languages can benefit from NLP applications of other states, especially those where the language is the majority language. Furthermore, given data in another language, these applications can easily be adapted to other regional or minority languages when the source code is made available to other research teams. If, for example, an administrative authority in a state develops a chatbot for its services, it can easily be used in other countries where the same regional or minority language is used or easily be adapted to another regional or minority language. That way, AI not only facilitates the implementation of the

Charter, but also favours and encourages transnational exchanges. Therefore, it can also contribute to the implementation of Article 18 of the Framework Convention.

Outlook

Due to the wide range of available applications and the rapid changes in the area of AI technologies, this report could only provide a first general overview and give some exemplary impulses on how different NLP applications can help states parties to implement the European Charter for Regional or Minority Languages and, to some extent, the Framework Convention for the Protection of National Minorities. This report claims in no way to be exhaustive since there are many other resources, applications and possibilities available to use AI for the Charter's implementation. Moreover, the impulses given in this report can all be starting points for new discussions and study projects.

The present report showed that artificial intelligence and natural language processing create several new possibilities for the protection, research and promotion of regional or minority languages. However, due to the current domination of English in the field of AI, regional or minority languages also face the risk of "Digital Language Extinction". To prevent this, resolute action is important, especially in the area of **data collection**. With sufficient data behind them, like parallel corpora of the regional or minority language and the official or majority language, the regional or minority languages are given numerous new possibilities to establish themselves as modern, relevant and living languages. This means that for almost every activity proposed in this report, the gathering of relevant natural language data is the first step and should therefore be addressed as early as possible to secure the language's status in modern language processing technologies.

The research field of AI and NLP is especially fast paced, and all applications mentioned in this report are constantly being improved. Most likely, many of the proposed half-automatic methods will be fully automated in the years to come and those methods that have been described as rudimentary today will produce human-like results in the future. AI is already of use to support the States Parties implementing the Charter, and the support that AI applications provide will most likely only be amplified in the coming years. It is therefore worth the effort to invest in NLP applications for regional or minority languages to profit from them already today and even more so tomorrow.

Statement of the Committee of Experts of the European Charter for Regional or Minority Languages on the promotion of regional or minority languages through artificial intelligence

Adopted on 16 March 2022

Since the drafting of the European Charter for Regional or Minority Languages in the 1980s, different new technologies have improved the conditions for its implementation by the states parties. The Committee of Experts has already examined how new social media help to increase the media offer in regional or minority languages.[6]

The rise of artificial intelligence (AI) marks a new era of technology which can also facilitate the everyday use and promotion of regional or minority languages and hence support states parties in implementing the Charter provisions which they have ratified. The Council of Europe is currently preparing a legal framework on AI, based on the Council of Europe's standards on human rights, democracy and the rule of law.

The Committee of Experts welcomes the development of AI applications using regional or minority languages. This requires the gathering of natural language data, which is of particular importance for the documentation of less-widely used languages. It needs to be borne in mind that AI constitutes an addition to the learning and use of regional or minority languages. Particular attention must be paid to developing and/or including appropriate administrative and legal terminology in each regional or minority language. Once developed, AI applications facilitate the daily use of regional or minority languages, disseminate such languages to larger audiences, raise their visibility and prestige, and encourage more people to learn, use and transmit them to the next generations.

With the help of AI applications, authorities can fairly quickly make available an offer for users of regional or minority languages, including in less-widely used languages. The use of AI therefore supports authorities in taking "resolute action to promote regional or minority languages in order to safeguard them", which is one of the central objectives and principles of the Charter.

Having taken note of the study "Facilitating the Implementation of the European Charter for Regional or Minority Languages through Artificial Intelligence"[7], the Committee of Experts encourage states

- to include the promotion of the use of regional or minority languages in their policies, legislation and practice on digitalisation,

- to promote the inclusion of regional or minority languages into research and study on AI with a view to supporting the development of applications facilitating their use in public and private life,

- to develop, in co-operation with the users of regional or minority languages and the private sector, a structured approach to the use of AI applications in the different fields covered by the Charter.

[6] Council of Europe (ed.): New technologies, new social media and the European Charter for Regional or Minority Languages, Report for the Committee of Experts, 2019

[7] Council of Europe Secretariat of the European Charter for Regional or Minority Languages (ed.): Facilitating the implementation of the European Charter for Regional or Minority Languages through artificial intelligence, 2022

Prems 031422

ENG

www.coe.int

The Council of Europe is the continent's leading human rights organisation. It comprises 46 member states, including all members of the European Union. All Council of Europe member states have signed up to the European Convention on Human Rights, a treaty designed to protect human rights, democracy and the rule of law. The European Court of Human Rights oversees the implementation of the Convention in the member states.

COUNCIL OF EUROPE

CONSEIL DE L'EUROPE